Aalto University
School of Science
Master's Programme in Mathematics and Operations Research

Sara Melander

Survival regression model for rolling stock failure prediction

Master's Thesis
Espoo, April 29, 2019

Supervisor: Professor Antti Punkka
Advisors: Ville Mattila M.Sc. (Tech.), VR Maintenance Ltd
Otto Sormunen D.Sc. (Tech.), VR Maintenance Ltd

Aalto University
School of Science

Aalto University
School of Science
Master's Programme in Mathematics and Operations Research

ABSTRACT OF MASTER'S THESIS

Author:	Sara Melander		
Title:	Survival regression model for rolling stock failure prediction		
Date:	April 29, 2019	**Pages:**	vii + 97
Major:	Systems and Operations Research	**Code:**	SCI3055
Supervisor:	Professor Antti Punkka		
Advisors:	Ville Mattila M.Sc. (Tech.) Otto Sormunen D.Sc. (Tech.)		

In railway operations where functioning machinery is a key to success, maintenance is an important part of the business. At the Finnish rolling stock maintenance company VR Maintenance Ltd, maintenance consists of repair due to failures and preventive maintenance actions. Failures happen unexpectedly, whereas the preventive maintenance actions follow a predetermined programme. Currently, over 50% of the maintenance at VR consists of repair. To improve the maintenance operations at VR, this thesis develops a prediction model for rolling stock failure prediction. The model serves as a tool for short-term maintenance scheduling, for budgeting, and for analysing the current failure process to help develop the maintenance programme.

The prediction model consists of accelerated failure time models (AFTM) developed for each failure category for each rolling stock fleet. AFTM is a survival regression model, where a number of factors found to affect the time to failure are linked to a failure distribution. In this thesis we use a Weibull AFTM, where the baseline failure distribution is a Weibull distribution. We found five factors affecting the failure process significantly: the month, the average speed, the kilometres since previous preventive maintenance action, the age of the rolling stock, as well as the split between "good" and "bad" rolling stock individuals. The model is based on failure and maintenance records from a two-year period. Using Monte Carlo simulation, the model outputs the predicted number of failures during the specified prediction period with confidence intervals. The model also predicts the total repair time for these failures as well as the number of critical failures.

The prediction accuracy of the model was found lacking. Some failure categories were accurately predicted, whereas others need to be improved further. Further model development and a pilot study to test its usefulness are encouraged, as utilising a prediction model in the maintenance operations is expected to boost fleet reliability and availability, and reduce maintenance costs. It can also be utilised to further develop the operations, e.g. as a building block in an opportunistic maintenance optimisation model or in enhancing condition based maintenance.

Keywords:	rolling stock, maintenance planning, failure prediction modelling, accelerated failure time model
Language:	English

Aalto-universitetet
Högskolan för teknikvetenskaper
Master's Programme in Mathematics and Operations Research

SAMMANDRAG AV DIPLOMARBETET

Utfört av:	Sara Melander		
Arbetets namn:	Överlevnadsregressionsmodel för att förutspå fel i tågutrustning		
Datum:	29 april 2019	**Sidantal:**	vii + 97
Huvudämne:	Systems and Operations Research	**Kod:**	SCI3055
Övervakare:	Professor Antti Punkka		
Handledare:	Diplomingenjör Ville Mattila Teknologie doktor Otto Sormunen		

Underhållsarbete är en viktig del av tågverksamheten, eftersom verksamhetens lönsamhet bygger på fungerande utrustning. Hos den finska underhållsoperatören för tågutrustning VR Underhåll Ab består underhållsarbetet av reparationer av fel samt av förebyggande underhåll. Fel uppstår oväntat medan det förebyggande underhållet följer ett program. I nuläget består över 50% av underhållsarbetet på VR av reparationer. För att förbättra VR:s underhållsverksamhet utvecklar det här arbetet en prediktionsmodell för att förutsäga fel som uppstår i tågutrustningen. Modellen fungerar som ett verktyg för kortsiktig underhållsplanering, för budgetering samt för att analysera hur och när fel uppstår, vilken stöder utvecklandet av underhållsprogrammet.

Prediktionsmodellen består av modeller för accelererad tid till fel (engelska: accelerated failure time model) (AFTM) som utvecklats skilt för varje felkategori och varje tågflotta. AFTM är en regressionsmodell för överlevnadsmodellering, där ett antal variabler som konstaterats ha ett samband till överlevnadstiden kopplas till en fördelning som beskriver tiden tills ett fel uppstår. I det här arbetet använder vi en Weibull AFTM, där basfördelningen är en Weibullfördelning. Vi identifierade fem relevanta variabler: månaden, medelhastigheten, den körda sträckan sedan senaste förebyggande underhållsarbete, tågets ålder samt en indelning av braöch dåligatåg. Modellen baserar sig på fel- och underhållsdata från en tvåårsperiod. Genom att använda Monte Carlo simulering producerar modellen en prognos med konfidensintervall över antalet fel som uppstår under en given period. Modellen förutsäger också den totala reparationstiden för felen, samt antalet kritiska fel.

Modellens prediktionsnoggrannheten är dock inte tillräckligt bra. En del felkategoriers modeller behöver vidareutvecklas för att garantera noggranna förutsägelser. Vidareutveckling av modellen och testning av dess nytta genom ett pilotprojekt uppmuntras eftersom användandet av en prediktionsmodell i underhållsverksamheten förväntas förbättra tågens pålitlighet och tillgänglighet samtidigt som underhållskostnaderna sjunker. Modellen kan också användas för att vidareutveckla verksamheten, till exempel som en del av en optimeringsmodell för opportunistiskt underhåll eller för att förbättra riskbaserat underhåll.

Nyckelord:	tågutrustning, underhållsplanering, felmodellering, accelerated failure time model
Språk:	Engelska

Acknowledgements

I wish to thank VR maintenance for granting me the opportunity to write this thesis. The excellent guidance and support from my advisors Otto and Ville have been invaluable throughout the process. A special thanks also to my supervisor Antti for his efforts in overseeing my thesis and giving me valuable feedback and encouragement in the finishing of it.

What a strange feeling closing this chapter of my life, which has come to give me so much more than just a degree. Otaniemi is such a unique place and the community at Aalto University and especially at Teknologföreningen has made this experience extraordinary. Words cannot express the gratitude I feel for the amazing opportunities these seven years have given me, not to mention all the wonderful people I have met along the way.

I especially want to thank my mother Åsa for the unconditional support throughout my studies and in the writing of this thesis. Without my study partner and dear friend Nina these years would have been so much harder and way duller. Finally, I thank my beloved Niklas for all his love and care, which has helped me make it this far.

Otaniemi, April 29, 2019

Sara Melander

Abbreviations and Acronyms

ABAO	"As bad as old", condition after repair
AFTM	Accelerated failure time model
AGAN	"As good as new", condition after repair
c-index	Concordance index
CBM	Condition based maintenance
CI	Confidence interval
CM	Corrective maintenance
FMECA	Failure modes, effect and criticality analysis
FTA	Fault tree analysis
HPP	Homogeneous Poisson process
KS-test	Kolmogorov-Smirnov test
NHPP	Non-homogeneous Poisson process
OM	Opportunistic maintenance
PDF	Probability distribution function
PdM	Predictive maintenance
PHM	Proportional hazard model
PLP	Power law process
PM	Preventive maintenance
RBD	Reliability block diagram
RP	Renewal process

Contents

Chapter 1

Introduction

Maintenance is a key function in many businesses, as it guarantees that the operations can run smoothly and ensures safety (Dekker, 1996). According to Stern et al. (2017), maintenance expenses account for approximately 50% of the overall costs in railway operations. Fleet availability and reliability are key areas of interest for railway operations. These can be optimised with an effective maintenance strategy, while still minimising maintenance costs (Duffuaa et al., 2015). Implementing a well suited maintenance strategy for the operations may also improve resource management efficiency and reduce failures (Cheng and Tsao, 2010).

The maintenance strategy defines the implementation of two main maintenance approaches: repair due to failure and preventive maintenance (Wang, 2002). Failures generally happen unexpectedly and are repaired as they occur or when needed. Preventive maintenance refers to pro-active maintenance actions taken to prevent failures while the equipment is still in working condition, mainly to increase reliability and reduce maintenance costs. Preventive maintenance is often cheaper than repair, as repairing a failure might require more extensive work than performing pro-active maintenance. Also, whereas failures are unexpected, pro-active maintenance follows a predetermined programme of scheduled maintenance actions (Stern et al., 2017). The preventive maintenance programme can be based on driven kilometres, observed condition of the system, opportunities for maintenance due to shut down of the system, number of production cycles run, or other predetermined factors (Duffuaa et al., 2015). The maintenance strategy should be tailored for the specific system, as the system's structure and interdependencies affect the suitability of different strategies.

Maintenance scheduling is an element of maintenance operations, which Duf-

fuaa et al. (2015) define as the process of assigning resources and manpower to maintenance jobs and determining when to perform them. The authors argue that the effectiveness of a maintenance system is greatly affected by the quality of the maintenance schedule and its ability to adapt as changes appear. Changes are due to happen, since the need for maintenance is dependent on the failures that occur. As failures occur unexpectedly, being able to accurately predict the need for repair will make the maintenance system more efficient.

At the Finnish rolling stock maintenance company VR Maintenance Ltd (henceforth referred to as VR), the maintenance strategy consists of preventive maintenance actions based on usage, time, or component condition. However, currently over 50% of the maintenance work is repair due to failures. The failures happen unexpectedly and even if repairing them accounts for the majority of the overall maintenance, failures are currently not being extensively predicted. This contributes to challenges with weekly and daily maintenance scheduling. Hence, rescheduling and delays are not uncommon, which increase maintenance costs and reduce efficiency. In addition, a comprehensive understanding of the failure process for the rolling stock fleets is currently lacking. Being able to predict failures would grant opportunities to improve several elements of the maintenance operations, resulting in reduced maintenance expenses and improved reliability and availability of the rolling stock fleets.

This thesis develops a failure prediction model for rolling stock failures at VR. The prediction model mainly provides a tool for enhanced weekly maintenance scheduling. It can also be utilised for the improvement of the preventive maintenance programme, as an accurate modelling of failure in combination with an understanding of the failure process are bases for developing an optimal programme. An estimate for upcoming failures would also serve as a basis for budgeting. Factors affecting the failure process are identified through expert interviews and analysis of maintenance and failure records from a two-year period. These records are also used for developing the forecasting model, which predicts failures per rolling stock fleet and failure category. The failures are categorised according to the location of the rolling stock they occur in, for example, doors and entrances, brakes, or air-conditioning.

For each failure category and each of 12 rolling stock fleets (6 locomotive fleets and 6 electric multiple unit fleets) we fit an accelerated failure time model (AFTM) to the observed inter-failure kilometres. The AFTM is a survival regression model, in which a number of factors having an effect on the failure process are linked to the underlying survival function. This allows to consider

the effect of these factors in the modelling of failures. For the rolling stock failure processes we identify five relevant factors, which we include in the model: the month, the average speed, the number of kilometres driven since previous preventive maintenance, the age of the rolling stock, and the split between identified "good" and "bad" individuals.

Figure 1.1 presents the structure of the thesis and illustrates how the chapters are connected. The rest of the thesis is structured as follows. Chapter 2 gives an introduction to maintenance planning and maintenance strategies, along with key concepts related to these as presented in the literature. The current maintenance planning protocol at VR is also presented in the chapter. A literature review on failure modelling and failure prediction is presented in Chapter 3. Both quantitative and qualitative methods and models as well as some case study examples are discussed. Thereafter, Chapter 4 presents the maintenance and failure records, which are used in Chapter 5 to implement a model for failure prediction at VR. In Chapter 5 we describe the development of accelerated failure time models and how these are combined into a complete model for predicting the overall failures during a given time period. The prediction accuracy of the model is presented and discussed in Chapter 6, which also reviews the utilisation of the failure prediction in the maintenance operations at VR, as well as the limitations of the prediction model. Finally, the conclusions and ideas for further development are presented in Chapter 7.

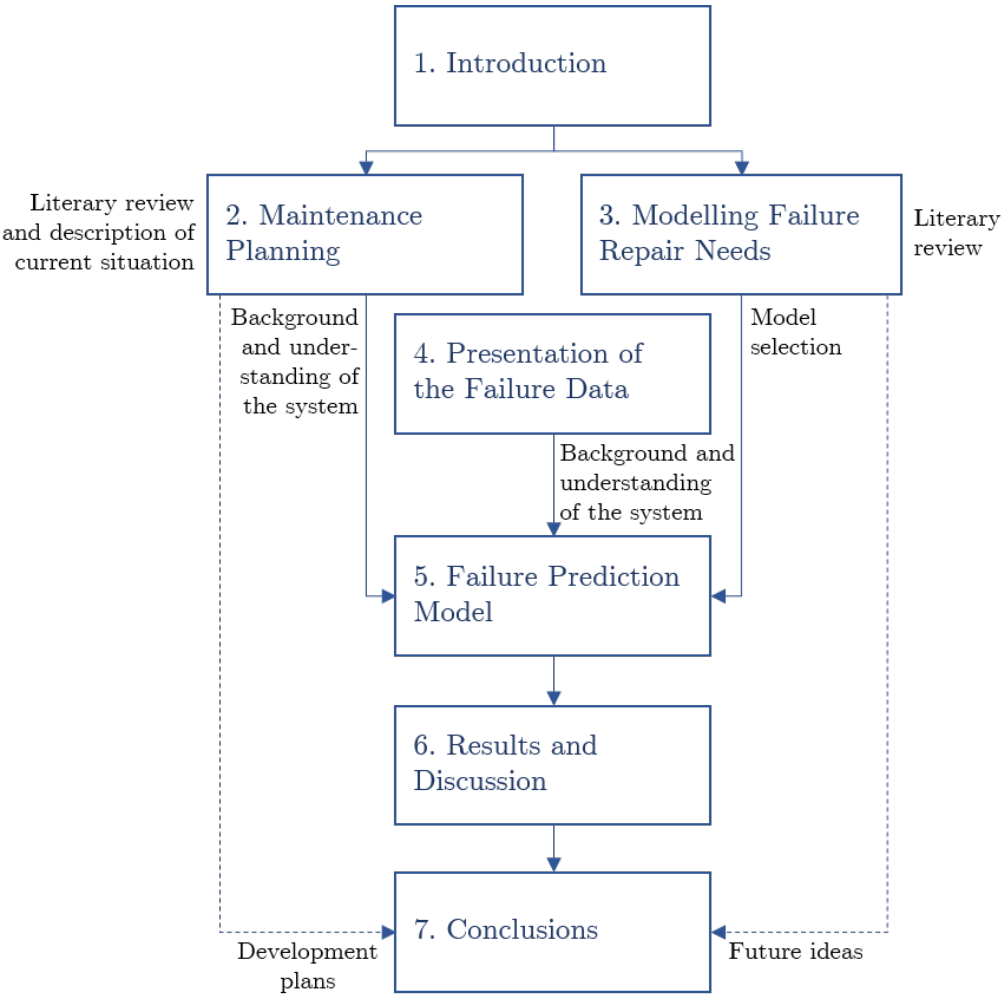


Figure 1.1: Illustration of the structure of the thesis and the connections between the chapters.

Chapter 2

Maintenance Planning

In this chapter, we discuss different maintenance strategies and concepts presented in the literature. Thereafter, we relate the maintenance strategy at VR to these concepts and describe the current maintenance planning process as well as future improvement plans. Based on interviews with key personnel we identify key development opportunities for the process and present how a failure prediction model would serve as a tool in different areas of the maintenance planning process.

2.1 Maintenance Strategies

Duffuaa et al. (2015) discuss how the maintenance strategy should be in line with the organisation's overall strategy and vision. The maintenance strategy defines issues such as maintenance outsourcing and organisation, but also includes maintenance methodology. The maintenance methodology refers to the equipment level maintenance strategy, that is, it describes the approach for maintaining the equipment, mainly defining the incorporation of pro-active maintenance. This is what is meant by maintenance strategy in this thesis. Table 2.1 presents the main maintenance strategies that are described further in this section. Generally, the overall maintenance strategy is a mix of these strategies.

The basic type of maintenance is corrective maintenance (CM), which is needed to restore the system to working condition when a failure occurs (Pham and Wang, 1996). Corrective maintenance can also be referred to as repair. As the need for corrective maintenance often is unforeseen, it may lead to unexpected and costly breaks in the operations. When a failure oc-

Table 2.1: A summary of the maintenance strategies presented in this section. CM and PM are the main maintenance approaches, whereas CBM, PdM and OM are approaches to define how PM is performed. (Stern et al., 2017)

Corrective maintenance	CM	Also referred to as run-to-failure strategy. The system is maintained (repaired) only at failure.
Preventive maintenance	PM	Pro-active maintenance actions are taken to prevent failures.
Condition based maintenance	CBM	The condition of the system determines what (if any) pro-active maintenance action should be taken.
Predictive maintenance	PdM	Pro-active maintenance is performed if the future condition of the system is predicted to fall below a predetermined threshold.
Opportunistic maintenance	OM	Maintenance is performed as an opportunity arises.

curs, there are different grades of restoration that can be sought with the corrective maintenance action: minimal repair, imperfect repair or perfect repair. Minimal repair refers to only repairing the system as much as needed for it to be in operating condition again, restoring the system to an "as bad as old" (ABAO) state. Perfect repair usually refers to replacing the system with a new one, hence bringing it back to an "as good as new" (AGAN) condition. Pham and Wang (1996) however address, that in reality a maintenance action usually restores the system to a state between ABAO and AGAN, indicating imperfect repair. Due to the stochastic nature of failures, implementing corrective maintenance as the only maintenance strategy may be costly as maintenance actions are not planned and consequently the maintenance needs are unexpected. (Duffuaa et al., 2015)

To minimise failures and consequently the need for corrective maintenance, as well as to reduce maintenance costs, pro-active maintenance actions can be taken. These actions are performed when the system is still in operating condition and are called planned or preventive maintenance (PM). Numerous maintenance policies have been developed to determine when to take which pro-active maintenance action. As in preventive maintenance strategies pro-active maintenance actions are by definition part of the maintenance strategy, the maintenance policy defines how these preventive actions are implemented (Duffuaa et al., 2015). Wang (2002) reviews traditional preventive maintenance policies developed and researched in the second half of the 20^{th}

century.

A widely used and extensively studied preventive maintenance policy is the age-dependent policy, where a component is replaced at failure or at a predetermined age T, depending on which occurs first. Numerous extensions and generalisations for the age-dependent policy have been developed, see for example Sheu et al. (1995) and Block et al. (1993). Another popular policy is the failure limit policy. Under this policy some reliability index, such as the failure rate, is audited. As the index falls below a predetermined threshold and the probability for failure grows larger than a tolerance level, the system is preventively maintained to reduce the probability of failure. Historical failure and maintenance data can be exploited to determine the system specific optimal maintenance policy. (Wang, 2002)

More advanced maintenance strategies defining when to perform preventive maintenance are

- Condition based maintenance (CBM)
- Predictive maintenance (PdM)
- Opportunistic maintenance (OM).

Condition based maintenance determines what preventive maintenance actions to take based on monitoring the condition of the system, for example, by sensors or with inspections. Based on pre-determined rules, the state of the system indicates if a preventive maintenance action should be performed or not. Extensive historical data on failures and condition states is needed to determine the condition thresholds for preventive maintenance actions. (Alaswad and Xiang, 2017)

Predictive maintenance is based on an estimation of the state of the system at a future time. If the system is expected to fall below a specified threshold before the next maintenance, an additional pro-active maintenance action is taken. Predictive maintenance requires not only monitoring of the condition of the system, but also monitoring of factors affecting the system. Utilising machine learning, notifications can be given when the system's condition is predicted to fall below the allowed threshold. This way, the system can be preventively maintained before a failure occurs. (Stern et al., 2017)

Opportunistic maintenance is, as the name indicates, performed as the opportunity arises. Opportunities include another maintenance occasion or a system shut down (Duffuaa et al., 2015). Urbani (2017) describes how opportunistic maintenance considers both preventive maintenance and corrective

maintenance. Utilising predictive maintenance, estimates on which maintenance task to perform at any maintenance occasion can be derived. He also suggests that a well implemented opportunistic maintenance strategy increases system availability and reduces maintenance costs as several maintenance actions are performed at the same time and additional set-up costs associated with any maintenance occasion can be avoided.

Many models and methods for determining an optimal or a well suited maintenance strategy have been developed. A review is presented by Sharma et al. (2011), while Dekker (1996) summarises case studies of applications of maintenance optimisation models. To optimise the maintenance policy, an understanding of the system is needed, which typically requires historical data on maintenance actions and failures. This indicates that the model needs to be customised for the specific case, as the parameters are system specific (Urbani, 2017). Wang (2002) presents that the objective for a well suited maintenance strategy is mainly to maximise system reliability or to minimise system maintenance costs. Generally however, a combination of these two is needed as a certain level of reliability needs to be guaranteed and the maintenance costs cannot grow too large. Hence, he defines the main objectives for finding the optimal maintenance strategy as either minimising system maintenance cost rate while reliability requirements are met, or as maximising reliability measures while system maintenance cost rate requirements are met.

According to their report on rolling stock maintenance Stern et al. (2017) identify the maintenance strategy for rolling stock greatly affected by strict safety regulations dictating the maintenance for security components and components leading to train failure, such as brakes. These components together with highly visible quality components, such as air-conditioning, are generally preventively maintained. Other components are maintained at failure, hence stoppage due to failure is inevitable at some point. They also conclude that the new digital developments enable the exploitation of condition based maintenance and predictive maintenance in rolling stock maintenance operations, which may come to increase efficiency of maintenance operations by 15-25%. However, these strategies have not yet been implemented to such an extent in railway operations and at least the leap to predictive maintenance will require extensive investments.

2.2 Current Maintenance Planning Process at VR

For the implementation of a failure prediction model to improve elements of the maintenance operations at VR, we have studied the current maintenance planning process. We have interviewed key personnel at VR to get an understanding of how the planning and scheduling are currently organised. Understanding the current process allows us to identify trouble points and how parts of the process can be improved with a failure modelling tool. The interviews in combination with an analysis on maintenance history and failure data contributes to the understanding of the current situation.

According to the insights of the maintenance planners Kaarela (2018), Lehtola (2018) and Levo (2018), the maintenance planning process at VR consists of three levels:

1. **Long-term planning**, where the preventive maintenance strategy and possible additional preventive maintenance actions are determined. Frequently occurring failures for each rolling stock fleet are identified, as well as seasonal fluctuations in the number and type of failures.

2. **Short-term planning**, which consists of scheduling the maintenance activities for the next week. Only preventive maintenance actions are scheduled, based on driven kilometres and time. Repair of failures preventing the rolling stock from being used are scheduled as these failures occur. Smaller failures not preventing usage, are usually repaired when the rolling stock arrives at the depot for a preventive maintenance or to repair a bigger fault. Short-term maintenance planners also consider the depot shift schedule, since not all maintenance staff have the knowledge or the certificates to perform all maintenance actions. Additionally, the maintenance planners take into account the current location of the rolling stock and direct it to a depot nearby with enough capacity, i.e. personnel, vacant maintenance slot and time.

3. **On-site planning**, where the exact maintenance actions are chosen based on the resources available and the criticality of the failures. Resources refer to maintenance slots, personnel, materials and time. Changes to the weekly schedule are common due to unexpected failures and delays in arrival times at the depot.

The current maintenance system and maintenance planning protocol at VR

have elements of corrective maintenance, preventive maintenance, condition-based maintenance as well as opportunistic maintenance.

The maintenance planners estimated that over 50%, even up to 80%, of the time spent on maintenance actions is used for repairs. This observation is backed up by analyses of the maintenance records. This means that corrective maintenance plays a big role in the current maintenance system. The short-term planning mainly focuses on scheduling preventive maintenance actions for the next week. Observed failures or estimated failure occurrences for the next week are only considered to a limited degree, as some time and capacity are reserved for corrective maintenance mainly based on expert judgement. Thus, this process has potential for improvement by utilising e.g. survival regression modelling for more accurately predicting the future needs for corrective maintenance.

Opportunistic maintenance is present in the current process as non-critical failures are repaired in conjunction with other maintenance actions. Smaller pro-active maintenance actions can also be conducted during a depot visit, if the maintenance schedule allows them and there are available resources. These pro-active maintenance actions can be based on the observed condition of a component. For example, if the thickness of the brake pad is measured to be below or close to a given threshold, it can be replaced.

The planning tool for weekly maintenance scheduling indicates preventive maintenance actions to be performed and this information dictates the schedule for the next week's maintenances. Levo (2018) expresses that unexpected critical corrective maintenance needs may result in planned preventive maintenance actions being moved. This is an issue since preventive maintenances actions need to be performed after a predetermined amount of driven kilometres or days, and a rescheduling might put the rolling stock on hold due to the kilometres or the number of days being reached, hence prohibiting the rolling stock from being used. On-site maintenance foreman Kehä (2018) also adds that currently preventive maintenance actions are scheduled very close to the maximum tolerance limit, resulting in little flexibility in the schedule and problems when unexpected critical failures happen. Kaarela (2018) and Levo (2018) believe that a closer examination of failures could reduce the need for preventive maintenance rescheduling due to capacity needed for critical corrective maintenance.

According to Lehtola (2018), material shortage is another issue resulting from the uncertain characteristics of the failure process. If materials are not available in stock, the repair is rescheduled, causing possibly a longer period of the rolling stock being unused before the needed material has been deliv-

ered. A failure prediction model could prevent possible material shortages, as needed materials could be ordered based on the failure estimate.

The maintenance system is under constant development and improvements are in the pipeline. Levo (2018) together with maintenance development manager Annala (2018) describe how condition based maintenance is already incorporated in e.g. maintaining of brakes with the aim of utilising condition based maintenance to an ever more complete degree in maintenance for all kinds of components. Another plan is to utilise machine learning to develop an understanding of which failures are correlated, namely, if certain failures can be predicted based on previous failures. Kaarela (2018) explains how the fleet engineers follow up on the failures and continuously develop the preventive maintenance programmes to reduce common failures. According to fleet engineer Parta (2018), there is an understanding of which failures become more frequent during the winter, and hence preventive maintenance is performed in autumn to reduce these failures and to reserve more maintenance capacity for the increased number of failures during the winter. In her bachelor's thesis Torpo (2019) examines different opportunistic maintenance modelling strategies. Preliminary plans have been made to combine a model like hers with an appropriate failure prediction model to produce optimal combinations of maintenance actions to perform at each depot stop. These combinations can be utilised in the on-site maintenance planning.

Based on the interviews, we decided to develop a model for failure prediction, which could shed some light on the characteristics of the failure process for different fleets and also provide an estimate for the expected number of failures during the following planning period. We found that this type of predictive model could serve as a tool for budgeting as well.

The planned use of the failure prediction model in the maintenance planning process is presented in Figure 2.1. The long-term maintenance planning provides insight into the factors affecting the failure process, and hence on which parameters to include in the failure prediction model. The output of the model depends on the user's input and on the historical failure and maintenance records from a specified time period.

The long-term maintenance planning would benefit from a failure prediction model, as it would serve as a tool for identifying rolling stock that are performing worse than others and as the model would provide insight on commonly occurring failures. Based on these insights, additional preventive maintenance actions could be incorporated in the maintenance program. The predicted number of failures could also serve as a basis for budgeting. The current budgeting process relies on historical failure and maintenance

records, when budgeting the maintenance needs for the following year. A failure prediction tool would allow for including the effect of relevant factors in the estimated number of failures. It could also be used for scenario analysis where e.g. the predicted failures with different number of driven kilometres for the next year could be examined.

In the short-term maintenance planning process, the failure prediction can be utilised in the scheduling and assigning of resources for next week's maintenance jobs. If a larger number of failures are predicted, more time and resources should be reserved for repair. Consequently, if the model predicts less failures, there is time to perform further preventive maintenance actions. If the short-term maintenance schedule is accurate, namely, adequate capacity has been reserved for corrective maintenance, the need for changes to the maintenance schedule is smaller, hence making the on-site maintenance planning easier.

Furthermore, in line with future plans, the output of the failure prediction model can also be exploited in an opportunistic maintenance model, defining optimal combinations of maintenance actions to be performed. Should such a model be developed, will the output of it together with the short-term maintenance schedule affect the on-site maintenance planning. Optimal opportunistic maintenance combinations and accurate weekly maintenance schedules enhance the maintenance work, which improves equipment reliability and reduces maintenance costs.

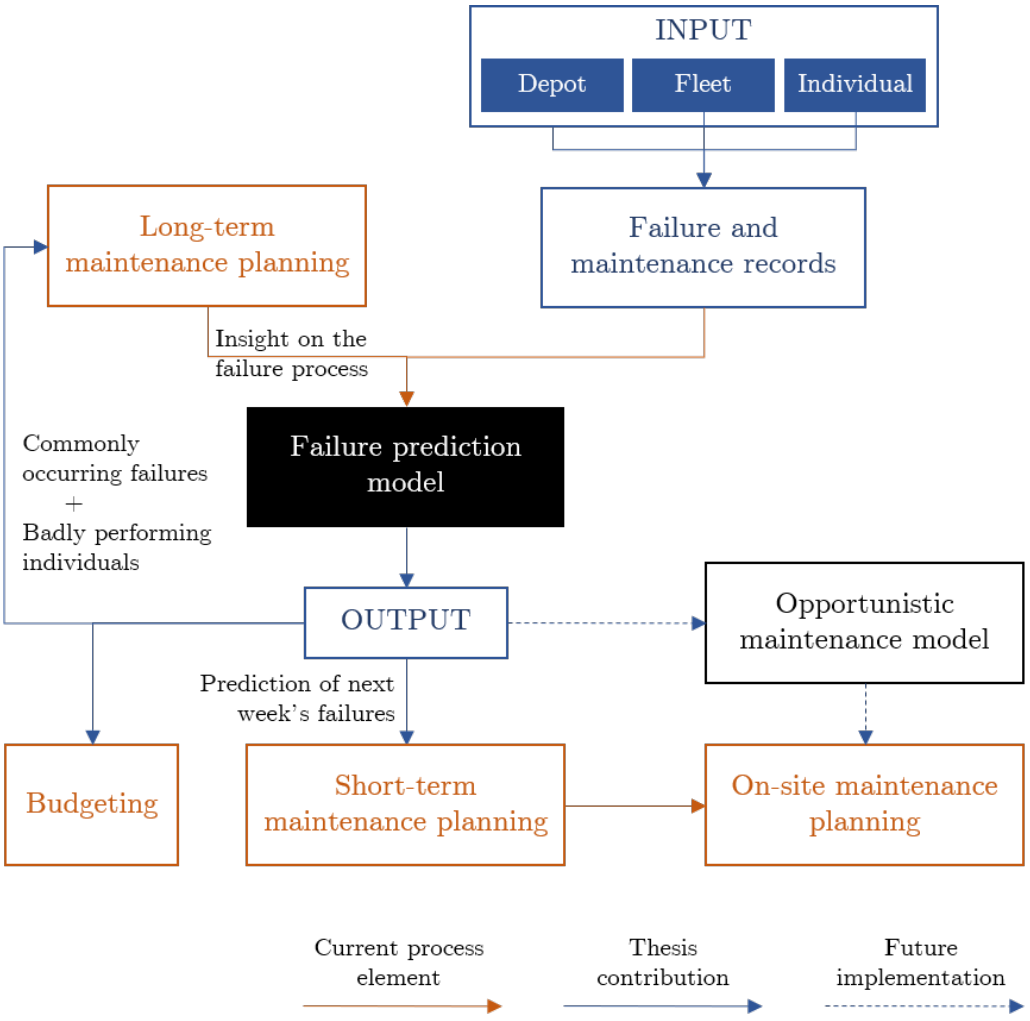


Figure 2.1: A graphical presentation of how the failure prediction model fits into the maintenance planning process.

Chapter 3

Modelling Failure Repair Needs

The total upcoming maintenance actions are the combination of preventive and corrective maintenance actions. Hence, to estimate the total maintenance work load both maintenance types must be considered (Duffuaa et al., 2015). In this thesis we only focus on modelling the uncertain part, namely the need for corrective maintenance. Thus, the emphasis of this chapter lies on failure modelling.

Yang et al. (2012) define a repairable system as one which generally is repaired at failure instead of being replaced. A rolling stock falls into the category of repairable systems, more specifically into the category of complex repairable systems. A systems is classified as complex if it consists of several components or subsystems, between which there are either economic, failure or structural dependence (Wang, 2002). Hence, the focus of this literary review is failure modelling of complex repairable systems.

Reliability refers to the ability of a system to function for a specified time period and reliability analysis deals with the modelling of a system's condition and failure risk. Numerous methods for modelling failures have been presented in the reliability analysis literature and this chapter provides a summary of different approaches. Figure 3.1 presents a scheme of the failure prediction models and methods introduced in this chapter and these are discussed further in Sections 3.1 and 3.2. Furthermore, Section 3.3 presents a few case studies relevant to VR's application.

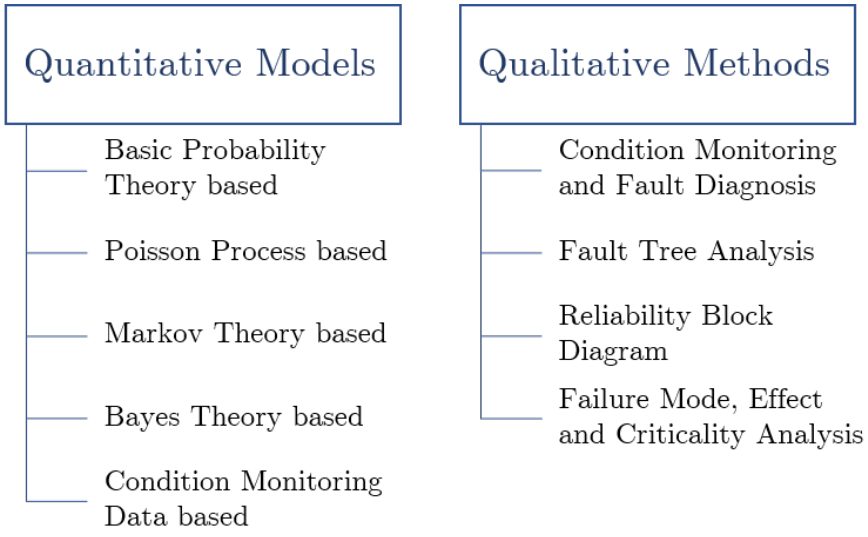


Figure 3.1: An overview of the methods for failure modelling presented in this chapter. The split is based on the literary review in the PhD thesis of Sun (2006).

3.1 Quantitative Models

A mathematical approach to failure modelling is popular in the reliability analysis of a system. Several quantitative failure models have been developed and are based on probability theory, statistics and stochastic processes. This section presents different quantitative approaches and also introduces the basic mathematical concepts related to reliability analysis.

3.1.1 Basic Principles of Probability Theory

The survival and failure distribution functions and the hazard function are concepts relevant for reliability analysis (Kalbfleisch and Prentice, 2002). In this section we present the relationships between these relationships of the probability distribution of the failure time T, as presented by Kalbfleisch and Prentice (2002) and Kaufmann et al. (1977). To illustrate the concepts, we present the survival, failure probability distribution, and hazard function for the Weibull distribution. The Weibull distribution is a commonly used distribution for modelling the reliability of a system.

The survival function $S(t)$ is a mathematical presentation of the expected reliability distribution of the system. The survival function gives the proba-

bility of a system surviving in the time interval (0,t], where $0 < t < \infty$, that is, the probability of the does not fail before time T

$$S(t) = \Pr(T \geq t).$$

The variable t does not necessarily correspond to time, but is rather a system specific variable best describing the use of the system and the change in the condition of the system. It can be, for example, operating time, number of production cycles or kilometres. Figure 3.2 presents three survival functions with different underlying survival distributions. The cumulative failure distribution function $F(t)$ is closely related to the survival function

$$F(t) = 1 - S(t) = \Pr(T < t).$$

The cumulative failure distribution function presents the probability of failure during a specific time interval (0,t), $0 < t < \infty$.

The failure probability density function $f(t)$ again presents the probability of failure at time t, hence with small dt corresponds to

$$f(t)dt \approx \Pr(t \leq T < t + dt) = F(t + dt) - F(t) = S(t) - S(t + dt),$$

as $f(t)$ is the derivative of the cumulative failure distribution function $F(t)$

$$F(t) = \int_0^t f(\tau)d\tau.$$

We also note that $f(t) \geq 0$, $\int_0^\infty f(t)dt = 1$ and $f(t)$ is related to $S(t)$ as follows

$$S(t) = 1 - \int_0^t f(\tau)d\tau = \int_t^\infty f(\tau)d\tau.$$

The failure rate $\lambda(t)$, also called the hazard rate or instantaneous failure rate, presents the instantaneous failure rate for a system which has survived until time t, i.e.

$$\lambda(t) = \lim_{dt \to 0^+} \frac{\Pr(t \leq T < t + dt | T \geq t)}{dt},$$

and the relationship to the survival function and the failure probability density function is

$$\lambda(t) = \frac{f(t)}{S(t)} = -\frac{\frac{dS(t)}{dt}}{S(t)}.$$

We can further write the hazard rate in the form

$$\lambda(t) = -\frac{d}{dt} \log S(t).$$

When we integrate with respect to t and use $S(0) = 1$, we get

$$S(t) = \exp\Big[-\int_0^t \lambda(\tau)d\tau\Big] = \exp[-\Lambda(t)], \tag{3.1}$$

where $\Lambda(t) = \int_0^t \lambda(\tau)d\tau$ is the cumulative hazard function. When we differentiate equation (3.1), we get the relationship between the failure probability density function and the hazard rate

$$f(t) = \lambda(t)\exp[-\Lambda(t)].$$

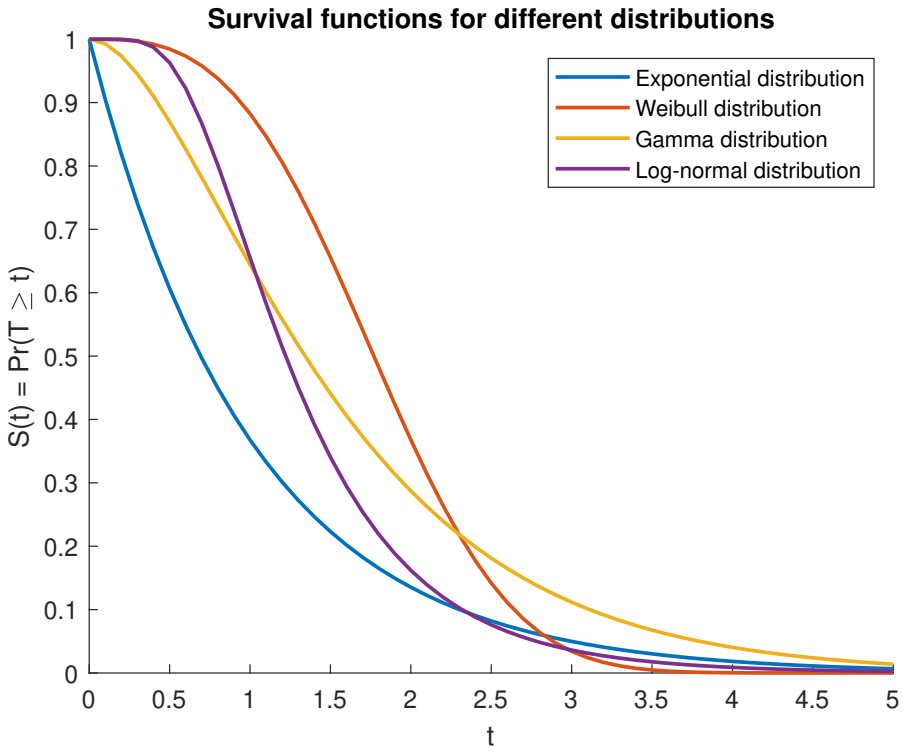


Figure 3.2: Examples of survival functions when t follows different failure distributions.

The Weibull distribution is one of the most commonly used distributions for reliability modelling due to the possibility to include a change in the failure rate over time. Other popular distributions for failure modelling are the exponential distribution, the Poisson distribution, the gamma distribution and the log-normal distribution. The failure distributions are discussed by, e.g., Kalbfleisch and Prentice (2002) and Kaufmann et al. (1977). We here present

the two-parameter Weibull distribution, the survival function of which has the form

$$S(t) = \exp[-(\lambda t)^{\gamma}], \tag{3.2}$$

where $\lambda > 0$ is the scale parameter and $\gamma > 0$ is the shape parameter, also referred to as the Weibull slope. Taking the derivative of the survival function yields the probability density function, which is

$$f(t) = \lambda\gamma(\lambda t)^{\gamma-1} \exp[-(\lambda t)^{\gamma}]. \tag{3.3}$$

The hazard function again has the form

$$\lambda(t) = \lambda\gamma(\lambda t)^{\gamma-1}. \tag{3.4}$$

To illustrate the effect of the Weibull shape parameter γ Figure 3.3 presents the probability density function for three different values for γ, while keeping scale parameter λ constant. The shape parameter also has a distinct effect on the failure rate, which is presented later in Figure 3.6. This is an important effect of γ in the Weibull distribution, as it indicates a decreasing ($\gamma < 1$), increasing ($\gamma > 1$) or constant ($\gamma = 0$) hazard rate.

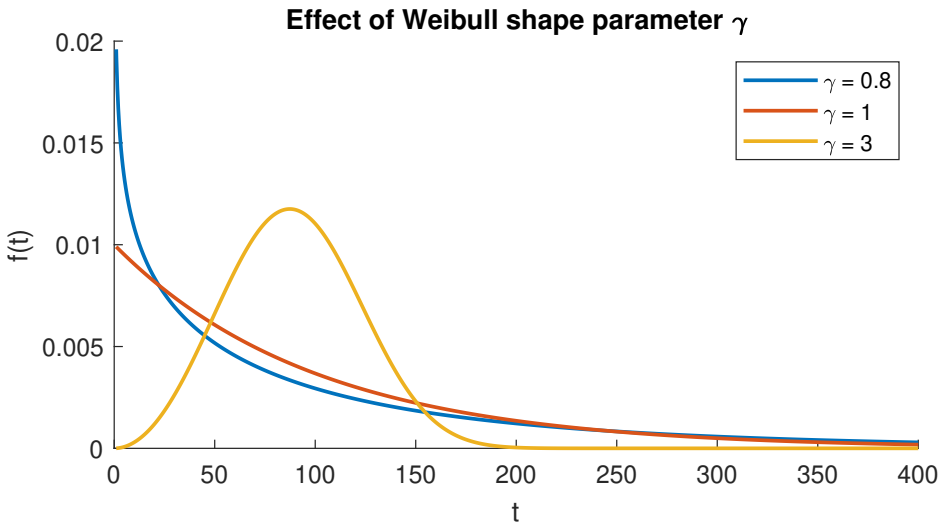


Figure 3.3: The effect of the Weibull shape parameter γ on the probability density function $f(t)$. The scale parameter is kept constant at $\lambda = 1/100$.

The scale parameter has a stretching effect on the Weibull probability density function, illustrated in Figure 3.4 as the probability density function is presented for different values on λ, while keeping γ constant. The peak of the probability density function is lower and appears at a larger value t for

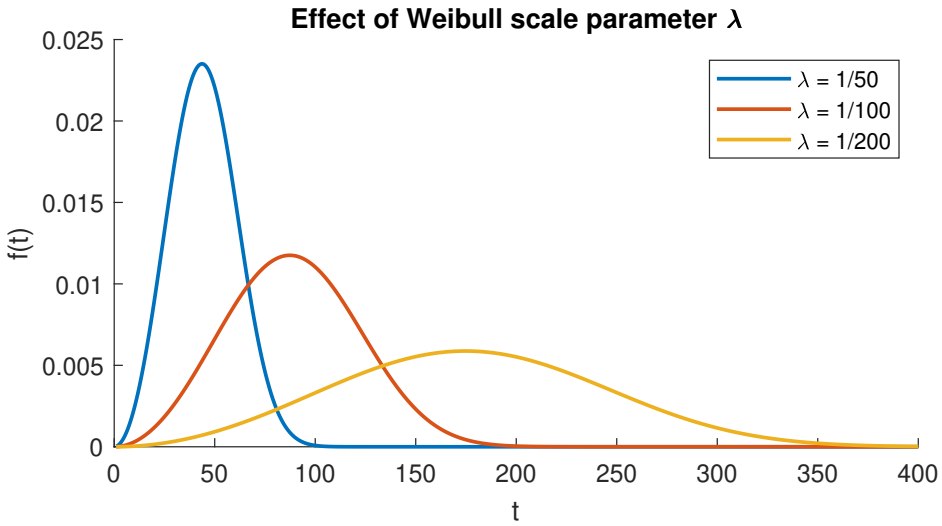


Figure 3.4: The effect of the Weibull scale parameter λ on the probability density function $f(t)$. The shape parameter is kept constant at $\gamma = 3$.

a smaller value on scale parameter λ. If λ is larger, the peak is higher and occurs at a earlier t-value.

Historical failure data can be used to determine a suitable distribution for modelling of failure times. Plotting the empirical distribution function, which is defined as

$$\hat{F}_n(x) = \frac{\text{no. sample values} < x}{n}$$

where n is the sample size, provide insight to what distribution the sample failure times could fit to (Kalbfleisch and Prentice, 2002). Most programming tools, such as Matlab, Python, or R, have functions to fit the failure times to different distributions. After determining the distributions, they can be used to estimate future maintenance needs, for example by examining failure probabilities at different times or by simulating failure times from the distributions with Monte Carlo simulation (Le Gat and Eisenbeis, 2000).

3.1.2 Models based on the Poisson Process

Lindqvist (2006) and Doyen and Gaudoin (2004) present the renewal process (RP), commonly used for modelling failures for repairable systems. RP includes both the homogeneous Poisson process (HPP) and the non-homogeneous Poisson process (NHPP). Poisson processes assume that the failure times are independent of each other and that they all follow the same

distribution. HPP assumes perfect repair, whereas NHPP assumes minimal repair. The difference can be seen from the failure intensity functions.

The HPP failure intensity function at time $t \geq T_{N_t}$ is of the form

$$\lambda_t = \lambda(t - T_{N_t}),$$

where T_{N_t} is the N:th failure time. Thus, $t - T_{N_t}$ $(t \geq T_{N_t})$ indicates that the intensity function starts at zero after each failure and does not consider earlier failures. The intensity function returning to zero after each failure captures the assumption of perfect repair, as it indicates that the system is returned to an AGAN state after repair. HPP can also be used to model a non-repairable system's failure process, where the system is replaced at failure.

As perfect repair often means replacement, HPP may be unsuitable for the modelling of repairable systems. Instead, NHPP can be used, where the assumption is minimal repair, returning the system to an ABAO state after repair. NHPP considers the cumulative lifetime of the system, not only the time since the previous failure. Hence, the failure intensity function is

$$\lambda_t = \lambda(t), \quad \forall\, t \geq 0.$$

As Doyen and Gaudoin (2004) discuss, NHPP allows for failure modelling of systems that improve or degrade over time, as the failure rate is time dependent. If the failure rate is constant, the failure process equals a HPP.

Figure 3.5 illustrates the failure probability density functions of a HPP and a NHPP. Both processes follow the same Weibull distribution. In the HPP the failure distribution is restored to the original state after each failure, whereas the failures in the NHPP do not change the distribution.

Doyen and Gaudoin (2004) describe the power law process (PLP), which is a popular type of NHPP with the failure intensity function

$$\lambda(t) = \lambda\gamma t^{\gamma-1},$$

where scale parameter $\lambda > 0$ and shape parameter $\gamma > 0$. An intensity function with $\gamma < 1$ represents an improving system, as this implies a decreasing failure intensity function. With $\gamma > 1$, the function is increasing, hence representing a deteriorating system with an increasing failure rate over time. We can see that with $\gamma = 1$, the intensity function becomes constant and as it is no longer time-dependent, reverting back to a HPP. The change in the PLP's failure rate for different γ values can be presented in a so called bathtub curve, illustrated in Figure 3.6.

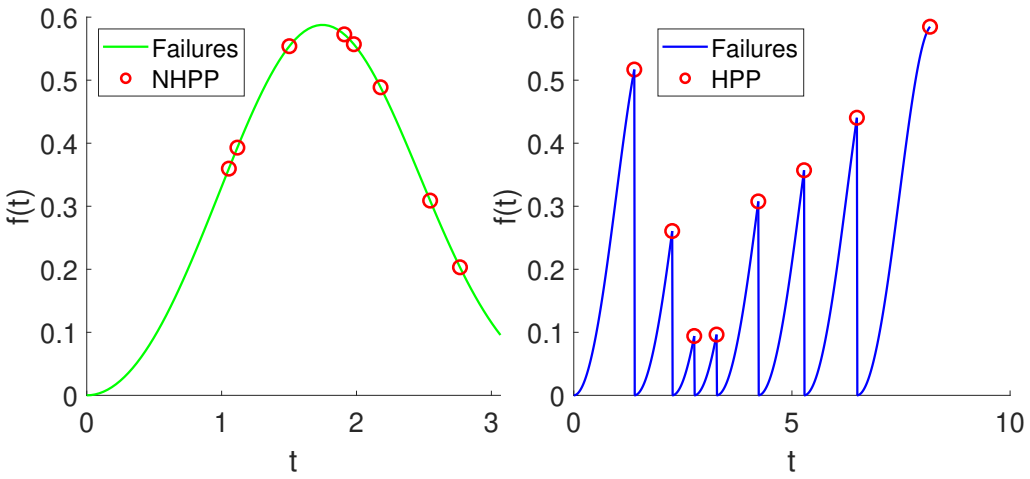


Figure 3.5: Failure probability density functions of a NHPP (left) and a HPP (right), with random inter-failure times simulated from a Weibull distribution.

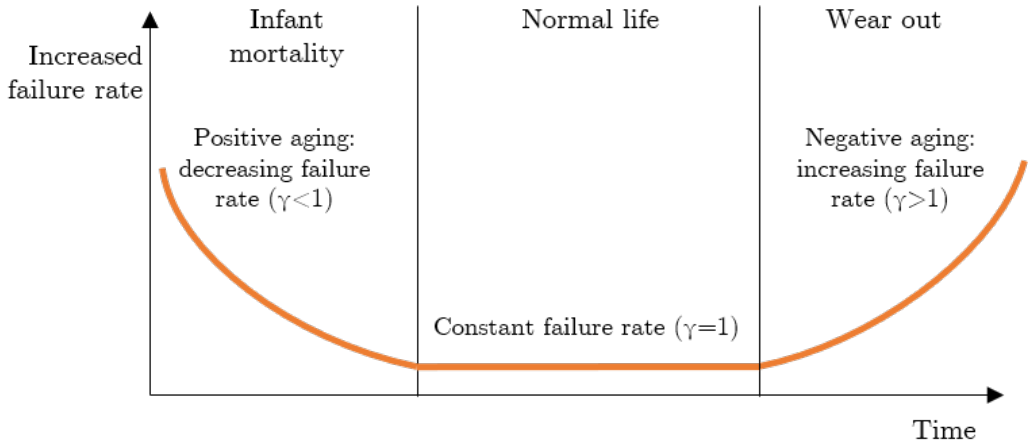


Figure 3.6: The so called bathtub curve presents how different values on the shape parameter γ affect the failure rate and hence the type of failure process. (Ahmad and Kamaruddin, 2012)

HPP and NHPP are basic models, with limited use for modelling failures in real systems due to these models' assumption of the system either being restored to an AGAN state or remaining at an ABAO state after repair. In reality however, the system is usually imperfectly repaired, wherefore the state is between these two after repair. A basic, widely referred method for imperfect repair modelling was developed by Brown and Proschan (1983).

The model assumes a perfect repair with probability p and minimal repair with a probability $1 - p$ at failure. Several generalizations of this classical model have later been developed. Another class of imperfect repair models are the virtual age models (Kijima, 1989). The virtual age models reduce the age of the system after repair by some factor. Hence, the failure intensity function value after repair at time T is not $\lambda(T)$, but rather $\lambda(T - x)$, where $0 \leq x \leq T$. For a review of imperfect repair models, see Peng et al. (2018). Figure 3.7 illustrates an imperfect repair process following the basic imperfect repair model of Brown and Proschan (1983). The underlying distribution is the same Weibull distribution as in Figure 3.5.

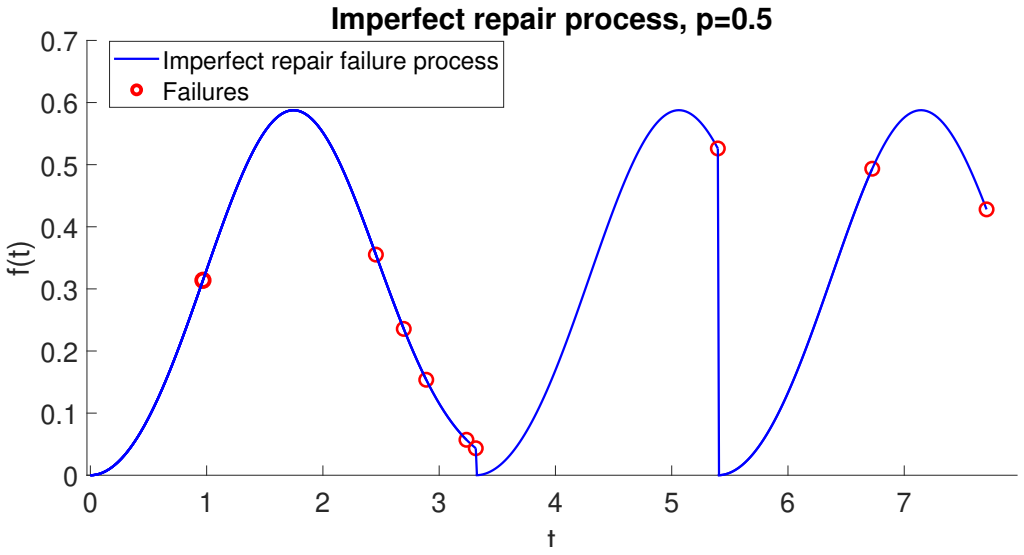


Figure 3.7: Failure process and failure probability density function of a failure process following the basic imperfect repair model presented by Brown and Proschan (1983), with probability of perfect repair $p = 0.5$. The underlying failure distribution is the same Weibull distribution as in Figure 3.5.

3.1.3 Models based on Markovian Theory

Buzacott (1970) presents how a Markov approach can be used for modelling of failure times for repairable systems. The Markov process is characterised by being non-hereditary and memoryless and it captures the stochastic process of a system transitioning from one state to another, so that the future probability behaviour only depends on its current state.

The system is assumed to have N possible states and $P_i(t)$ describing the probability of the system being in state i at time t. The states can correspond

to functions, failure modes, standby and different maintenance activities of the system, and the holding times in each state are exponentially distributed.

However, the transition probabilities can be difficult to estimate and the number of transition probabilities grow exponentially as the size of the system grows. Hence, the Markov approach is best suited for smaller systems.

3.1.4 Models based on Bayesian Theory

The Bayes' theorem, introduced by Thomas Bayes in the 18^{th} century, presents the probability of an event based on some prior conditional information relevant to the event. The mathematical presentation of the theorem is

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)},$$

where $P(A)$ and $P(B)$ refer to the probabilities of event A and B occurring, respectively. $P(A|B)$ is the probability of event A given that event B occurs, and $P(B|A)$ is the probability of event B given A. (Bayes, 1763)

In failure modelling, Bayesian models allow for the use of expert knowledge on the system's reliability to reduce uncertainties in the model. Observed values, such as, thickness of brake pads, amount of precipitation, or type of previous failure can be used to estimate the failure probabilities. Beiser and Rigdon (1997) model a failure process with HPP and PLP and use a Bayesian approach to incorporate prior knowledge affecting the failure process. The model is based on historical data and is used to estimate the number of failures during some given future time interval.

3.1.5 Models based on Condition Monitoring Data

Monitoring the current state of a system gives important insight into pending failures. There are several models for utilising the condition information in failure modelling, here we present the proportional hazard model (PHM), also called Cox's model, and the accelerated failure time model (AFTM). These are classified as survival regression models and also other parameters than the system condition can be included as parameters in the model. Hence, these models allow for including the effect of parameters affecting the failure process, such as weather condition, age or load. The models are presented by among others Kalbfleisch and Prentice (2002).

PHM is an extension to the homogeneous and non-homogeneous Poisson processes (HPP and NHPP), where explanatory variables affecting the failure

rate are added. As p explanatory variables (covariates) $\mathbf{x}' = (x_1, x_2, ..., x_p)$ are linked to the failure time T, the hazard function will be of the form

$$\lambda(t|\mathbf{x}) = \lambda_0(t) \exp(\mathbf{x}'\beta),$$

where $\lambda_0(\cdot)$ is an arbitrary baseline hazard function, $\mathbf{x}$ is a vector of the covariates, corresponding to, e.g. observed state of the system, the temperature, or other factors affecting the failure process, and $\beta' = (\beta_1, \beta_2, ..., \beta_p)$ is a vector of unknown regression parameters. In PHM, the covariates have an multiplicative effect on the hazard function. If we have $\lambda_0(t) = \lambda$, PHM reduces to an exponential proportional hazard function, as the baseline hazard function is constant. We get Weibull PHM if we have $\lambda_0(t) = \lambda\gamma(\lambda t)^{\gamma-1}$. In the case of an arbitrary baseline hazard function, the model is suitable for many applications, due to its flexibility.

The conditional survival function of a PHM is of the form

$$S(t|\mathbf{x}) = [S_0(t)]^{\exp(\mathbf{x}'\beta)},$$

and the probability density function is

$$f(t|\mathbf{x}) = \lambda_0(t) \exp(\mathbf{x}'\beta) \exp\big[- \exp(\mathbf{x}'\beta) \int_0^t \lambda_0(\tau) d\tau \big].$$

Whereas the covariates in the PHM act multiplicatively on the hazard function, the accelerated failure time model (AFTM) models the covariates' effect as an acceleration or deceleration of the time to failure. The hazard function of the AFTM is of the form

$$\lambda(t|\mathbf{x}) = \exp(-\mathbf{x}'\beta)\lambda_0(te^{-\mathbf{Z}_j\beta}). \tag{3.5}$$

This leads to a survival function

$$\begin{aligned} S(t|\mathbf{x}) &= \exp\big[- \int_0^t \exp(-\mathbf{x}'\beta)\lambda_0(\tau e^{-\mathbf{x}'\beta}) d\tau \big] \\ &= \exp\big[- \Lambda_0(te^{-\mathbf{x}'\beta}) \big]. \end{aligned}$$

The failure probability density function is the product of the hazard function and the survival function.

We can see the covariates' effect on the failure process in equation (3.5). The covariates affect the failure time rather than the hazard function as in the case of PHM. If we have covariates $\mathbf{x} = 0$, we are left with the baseline hazard function $\lambda_0(t)$. The regression parameters β specify whether the covariates accelerate or decelerate the failure time. If $\beta_j < 0$, covariate x_i has an accelerating effect on the failure time, whereas $\beta_i > 0$ indicates a decelerating effect of covariate x_i.

3.2 Qualitative Methods

A quantitative model is not always possible to apply for the modelling of failures. In this section, we present commonly used qualitative methods for reliability modelling, some of which can be used for quantitative analysis as well.

3.2.1 Condition Monitoring and Fault Diagnostics

The condition based maintenance strategy was presented in Section 2.1. Duffuaa et al. (2015) discuss the characteristics of condition based maintenance, which is an effective preventive maintenance strategy, as the current condition of a system is considered when maintenance decisions are made. This reduces the number of unnecessary maintenance actions. The condition of a system's components can be monitored continuously and maintenance actions are taken when the condition of a component falls below a predetermined threshold. Also, the condition can be inspected in conjunction with another maintenance action and the component can be opportunistically maintained if the condition falls below the threshold. The threshold is determined based on data from previous failures. Criticality of failures should also be assessed to determine the threshold for different maintenance actions. The failure modes, effect and criticality analysis (FMECA) presented in Section 3.2.4 can be used for determining failure criticality.

Sensors can be used for monitoring the condition of components by measuring, for example, vibrations, noise level, lubricating oil contaminants or running temperature. Condition based maintenance has become more popular as the technical development allows for more cost effective and reliable monitoring. This method, when implemented correctly, gives great insight into pending failures. A good understanding between the condition of components and failures is needed, which requires historical data. Condition based maintenance is especially suitable for systems where the change in system condition can be observed before failure, i.e. when the condition of the system can indicate imminent failures. However, as concluded in Section 2.1, extensive implementation of condition-based maintenance has not yet been incorporated in railway operations. Currently this strategy is limited to maintaining only specific component, such as brake pads.

3.2.2 Fault Tree Analysis

A classical method for analysing failures is fault tree analysis (FTA) (Watson et al., 1961). This is a method with a top-down approach, where a top event (failure) is analysed by graphically and logically presenting different combinations of possible events affecting the system, all leading up to this top event. After defining the top event to be analysed, events that are immediate, sufficient and necessary causes for the top event are identified and connected to the top event by logical operators.

An example of a simple fault tree is presented in Figure 3.8. The logical operators AND and OR, together with how the components are connected, describe possible paths to the top event. In the presented fault tree, examples of cases leading to the top event are: component A fails, or components D and E fail, or component H fails.

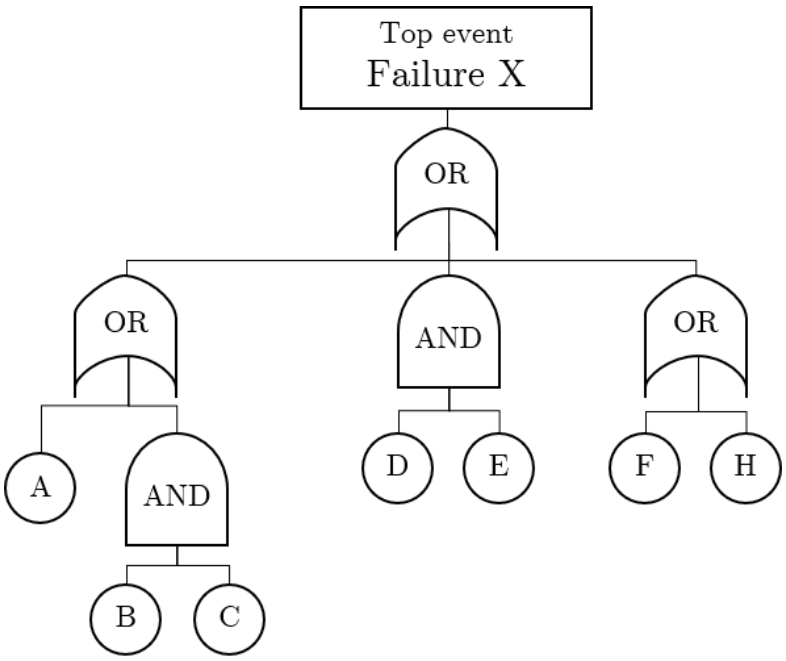


Figure 3.8: An example of a fault tree. In the tree, A-H are components whereas AND and OR are logical operators, connecting the components to the top event.

FTA can be used for quantitative analysis to provide information on, for example root causes, failure paths, and weak areas of the system. It can also be used for qualitative analysis, yielding information on the probability of the top event happening. Hence, probabilities for the top event in different

scenarios can be studied with a fault tree. Mancuso et al. (2017) use fault tree analysis to derive combinations of events leading to system failure, which they utilise for further analyses on system risk.

The fault or event tree analysis technique is however time consuming as numerous events must be considered. Another limitation is that the fault tree cannot describe the behaviour of a system with multiple states, for example a leakage of different gravity (Mancuso et al., 2017). It can also be difficult to find enough relevant data for the analysis to be reliable.

Traditional FTA is based on static fault trees, which is a limiting factor in the utilisation of fault trees for analysing system failure, since component failure may be affected by the state of other components. Dynamic fault trees can model such failure sequences, as they include additional logical operators that capture the relationships between component failures. (Distefano and Puliafito, 2007)

3.2.3 Reliability Block Diagram

Whereas FTA visualises the combination of component failures leading to system failure, a reliability block diagram (RBD) presents the reliability relationships and functions between components. Reliability relationships refer to how the reliability of one component affects the reliability of others. Another factor differentiating RBD from FTA is that RBD is success oriented, while FTA models the system failure. (Distefano and Puliafito, 2007)

RBD was the first model developed for reliability assessment. An RBD is a logical network showing the connections of functioning components needed for a specific system function to work. An example of a RBD is presented in Figure 3.9. In the presented RBD, component C_1 must function for components $C_{2.1}$, $C_{2.2}$ and $C_{2.3}$ to function. Similarly, for component C_3 to function, $C_{2.1}$, $C_{2.2}$, $C_{2.3}$ as well as C_4 need to be working. λ_i refers to the reliability measure for component i, illustrating the probability of the component to be in a working condition. If the λ:s are known for all components, the RBD can be used for quantitative analysis.

If a system has several functions, multiple reliability block diagrams might be used. The qualitative information the RBD provides, is information on the system's state (functioning or failed) under given conditions. When RBD is used for quantitative analysis, the system's reliability at a given time t is measured. A reliability block diagram can be converted to a fault tree and vice versa, but the fault tree is generally more suitable for quantitative

analysis. (De Souza, 2012)

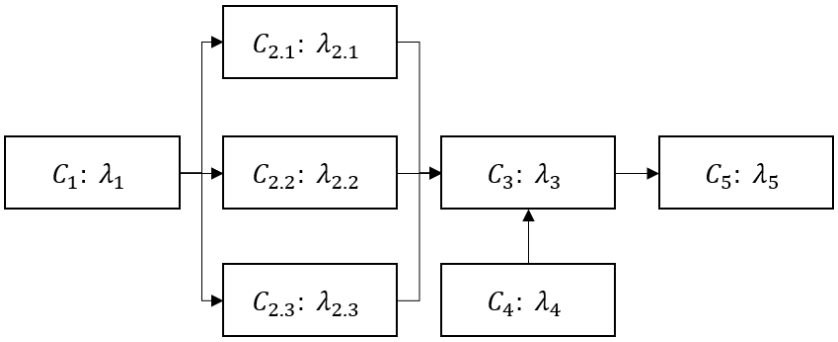


Figure 3.9: An example of a reliability block diagram, illustrating the reliability relationships between components. λ_i is the reliability measure for component C_i.

3.2.4 Failure Modes, Effect and Criticality Analysis

Bowles and Peláez (1995) describe the failure modes, effect and criticality analysis (FMECA), the goal of which is to quantify and rank critical failures, for prioritisation of corrective actions. The approach was, like FTA, developed in the aerospace industry for design reliability assessment. FMECA is often conducted by a group of experts identifying for each of the system's component or subsystem

- the failure modes,
- the consequences of the failure modes,
- and the criticality of each failure.

The criticality assessment is done by evaluating the gravity of each failure's negative effects on the function and operation of the system, on other components, on the environment and on people. After the evaluation of all identified failure modes, these are typically ranked according to probability of failure and severity of the effect. The highest ranked failure modes are then given higher priority, as they are assumed to be more important. FMECA helps to identify critical parts of the system and it's design, providing an opportunity to improve the reliability of those parts. Identifying the failure modes, their consequences and the criticality of the failures also allows for determining the most suitable maintenance policy and serves as a basis for determining

the priority of maintenance tasks. However, FMECA analysis is difficult to implement even for a rather simple system, due to the exponential growth of possible failure modes as the system's complexity grows.

3.3 Examples of case studies

In this section, we present some case studies where the approaches presented earlier in this chapter are used for modelling and analysing failures. The cases are related to rolling stock, aviation industry and water networks.

Cheng and Tsao (2010) present an approach for first determining a suitable strategy for rolling stock maintenance and then estimating the needs for spare parts. In their calculations, they assume the failure time to follow a Weibull distribution, the parameters of which are determined based on historical failure data. The failure times generated from the Weibull distribution are used for estimating component replacement intervals. They also discuss how FMECA and FTA can be utilised for determining the most critical components of rolling stock operations. This is later done by Dinmohammadi et al. (2016), as they introduce a FMECA based approach to evaluate the risks associated with unexpected failures of components of a rolling stock. The failures are ranked according to likelihood of occurrence and the severity of damage. For the failure ranked as most critical, mitigating actions can be identified. The authors apply their approach on passenger door units of rolling stock operating in Scotland. They define potential failure modes and their root causes and conclude that 12% of the failure modes are of high criticality to the functionality of the door system. They utilise their findings for developing a preventive maintenance programme.

Reliability analyses and predictions are an important part of the aviation operation's research, which is also indicated by the several reliability models originally developed for the aviation industry. In a case study on jet engine life modelling, Weckmann et al. (2001) apply the power law process (PLP). They use jet engine removal data from two airline databases for the construction of the model, which they use to forecast future engine removal instances. They also include a seasonal effect in their forecasting model by adjusting the estimated jet engine removal times according to a month specific adjustment factor. The model is evaluated based on data fit and accuracy of the forecast. They conclude that even if data from a longer time period (around 15 years) would be needed for the model to be more accurate, PLP is an effective forecasting tool providing a more accurate forecast compared to

other used forecasting techniques. A PLP forecasting model able to predict with an accuracy of 5% could provide substantial cost savings to an airline's maintenance expenses.

Water network failure prediction is addressed in several papers. In his PhD thesis Røstum (2000) presents a non-homogeneous Poisson process (NHPP) with covariates and a modified Weibull proportional hazard model (PHM) to predict the pipe failures for each individual pipe in a water distribution network. The models are based on data on the water distribution network in Trondheim, Norway. Covariates which he found significant for the Weibull PHM were pipe length, age and dimension, soil condition, and the number of previous failures. He concludes that both models are capable of modelling failures and, as the model results are more accurate at network level, the pipe level results are also satisfactory. He observes that the Weibull PHM tends to overestimate failures compared to NHPP. The author argues that predictive models should become part of water network maintenance decision processes, as such tools would lead to a reduction of maintenance expenses. The actual implementation of a prediction model in the process is however not described.

In their model for water network failure forecasting, Le Gat and Eisenbeis (2000) use a Weibull PHM and Monte Carlo simulation to estimate the expected number of failures for each pipe in the network. The prediction model is based on data from the water company in Charente-Maritime, France. They include both environmental covariates, such as traffic, humidity, and acidity, as well as internal factors, such as pipe diameter, length, and age. They conclude that the Weibull PHM provides an efficient method even with shorter maintenance records. Without specifying the type of prediction model, the authors conclude that the use of a model predicting pipe condition deterioration would enhance the water network maintenance decision making process and would support the analyses of, for example water quality and water loss.

Further applications of mathematical models in maintenance are presented by Scarf (1997). He presents different approaches to failure modelling as part of modelling maintenance. He also underlines the system specificity characteristic of maintenance modelling, namely how the focus should lie on the understanding of the system and decision-maker's interests rather than on the model used to reach the goal or solution.

Chapter 4

Presentation of the Failure Data

This chapter provides background to the failure and maintenance records at VR. Here we also present an analysis of the failure times from a two-year period for one rolling stock fleet to get an understanding of the failure process. This data is also used for developing the prediction model described in the next chapter. Analysis is conducted separately for the locomotive fleets and multiple unit fleets presented in Table 4.1 before developing the system specific failure prediction models. However, we only present the results for one fleet in this thesis to illustrate the procedure.

4.1 Failure and Maintenance Records

Fleet sizes, i.e. the number of rolling stock individuals in the fleet and nick names of VR's fleets are presented in Table 4.1. Types SRx are electric locomotive classes, DV12 and DR14 are diesel locomotive classes and DR16 refers to a class of diesel-electric locomotives. The SMx classes are electric multiple unit train fleets, where the rolling stock consists of a combination of carriages with electric motors incorporated in one or several carriages, hence not needing a separate locomotive. Fleets SM2, SM4 and SM5 operate the local traffic in the Helsinki region. SM3 is a Pendolino fleet operating on domestic routes, whereas the Pendolino fleet SM6, called Allegro, operates the route between Helsinki and Saint Petersburg. The DM12 fleet consists of single unit diesel rail cars, mainly operating on secondary routes. Figure 4.1 presents an electric locomotive of type SR3 and Figure 4.2 presents a diesel locomotive of type DV12, the largest locomotive fleet. A SM3 type Pendolino is presented in Figure 4.3. In addition to the fleets listed in Table 4.1, VR

also has carriage fleets, but the failure prediction model is not developed for these and hence are not presented here.

Table 4.1: Locomotive and multiple unit fleets for which failure prediction models are developed.

Type	**Name**	**Nick name**	**Fleet size**
Locomotive	SR1	Siperian susi	109
Locomotive	SR2	Marsu	46
Locomotive	SR3		18
Locomotive	DV12	Deeveri	192
Locomotive	DR14	Seepra	24
Locomotive	DR16	Iso-Vaalee	23
Multiple unit	SM2	Sami	50
Multiple unit	SM3	Pendolino	18
Multiple unit	SM4	Pupu	30
Multiple unit	SM5	Flirt	81
Multiple unit	SM6	Allegro	4
Multiple unit	DM12	Lättä	16

Figure 4.1: An electric locomotive of type SR3, VR's newest locomotive fleet. (VR Group, 2019b)

Figure 4.2: A diesel locomotive of type DV12, VR's largest locomotive fleet. (VR Group, 2019c)

Figure 4.3: An electronic multiple unit train of type SM3, operating domestic routes. (VR Group, 2019a)

The data consists of failure and maintenance records with the information presented in Table 4.2. The model and analyses are based on the presented data entries.

Table 4.2: The information in the maintenance and failure records used for the analysis.

Data	**Explanation**
Equipment ID	ID of the rolling stock.
Fleet name	The name of the fleet the rolling stock belongs to (names presented in Table 4.1).
Fault code	Code with 18 categories, categorising the failures on a high level according to the location or system of the rolling stock where the failure occurs.
Fault creation time	Time stamp when the failure report is made, in this thesis used as failure time.
Repair start time	Time stamp when the fault is taken under repair.
Fault completion time	Time stamp when the repair of the fault is finished.
Maintenance type	Categorisation of preventive maintenance actions.
Maintenance start time	Time stamp when preventive maintenance is started.
Maintenance end time	Time stamp when preventive maintenance is completed.
Mileage values	Rolling stock total driven kilometres at different times. For the analyses, the kilometres is mapped to the failure times and repair completion times, so that the kilometre value is from the same day, or a maximum of five days before the time stamp.
Criticality	The criticality of the fault expressed on a scale of 1-3. 1 prohibits use before the fault is repaired, 2 means the fault must be repaired at the next depot stop and 3 does not prevent use.
Depot	The depot where the fault has been repaired.
Speed	The maximum speed limit for the train unit a rolling stock has been part of at any time, in addition to the operating time and km for the particular train unit.

The unit of measure best describing the change in the rolling stock condition is the driven kilometres. Hence, we must calculate the kilometres between the failures to illustrate the failure times. As most of the failures that occur are small and do not prevent the use of the rolling stock, the failures are not necessarily repaired at once. Therefore, additional failures in the same category might occur before the previous failures have been repaired. To make the inter-failure kilometres comparable, the calculation of the inter-

failure kilometres is different depending on when the failure occurs

- **The failure occurs before the previous failure is repaired:** The inter-failure kilometre is the difference between the kilometre mark at failure and the kilometre mark at the previous failure.
- **The failure occurs after the previous failure is repaired:** The inter-failure kilometre is the difference between the kilometre mark at failure and the kilometre mark at the repair of the previous failure.

An example of a failure process for a rolling stock individual for one failure category is presented in Figure 4.4. The inter-failure kilometres for failure 2 (F_2) equal the difference in driven kilometres between the occurrence of the first and the second failure, as the first failure (F_1) has not been repaired before the second failure arises. After the second failure, both F_1 and F_2 are repaired (marked as R_1 and R_2). The repair restores the stock to a condition with no failures. Thereafter, the third failure (F_3) occurs. The inter-failure kilometres for the third failure correspond to the difference in driven kilometres at the third failure and the repair of the second failure. The inter-failure kilometres for the fourth failure (F_4) are calculated similarly. The fifth failure (F_5) is found as the fourth failure is repaired, hence the inter-failure kilometres for the fifth failure are zero.

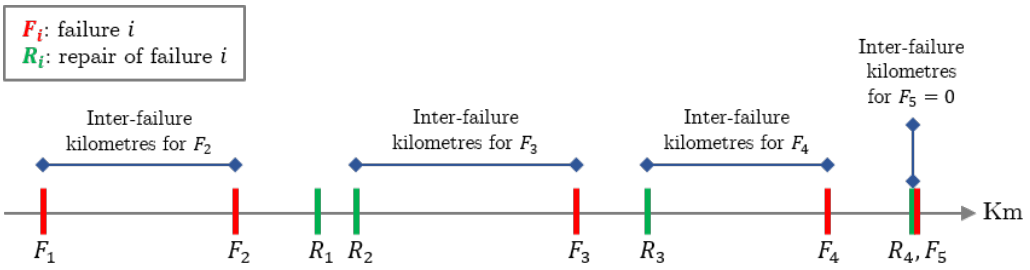


Figure 4.4: Example of a failure process for one failure category for a rolling stock individual.

There are however, some limitations associated with the data (Annala, 2018). The failures are not always correctly categorised and some actions associated with condition based preventive maintenance are systematically recorded as failures. Also, the reporting of a failure is not necessarily done when the failure occurs, but rather when it is noticed or even after it has already been fixed. This, in combination with the total driven kilometres not being marked daily, may lead to inaccuracies in the inter-failure kilometres.

The broad failure categories represent an additional limitation (Annala, 2018). The categorisation only categorises the failures at a higher level without specifying the exact type of failure or the failed component. This also affects the calculation of inter-failure kilometres, as the repair of a failure in a specific category might not affect the timing of the next failure in the same category at all. For example, say a failure occurs in category N, referring to doors and entrances. The failure is located in door number one of the rolling stock and it is repaired. The next day, another category N failure occurs - this time in door number two of the rolling stock. Repairing the failure in door number one does not necessarily affect the condition or probability of failure of door number two, hence leading to a misleading inter-failure kilometre for the second failure. The data at hand is limited to the described categorisation within most failure categories. There is some component specific failure history, which would allow for more accurate and descriptive inter-failure kilometres. However, this thesis focuses on the high level categorisations, as the low number of data points in the more specific failure categories hinder the development of a component specific prediction model.

Levo (2018) expresses that the repair lead times marked in the system also have some limitations. These are not necessarily consistent with the realised repair times, due to a number of reasons. Firstly, the workers may work on several failures simultaneously and mark them as completed at the same time. This results in the data indicating that X hours have been spent on repairing each of the failures, whereas in reality X hours have been spent on the repairs combined. Secondly, a failure ticket may be opened in conjunction with a preventive maintenance, but the failure is not repaired before the rolling stock leaves the depot. The ticket then stays open until the failure is repaired and marked as completed in the system, which can be several days, or even months later. The actual lead time for the repair has not been more than a few hours, but the data indicates it has taken several days. This however, is the case only for non-critical failures, which do not limit the rolling stock from operating. The marked repair times for failures prohibiting the rolling stock from being used are more accurate, even though they may include waiting times due to shortage in components.

4.2 Data Analysis

Next, we present an analysis of the failure data of a single fleet X. Figure 4.5 presents the number of failures per failure category for fleet X. The largest categories are F, R, E, J and Q; these failures combined correspond to 56%

of all failures during the two-year inspection period. The total number of failures during the two-year period is 16 024. These are mainly small failures not preventing the rolling stock from being used. Critical failures are only 1% of the total number of failures.

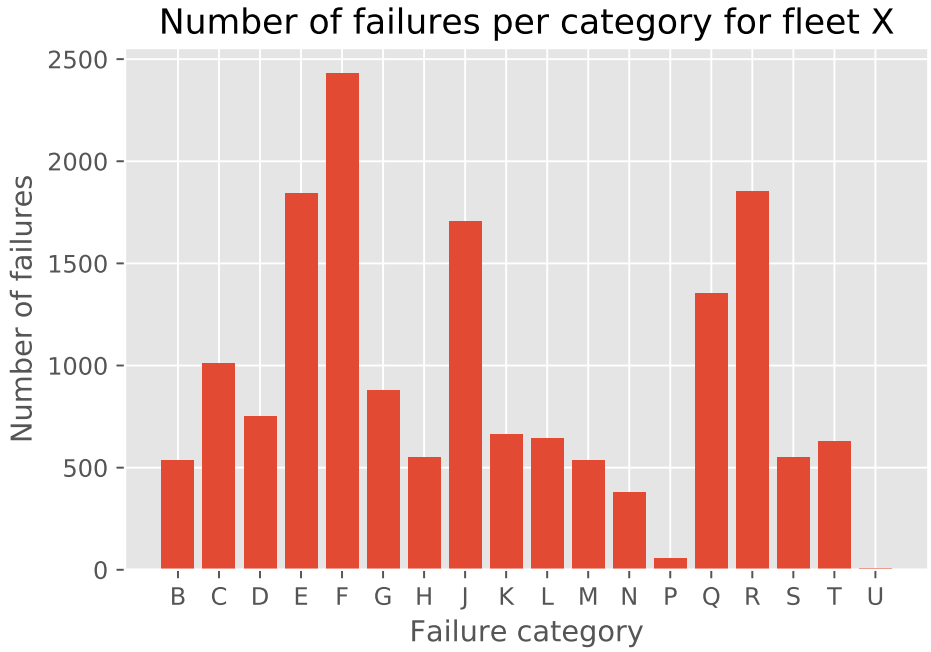


Figure 4.5: Number of failures per failure category for fleet X during a time period of two years.

A histogram of all inter-failure kilometres larger than zero are presented in Figure 4.6. For this fleet, approximately 9% of the failures are found during a depot stop, which results in the inter-failure kilometre being zero. The incorporation of these failures in the failure prediction model has to be done separately, the approach for which is described in Section 5.5. These have to be included separately in the model, since inter-failure kilometres equal to zero prevents the fitting of a Weibull or log-normal distribution to the inter-failure kilometres. We will use a Weibull distribution for the failure modelling, hence we now focus only on the inter-failure kilometres larger than zero. These are presented in the histogram in Figure 4.6, and correspond to 91% of all inter-failure kilometres.

The shape of the histogram presenting the inter-failure kilometres resembles an exponential probability distribution function. A Weibull or log-normal distribution could also fit the data. In Figure 4.7, we have fitted an exponen-

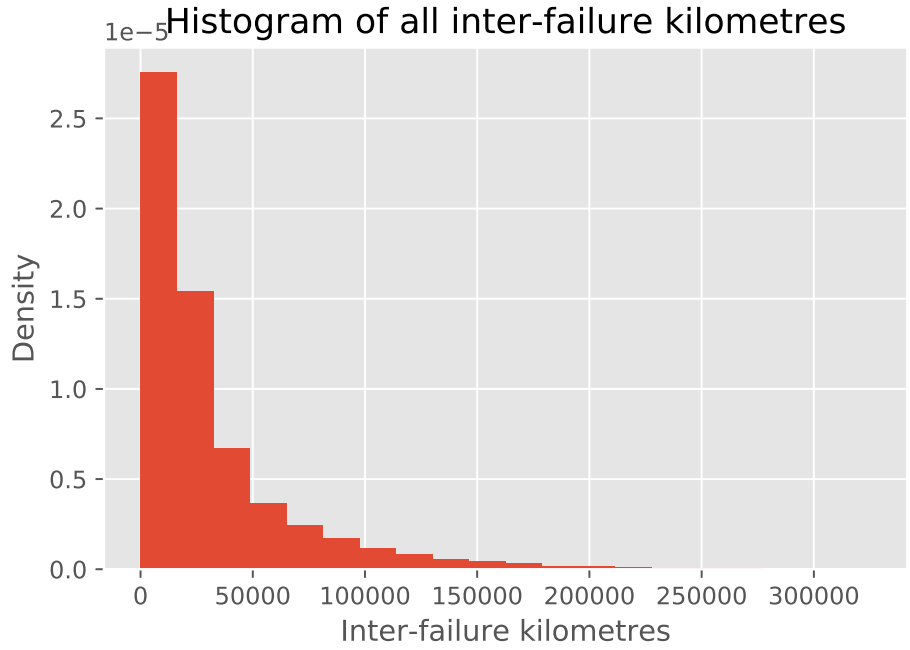


Figure 4.6: A histogram of all inter-failure kilometres larger than zero for fleet X, during the two-year period.

tial, Weibull and log-normal distribution to the inter-failure kilometres. We perform a Kolmogorov-Smirnov goodness-of-fit test (KS-test) (Massey Jr, 1951) to see whether the failure kilometres seem to fit the distributions or not. The test hypotheses are as follows:

- H_0 : The inter-failure kilometres follow the specified distribution.
- H_1 : The inter-failure kilometres do not follow the specified distribution.

We use a significance level $\alpha = 0.05$, indicating that the null hypothesis is rejected if the p-value falls below 0.05. Table 4.3 presents the results of the KS-tests. All p-values are below the 0.05 significance level, indicating that the inter-failure kilometres do not fit any of the distributions.

As we cannot fit one distribution to all inter-failure kilometres, we look at the inter-failure kilometres for specific failure categories. The histograms of the inter-failure kilometres for the the five largest failure categories (F, R, E, J and Q) with fitted exponential, Weibull and log-normal probability distribution functions are presented in Figure 4.8. We perform the Kolmogorov-Smirnov test to examine the fit of the distributions. The results are presented

in Table 4.4. With a significance level $\alpha = 0.05$ the KS-test null hypothesis of distribution fit is rejected for all but E and J failures coming from the Weibull distribution.

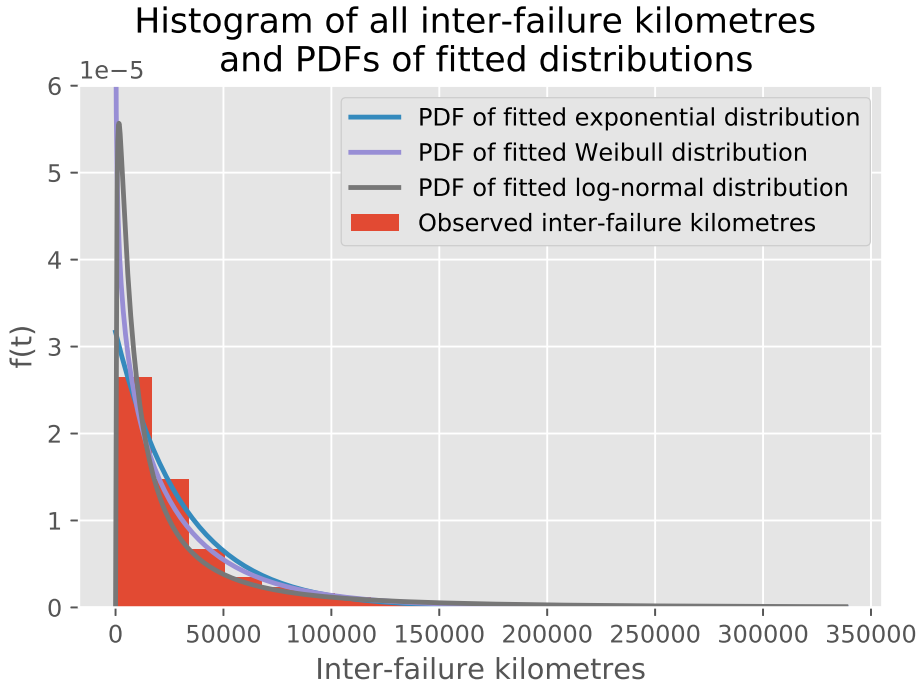
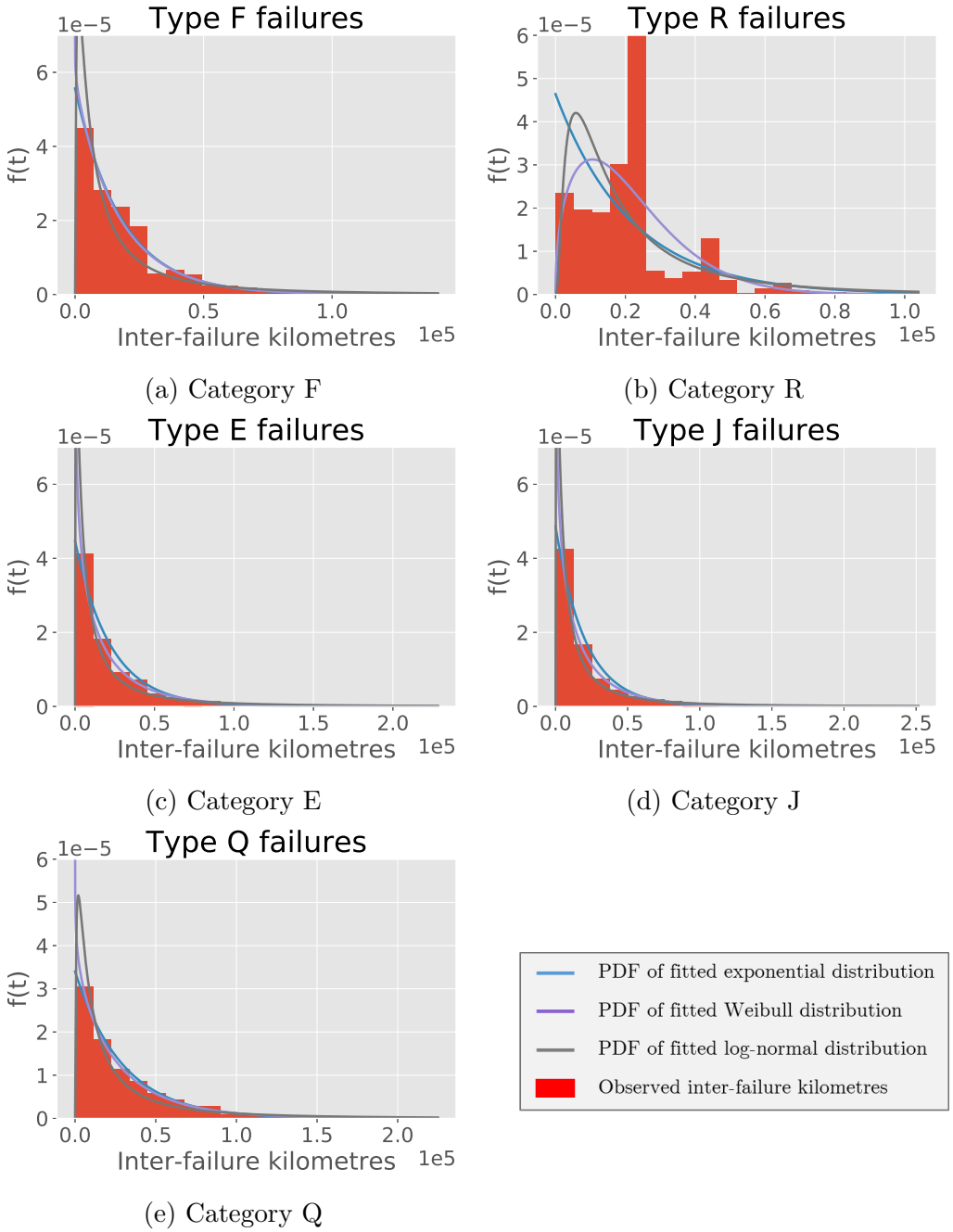


Figure 4.7: Probability density functions (PDF) $f(t)$ of distributions fitted to all inter-failure kilometres.

Table 4.3: P-value of KS-test performed for three distribution on all inter-failure kilometres larger than zero.

Distribution	p**-value**	**Null hypothesis**
Exponential	1.8e−65	Rejected
Weibull	1.9e−15	Rejected
Log-normal	1.8e−86	Rejected



(a) Category F

(b) Category R

(c) Category E

(d) Category J

(e) Category Q

Figure 4.8: Histograms of category specific inter-failure kilometres and probability density functions $f(t)$ of fitted distributions.

Table 4.4: *P*-values of KS-test performed for three distribution on inter-failure kilometres larger than zero for failure categories F, R, E, J and Q.

	Distribution	***p*-value**	**Null hypothesis**
F	Exponential	0.03	Rejected
	Weibull	0.01	Rejected
	Log-normal	0.0	Rejected
R	Exponential	0.0	Rejected
	Weibull	0.0	Rejected
	Log-normal	0.0	Rejected
E	Exponential	2.3e−14	Rejected
	Weibull	0.14	Accepted
	Log-normal	7.9e−10	Rejected
J	Exponential	0.0	Rejected
	Weibull	0.11	Accepted
	Log-normal	2.0e−4	Rejected
Q	Exponential	0.02	Rejected
	Weibull	0.04	Rejected
	Log-normal	1.7e−12	Rejected

As it appears that all failures cannot be modelled only with a failure distribution, we want to identify parameters affecting the failure process and link them to the failure distribution. This can be done with an accelerated failure time model (AFTM). In the analysis conducted in this chapter, we found that the Weibull distribution is the best fit for most of the failure types' inter-failure kilometres. Hence, we implement Weibull AFTMs, where the underlying failure distribution is the Weibull distribution. The implementation of the Weibull AFTM is presented in the next chapter.

Chapter 5

Failure Prediction Model

As concluded in the previous chapter, the failures for the rolling stock fleet X cannot be modelled only with a failure distribution. Analyses of the failures of the other fleets gave similar results. Hence, we construct a survival regression model, which allows us to include the effect of parameters identified as relevant in the modelling of failures.

In this chapter we illustrate how an accelerated failure time model is fitted to the failure data presented in the previous chapter and how the model produces an estimate for upcoming repair needs and repair times. We also provide a detailed mathematical presentation of the accelerated failure time model and present how the model was validated. The model is developed for failure prediction for 12 fleets, but in this thesis we demonstrate the analysis and modelling with the results for only fleet X. The model is implemented using Python coding language, specifically utilising packages scipy.stats (The SciPy community, 2019) and lifelines (Davidson-Pilon, 2019).

5.1 Model Structure

We develop a model for estimating the number of failures during a given time period. The structure of the model is presented in Figure 5.1. The user specifies whether the output should be for a specific rolling stock or for the whole fleet. The user can also give a depot as the model input. In case a depot is specified, the number of failures can be predicted either for a specific fleet or rolling stock repaired at the given depot, or for the sum of the failure repairs for all fleets to be conducted at that depot. The user must specify the prediction period, by defining the starting date of the prediction period

and the number of days to predict. The can also also specify the number of kilometres a rolling stock operates during the prediction period. If the user only specifies the number of kilometres or only the prediction period, the model bases the average kilometres a rolling stock operates per day on historical records on driven kilometres for that specific fleet.

When using this model in the scheduling of next week's maintenance jobs, the user wants to have a prediction on next week's expected number of failures. The user typically knows the number of kilometres the rolling stocks are expected to operate during the week. Hence, the user inputs the dates for the next week and the average expected operating kilometres per rolling stock. For budgeting purposes on the other hand, the prediction period input would probably be a specific month and the operating kilometres could be based on an estimate for that month's average operating kilometres per rolling stock. A fleet engineer examining the failure pattern for the whole fleet or a specific rolling stock individual, could look at predictions for different time intervals and comparing the predictions. In that case, the historical average operating kilometres would probably suffice for the prediction.

The user also inputs dates to define the time period from which data is used for model estimation. In addition, the user specifies whether to take the timing of the previous failure into account in the simulation of failures. If the timing of the previous failure is not to be taken into account, the model assumes that a failure has just occurred, i.e. the kilometres since the previous failure are zero.

The prediction model is based on modelling the kilometres to failure with an accelerated failure time model (AFTM). The AFTM is selected as the failure modelling approach since we want a quantitative model, with the possibility to include the effect of other parameters. The AFTM makes it possible to easily add, alter and remove covariates and provides a quantitative prediction.

An AFTM is fitted separately to the inter-failure kilometres of each fleet and each failure category. Given that there are 18 failure categories for each fleet, predicting the number of failures for all 12 fleets requires generating 216 unique models. If a regular exponential, Weibull or log-normal distribution is a better fit to the data than an AFTM, we choose that distribution for the prediction instead. This is the case, if the inspected factors do not significantly affect the failure process, or if the number of observations is too small. To determine the best suited model or distribution, we compare the maximum log likelihood values of each model and choose the one with the highest log-likelihood.

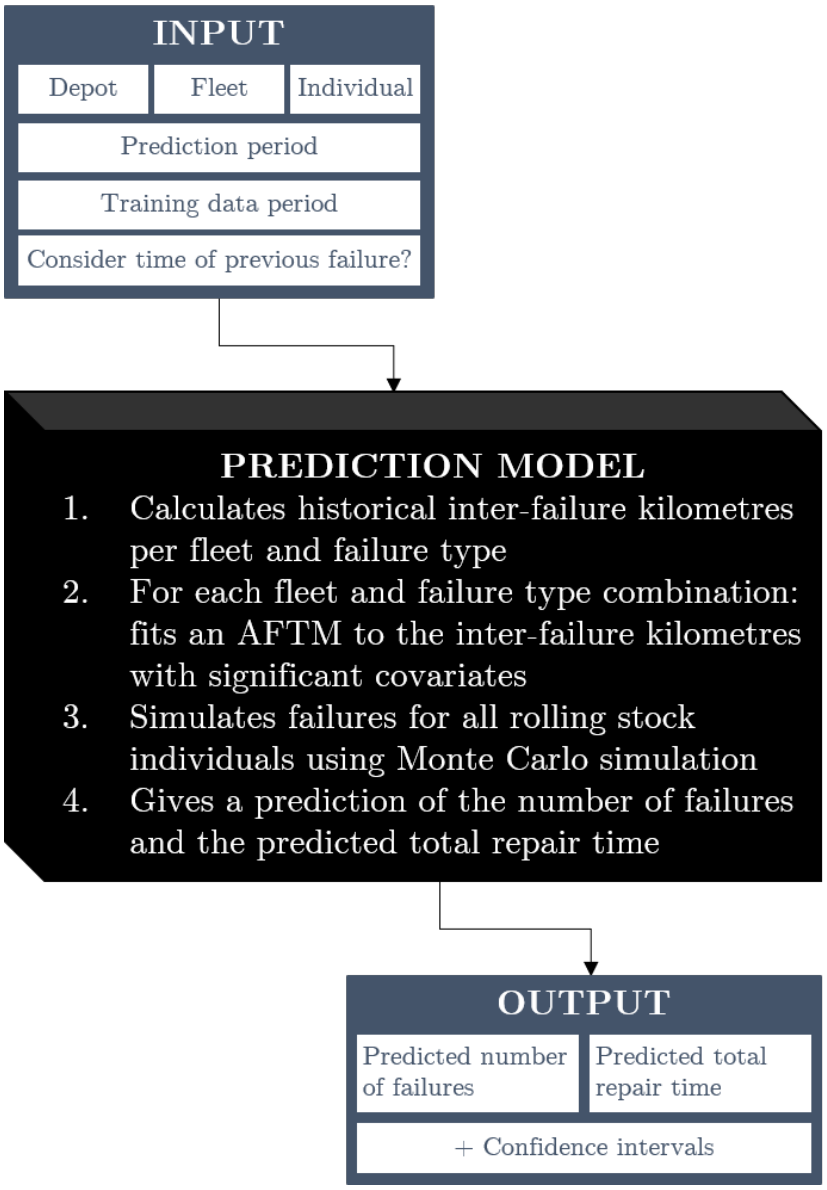


Figure 5.1: The structure of the model used for failure prediction.

The prediction is conducted with Monte Carlo simulation with 100 iterations where failure times are drawn from the generated AFTMs. The model simulates failures in all failure categories for all individuals. The predicted number of failures and repair times are presented as averages across the simulation replications as well as associated confidence intervals. In addition to the expected number of failures, the model also expresses the estimated total repair time for the failures.

It should be noted that there are some computational limitations associated with Monte Carlo simulation. Enough data is needed in order to determine the distributions accurately; a distribution that does not describe the phenomena or system being simulated, leads to incorrect results. Also, to guarantee that the simulations converge to a result with a small enough error term, the number of simulations needs to be large. Depending on the complexity of the simulation model, increasing the number of simulations may lead to a much longer run time. Hence, the number of iterations must be selected so that the result is accurate enough, while still fitting the possible run time limitations. In this thesis, these limitations are not restricting factors for the use of Monte Carlo simulation.

5.2 Mathematical presentation

We presented the accelerated failure time model in Section 3.1.5. We further present the theory of the model as presented by Lee and Wang (2003), focusing on the Weibull AFTM, which is used for failure modelling in this thesis. In the Weibull AFTM, covariates affecting the failure process are linked to the Weibull distribution. A linear relationship between the logarithm of the survival time T and the covariates is an assumption for the AFTM and can be used for analysing survival times. Given a Weibull AFTM, this linear relationship for the a survival time T_i with covariate values $\mathbf{x}_i$ is

$$\log T_i = \beta_0 + \mathbf{x}_i'\beta + \sigma\epsilon_i = \mu_i + \sigma\epsilon_i, \tag{5.1}$$

where β_0 and $\beta = (\beta_1, \beta_2, ..., \beta_p)$ correspond to the regression coefficients, $\mathbf{x}_i' = (x_{1i}, x_{2i}, ..., x_{pi})$ are the covariate values associated with failure time T_i, and ϵ_i is an independently and identically distributed random variable with an extreme value distribution with the density function

$$g(\epsilon) = \exp(\epsilon - \exp(\epsilon))$$

and survival function

$$G(\epsilon) = \exp(-\exp(\epsilon)).$$

The survival time T is Weibull distributed with

$$\gamma = \frac{1}{\sigma} \tag{5.2}$$

and

$$\lambda_i = \exp\Big(-\frac{\mu_i}{\sigma}\Big) = \exp\Big(-\frac{\beta_0 + \mathbf{x}_i'\beta}{\sigma}\Big) = \exp\big(-\gamma(\beta_0 + \mathbf{x}_i'\beta)\big). \tag{5.3}$$

Combining equations (5.3) and (5.2) with the Weibull survival function (equation (3.2)), gives the Weibull survival function with covariates

$$S(t|\lambda_i) = \exp(-\lambda_i t^\gamma). \tag{5.4}$$

Similarly, inserting equations (5.3) and (5.2) into the Weibull hazard function (equation (3.4)) and into the Weibull probability density function (equation (3.3)), presents the Weibull hazard function with covariates

$$\lambda(t|\lambda_i) = \lambda_i \gamma t^{\gamma-1},$$

as well as the Weibull probability density function with covariates

$$f(t|\lambda_i) = \lambda_i \gamma t^{\gamma-1} \exp(-\lambda_i t^\gamma).$$

The relationship presented by equation (5.1) shows the effect of the covariates on the failure time. We assume a simple example to illustrate it. We have one covariate, which can take the value 0 or 1. If $x = 0$, the covariate does not effect the time to failure and we get $T_0 = \exp(\beta_0 + \sigma\epsilon)$. If the covariate is set to $x = 1$, the time to failure is $T_1 = \exp(\beta_0 + \beta_1 + \sigma\epsilon)$, which is equivalent to $T_1 = T_0 \exp(\beta_1)$. This illustrates the effect of the covariate on the time to failure:

- if $\beta_1 > 0$: decelerating effect as T_1 becomes larger than T_0,
- if $\beta_1 < 0$: accelerating effect as T_1 becomes smaller than T_0,
- if $\beta_1 = 0$: no effect as T_1 becomes equal to T_0.

The parameters β_0, β and γ in equations (5.3) and (5.2) are estimated through the maximum likelihood method. The log-likelihood function for the observed n failure times to be maximised is

$$\begin{aligned} l(\beta_0, \beta, \gamma) &= \sum_{i=1}^{n} \log(f(t_i|\lambda_i)) \\ &= \sum_{i=1}^{n} \big[\log\gamma + (\gamma - 1)\log t_i - \gamma(\beta_0 + \mathbf{x}_i'\beta) \\ &\qquad - t_i^\gamma \exp(-\gamma(\beta_0 + \mathbf{x}_i'\beta)) \big]. \end{aligned}$$

The parameters $\mathbf{b} = (\beta_0, \beta, \gamma)$ that maximise the log-likelihood functions correspond to the maximum likelihood estimate (MLE) $\hat{\mathbf{b}}$, which are derived using the Newton-Raphson iterative procedure to find

$$\frac{\partial(l(\mathbf{b}))}{\partial b_i} = 0 \quad \forall\, b_i \in \mathbf{b}.$$

The significance of the covariates' effect on the failure time can be examined using the log-likelihood function. To check if any of the given p covariates have a significant effect on the failure time, we consider the following chi-squared distributed X_L with p degrees of freedom

$$X_L = -2\big(l(\hat{\beta}_0(\mathbf{0}), \mathbf{0}) - l(\hat{\beta}_0, \hat{\beta})\big),$$

where $\hat{\beta}_0(\mathbf{0})$ is the MLE of β_0 given $\beta = 0$. The null hypothesis is

$$H_0 : \beta = 0$$

and is rejected if the 100α percentage point of the chi-squared distribution with p degrees of freedom is smaller than X_L.

Similarly, we can test the significance of a single covariate on the failure time. For the case with $p = 2$, where the first covariate has been found significant, we can test the significance of the second covariate with

$$X_L = -2\big(l(\hat{\beta}_0(\mathbf{0}), \hat{\beta}_1(0), 0) - l(\hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2)\big),$$

where $\hat{\beta}_0(0)$ and $\hat{\beta}_1(0)$ are the MLE of β_0 and β_1 given $\beta_2 = 0$. The null hypothesis

$$H_0 : \beta_2 = 0$$

and is rejected if the 100α percentage point of the chi-squared distribution with p degrees of freedom is smaller than X_L.

5.3 Covariate selection

Choosing relevant covariates for the AFTM can be a challenging process (Lee and Wang, 2003). In this section, we perform a univariate analysis of the impact on the failure times for five parameters chosen based on suggestions from maintenance experts at VR. In addition to the results of the univariate analysis for each chosen parameter, the effect of the parameter as a covariate in the Weibull AFTM and the significance of the covariate are presented in this section. The chosen covariates are:

- Seasonality, examined on a monthly basis
- Average speed of the train units the rolling stock has been part of since the previous failure
- Kilometres since previous preventive maintenance
- Rolling stock age, expressed in total driven kilometres
- Good and bad individuals, differing significantly in performance from the fleet average.

5.3.1 Impact of Season

We start by analysing the season's impact on the inter-failure kilometres. We define the season as the month when the failure occurs. Figure 5.2 presents the average kilometres to failure with standard errors for each month for failures of type F for fleet X.

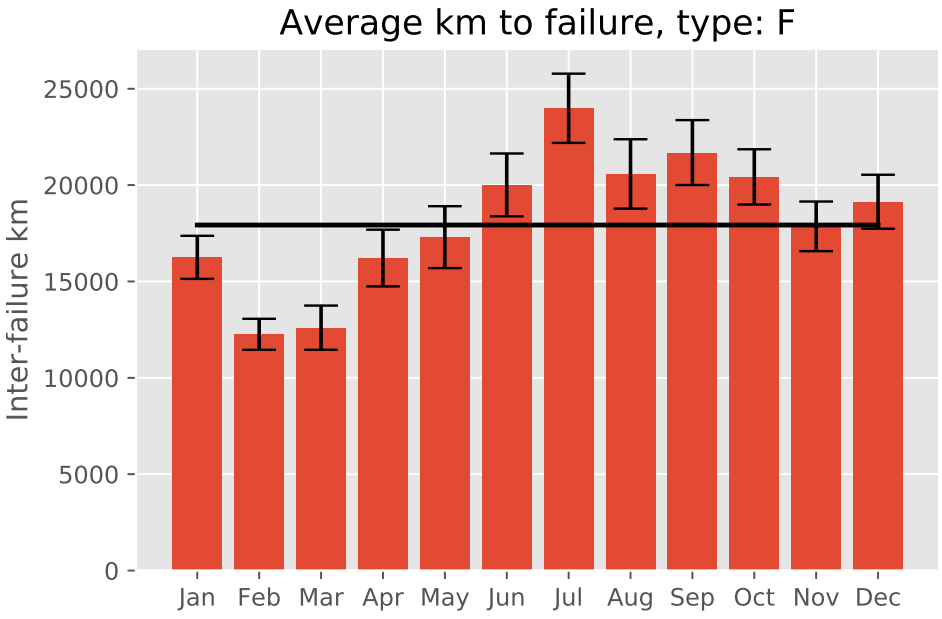


Figure 5.2: The average kilometres to failures each month for failure type F, fleet X.

A larger value on the average inter-failure kilometres is preferred, since it means the average kilometres to failure have been longer. The bar chart

shows that the average inter-failure kilometres have been the longest in July and the shortest in February and March. This indicates that the number of failures per driven kilometre is largest in February and March and smallest in July. This is consistent with the intuitive hypothesis of the failure rate being larger in Finland during the winter and smaller during the summer.

The average inter-failure kilometres for failure category F for fleet X is approximately 18 000 km. This corresponds to around 40 days on average, implying that a rolling stock does not experience a type F failure every month. This might lead to some problems with the seasonal analysis, as the timing of the failures may fall more often on one month than another. However, as the failure records are from a two-year period, the timing issue is likely to be somewhat evened out. Failure records from a longer period would be needed for an accurate analysis of the seasonal effect.

In addition to the graphical analysis of the season's effect on the failure rate, we also perform a two-sample t-test to test whether the average time to failure for each month differs significantly from the average time to failure during the rest of the year. The results for the t-test are presented in Table 5.1. The null hypothesis is that there is no significant deviation from the yearly average kilometres to failure, meaning that the mean values of the two samples are equal:

- $H_0 : \mu_m = \mu_{M\backslash\{m\}}$, the investigated month's ($m \in M$) average inter-failure kilometres are equal to the average inter-failure kilometres for the rest of the year ($M\backslash\{m\}$).
- $H_1 : \mu_m \neq \mu_{M\backslash\{m\}}$, the investigated month's ($m \in M$) average inter-failure kilometres differ from the average inter-failure kilometres for the rest of the year ($M\backslash\{m\}$),

where $M = \{1, 2, ..., 12\}$.

The null hypothesis is rejected if the p-value is below the significance level $\alpha = 0.05$. Before performing the two-sample t-test, we test if the variances for the two samples are equal or not. If the test indicates equal variance, we use the t-test assumption of equal variance, i.e. $\sigma_m = \sigma_{M\backslash\{m\}}$. If the variances for the two samples cannot be assumed equal, we perform a t-test with the assumption of unequal variance, i.e. $\sigma_m \neq \sigma_{M\backslash\{m\}}$. The test for variance equality is performed before the two-sample t-tests for the other covariates as well.

The results in Table 5.1 show that the average inter-failure kilometres differ significantly from the mean of the rest of the year in February, March, July, September and October.

Table 5.1: Results from the t-test for testing of significant deviation from the yearly average inter-failure kilometres for each month. A rejected null hypothesis indicates significant deviation.

Month	**Sample size**	p**-value**	**Null hypothesis**
January	278	0.092	Accepted
February	183	5.1e−06	Rejected
March	131	0.001	Rejected
April	92	0.278	Accepted
May	109	0.573	Accepted
June	107	0.264	Accepted
July	135	4.4e−04	Rejected
August	129	0.128	Accepted
September	130	0.016	Rejected
October	159	0.038	Rejected
November	176	0.781	Accepted
December	141	0.507	Accepted

The season's effect on the survival function of the AFTM is presented in Figure 5.3. The covariate value is included as the historical average inter-failure kilometres for the specified month divided by the longest historical average inter-failure kilometres. Hence, the covariate value equals 1 for the month with the longest inter-failure kilometres and is smaller than 1 for all other months. The two cases of covariate values presented in Figure 5.3 correspond to July (season $= 1$) and February or March (season ≈ 0.6), which can be concluded from the average inter-failure kilometres per month presented in Figure 5.2. As the value for the season-covariate increases the kilometres to failure increase as well, indicating that the covariate has a decelerating effect on the failure time.

Testing the significance of the season covariate in the AFTM model for F type failures for fleet X generates a p-value $p < 0.005$, indicating that the effect of the season covariate is significant.

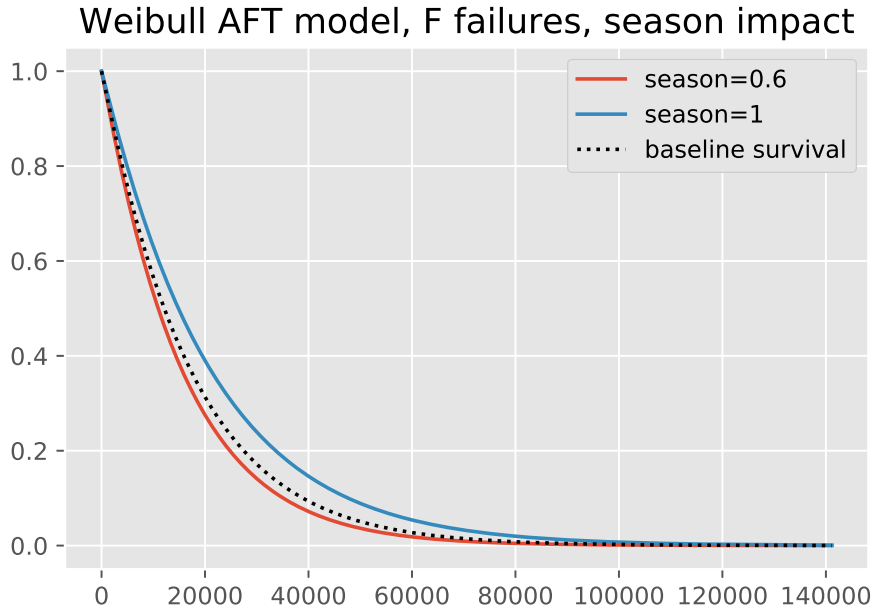


Figure 5.3: Effect of the season covariate on the survival function in the AFTM for F-failures for fleet X. Season = 1 represents July's survival function, and season = 0.6 is an approximate representation of the survival function in February and March.

5.3.2 Impact of Average Speed

Figure 5.4 presents the difference in average inter-failure kilometres at different average speed of the rolling stock since the previous failure. This is a parameter, which may be more relevant for locomotives as, intuitively, the speed affects the stress and performance of the locomotive. The bar chart indicates that the higher the average speed, the shorter the inter-failure kilometres. This follows the intuitive conclusion of a higher speed resulting in an increased number of failures.

The results of the t-test to check for significant deviations between the speed groups are presented in Table 5.2. The hypotheses for the t-test are:

- $H_0 : \mu_s = \mu_{S\backslash\{s\}}$, the investigated speed group's ($s \in S$) average inter-failure kilometres do not differ significantly from the average inter-failure kilometres for the rest of the population ($S\backslash\{s\}$).

- $H_1 : \mu_s \neq \mu_{S\backslash\{s\}}$, the investigated speed group's ($s \in S$) average inter-failure kilometres differ significantly form the average inter-failure kilometres for the rest of the population ($S\backslash\{s\}$),

where $S = \{[0, 80[, [80, 100[, [100, 120[, [120, \infty[\}$.

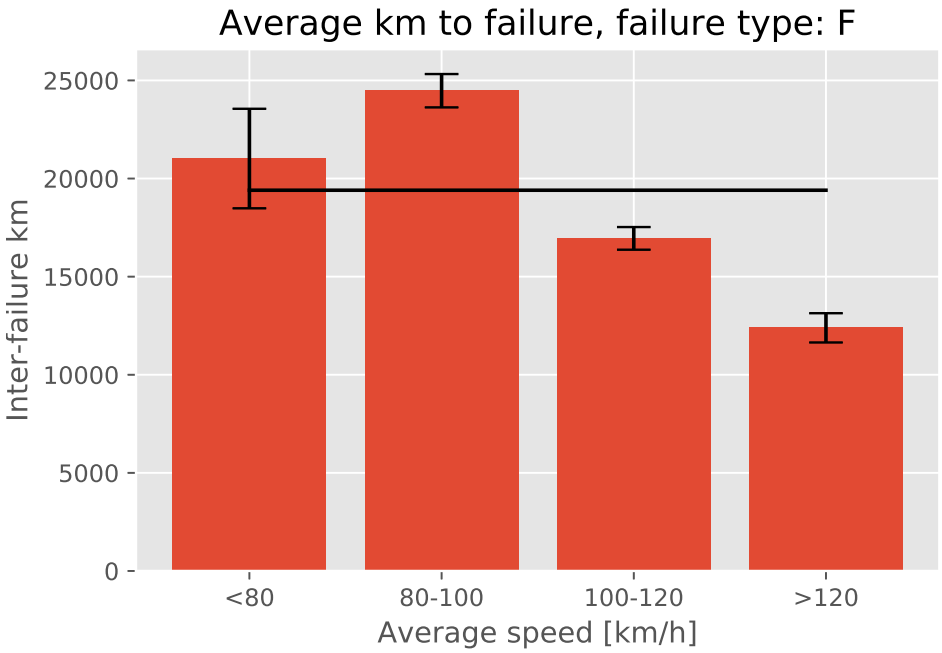


Figure 5.4: The average kilometres to failures for failure type F, fleet X, at different average speeds.

Table 5.2: Results from the t-test for testing for significant deviation from the overall average inter-failure kilometres for different average speeds. A rejected null hypothesis indicates significant deviation.

Speed	**Sample size**	***p*-value**	**Null hypothesis**
$<$ 80 km/h	81	0.052	Accepted
80-100 km/h	685	6.6e−11	Rejected
100-120 km/h	880	2.5e−05	Rejected
$>$ 120 km/h	206	7.7e−07	Rejected

The speeds are calculated as the weighted average rolling stock unit speed limits since the previous failure repair, or since the previous failure if it has not been repaired yet. In the calculations, the speed limit rather than the

actual speed is considered. Hence, the weighted average speed corresponds to the average speed given that the train unit would drive at the speed limit all the time. The results of t-test also indicates that the larger the average speed, the shorter the inter-failure kilometres.

The effects of the average speed on the AFTM survival function are presented in Figure 5.5. Covariate value speed = 0.5 corresponds to the average speed limit being 50 km/h and speed = 1.5 corresponds to the average speed being limit 150 km/h. A lower average speed decelerates the kilometres to failure, whereas a higher speed has an accelerating effect. The test for covariate significance gives the p-value $p < 0.005$, resulting in a rejection of the null hypothesis, indicating the average speed having a significant effect on the failure times of F type failures.

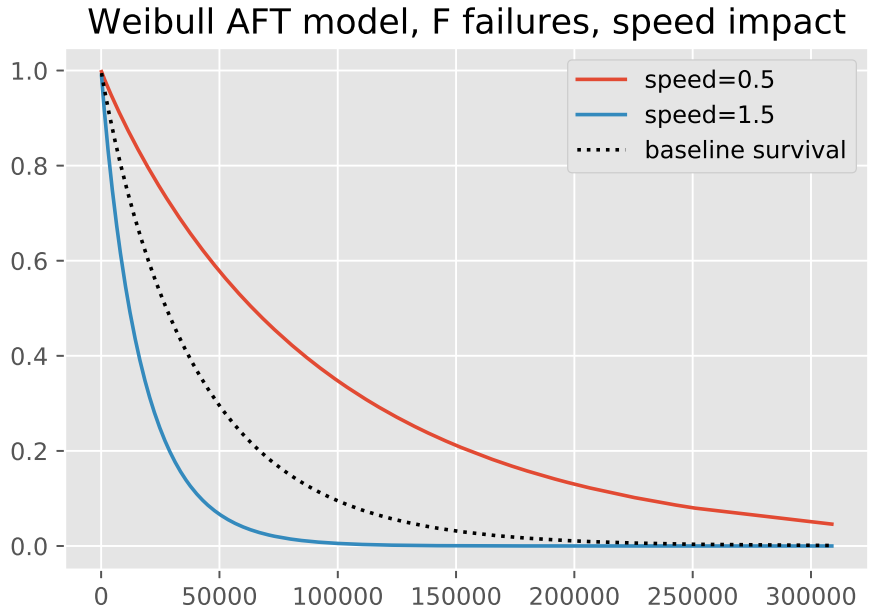


Figure 5.5: Effect of the speed covariate on the survival function in the AFTM for F-failures for fleet X.

5.3.3 Impact of Kilometres since Preventive Maintenance

The effect on the failure process of the number of driven kilometres since the previous preventive maintenance action is presented in Figure 5.6. The

previous preventive maintenance refers to any pro-active maintenance action performed and is not necessarily associated with the failure category. However, several preventive maintenance actions are typically performed at the same maintenance occasion. We can see that the average kilometres to failure are shorter as the kilometres since the previous preventive maintenance action are shorter. This indicates that the failure rate is higher after a preventive maintenance and decreases as the kilometres since the previous preventive maintenance increase, which may appear counter-intuitive. This could be the result of poor maintenance work quality. However, there is probably some other explanation to the counter-intuitive result, such as the users paying more attention to the condition of the rolling stock directly after it has been maintained, hence noticing more failures. Also, this might be the infant mortality effect presented in the bathtub curve in Figure 3.6, that is, the failure probability is larger in the earliest stage and decreases as the system ages.

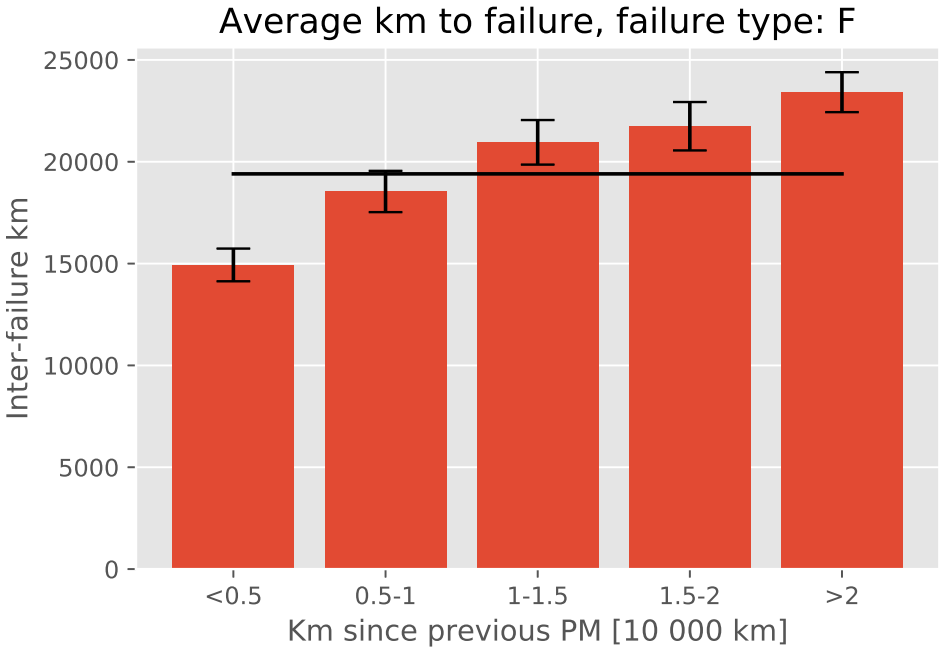


Figure 5.6: The average kilometres to failures for failure type F, fleet X, for different kilometres since the previous preventive maintenance.

We perform a t-test to check for significant deviations between the groups and the results are presented in Table 5.3. The hypotheses for the t-test are:

- $H_0 : \mu_m = \mu_{M\backslash\{m\}}$, the average inter-failure kilometres for the investigated kilometres since previous preventive maintenance ($m \in M$) do not differ significantly from the average inter-failure kilometres for the rest of the population ($M\backslash\{m\}$).
- $H_1 : \mu_m \neq \mu_{M\backslash\{m\}}$, the average inter-failure kilometres for the investigated kilometres since previous preventive maintenance ($m \in M$) differ significantly from the average inter-failure kilometres for the rest of the population ($M\backslash\{m\}$),

where $M = \big\{[0, 0.5[,\ [0.5, 1[,\ [1, 1.5[,\ [1.5, 2[,\ [2, \infty[\big\}$.

Table 5.3: Results from the t-test for testing for significant deviation from the overall average inter-failure kilometres for different kilometres since the previous preventive maintenance. A rejected null hypothesis indicates significant deviation.

Km since PM	**Sample size**	p**-value**	**Null hypothesis**
< 5000 km	512	2.0e−11	Rejected
$5000 - 10000$ km	362	0.142	Accepted
$10000 - 15000$ km	332	0.133	Accepted
$15000 - 20000$ km	301	0.019	Rejected
> 20000 km	345	4.2e−06	Rejected

We can conclude that there are significant deviations between the groups. The effect of the kilometres since previous preventive maintenance as a covariate in the AFTM is presented in Figure 5.7. The estimated survival function for an individual with 5 000 km since the previous preventive maintenance falls below the baseline survival function, whereas it is above the baseline function for an individual with 25 000 km since the previous preventive maintenance. The test for covariate significance gives the p-value $p < 0.005$, resulting in a rejection of the null hypothesis, indicating that the kilometres since previous preventive maintenance have a significant effect on the failure times.

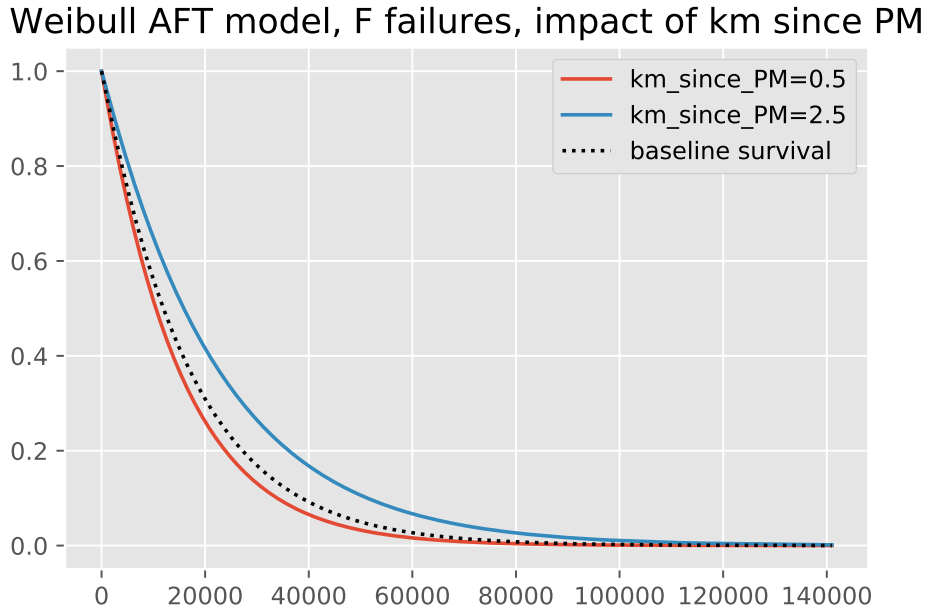


Figure 5.7: Effect on the survival function of the covariate illustrating the kilometres since previous preventive maintenance in the AFTM for F-failures for fleet X.

5.3.4 Impact of Age

Next, we examine what effect the rolling stock's age, expressed in total driven kilometres, has on the inter-failure kilometres. Figure 5.8 presents the average inter-failure kilometres with standard errors for type F failures for fleet X, for groups with different total driven kilometres representing the age of the rolling stock. The bar chart does not show the age having a clear significant impact on the failure process for this fleet for F failures. Only the youngest group seems to have an average inter-failure kilometre differing significantly from the other age groups. This group has a smaller average inter-failure kilometre, indicating that the failure rate is higher for this group.

We conduct a t-test to check for significant deviation of the average kilometres to failure for each age group compared to the rest of the population's inter-failure kilometre average. The hypotheses for the test are:

- $H_0 : \mu_a = \mu_{A\setminus\{a\}}$, the investigated age group's ($a \in A$) average inter-

failure kilometres do not differ significantly from the average inter-failure kilometres for the rest of the population ($A \backslash \{a\}$).

- $H_1 : \mu_a \neq \mu_{A\backslash\{a\}}$, the investigated age group's ($a \in A$) average inter-failure kilometres differ significantly form the average inter-failure kilometres for the rest of the population ($A \backslash \{a\}$),

where $A = \big\{[0, 1.5[, [1.5, 2[, [2, 2.5[, [2.5, 3[, [3, \infty[\big\}$.

The results are presented in Table 5.4. The t-test shows a significant difference in average inter-failure kilometres between the youngest group of rolling stocks and the rest of the population, which is consistent with the observation based on Figure 5.8. This could be explained by the rolling stocks experiencing more failures spend more time at the depot, hence not accumulate as much kilometres as the other rolling stocks, which places them in the youngest group in this analysis. However, combining the results from this covariate analysis with the results on "good" and "bad" individuals presented in the next section does not place the significantly worse performing individuals in the youngest group.

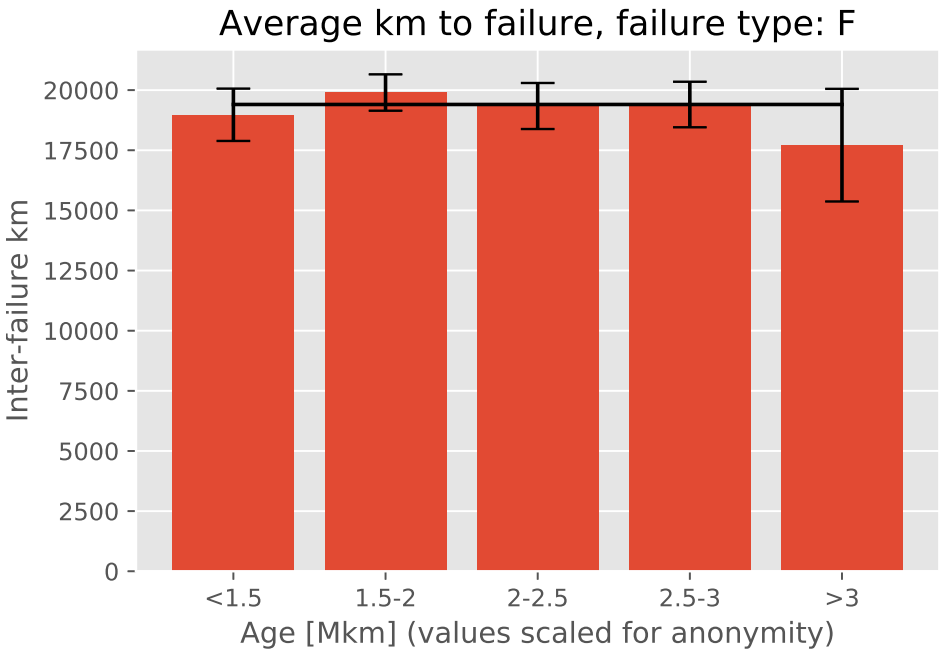


Figure 5.8: The average kilometres to failures for failure type F, fleet X, for different age groups (the total driven kilometres scaled for anonymity).

Table 5.4: Results from the t-test for testing for significant deviation from the overall average inter-failure kilometres for different age groups. A rejected null hypothesis indicates significant deviation.

Age (total driven km)	**Sample size**	p-**value**	**Null hypothesis**
$<$ 1.5 Mkm	326	0.022	Rejected
1.5-2 Mkm	632	0.104	Accepted
2-2.5 Mkm	459	0.775	Accepted
2.5-3 Mkm	350	0.480	Accepted
$>$ 3 Mkm	85	0.879	Accepted

Figure 5.9 illustrates the effect of the age covariate on the survival function. The survival function is very similar for both presented age groups. The test for covariate significance gives a p-value $p = 0.83$, based in which the null hypothesis of insignificance is accepted. Therefore, the age covariate do not have a significant effect on the F failures for fleet X.

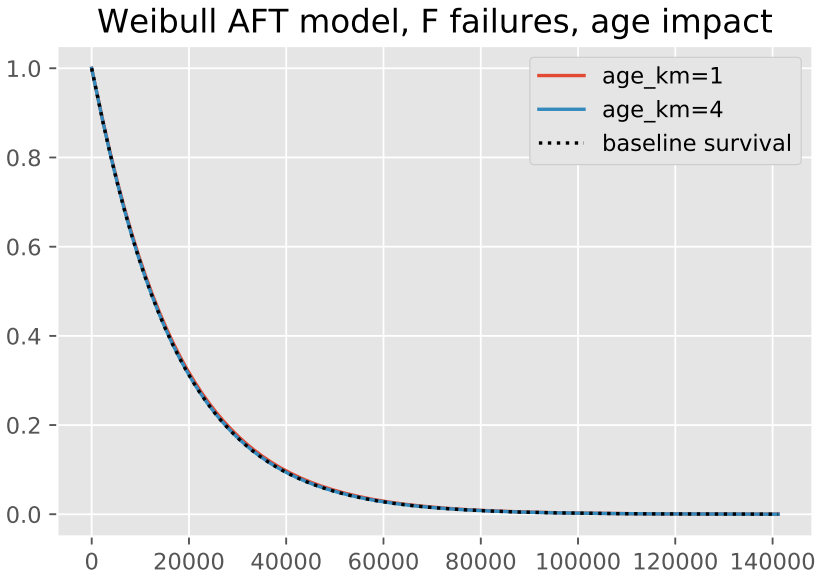


Figure 5.9: Effect of the age covariate on the survival function in the AFTM for F-failures for fleet X.

5.3.5 Difference between individuals

Finally, we want to examine if there are significant differences between the failure processes of different individuals in the fleet. We want to identify such individuals that perform better than the rest of the population, as well as individuals that perform worse than others. The data consists of a minimum of 8 failures for each rolling stock, hence we can perform a two-sample t-test to test for significant differences.

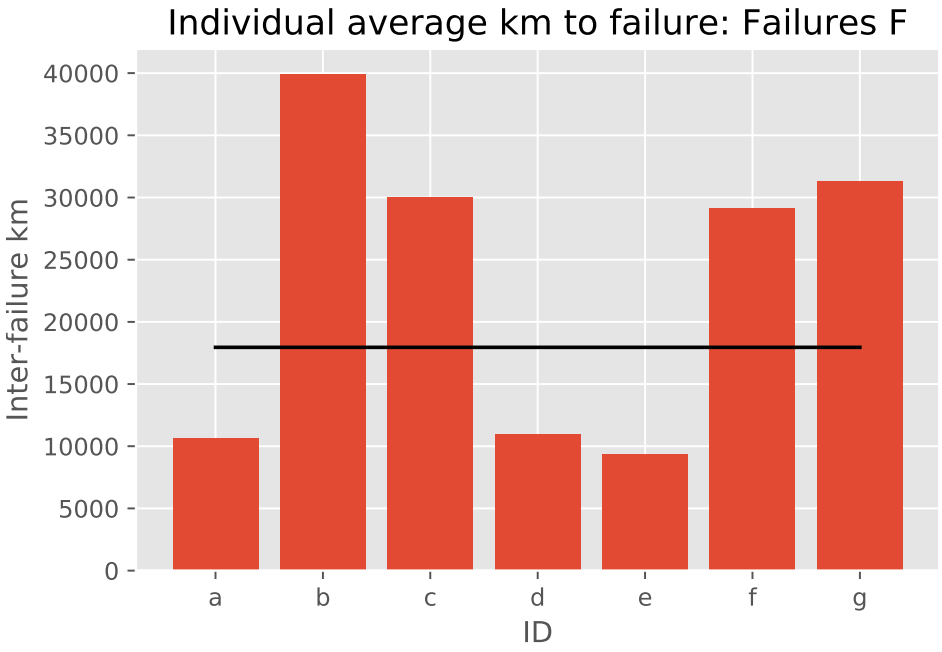


Figure 5.10: The average kilometres to failure for seven individuals from fleet X that were found to differ significantly from the rest of the population.

A two-sample t-test helps us identify individuals, the average inter-failure kilometres of which are significantly longer or shorter than those of the rest of the population. The hypotheses for the test are:

- H_0: $\mu_i = \mu_{I\backslash\{i\}}$: the investigated individual's ($i \in I$) average inter-failure kilometres do not differ significantly from the average inter-failure kilometres for the rest of the population ($I\backslash\{i\}$).
- H_1: $\mu_i \neq \mu_{I\backslash\{i\}}$: the investigated individual's ($i \in I$) average inter-failure kilometres differ significantly from the average inter-failure kilometres for the rest of the population ($I\backslash\{i\}$),

where I corresponds to the set of all fleet individuals.

For fleet X, we identify seven individuals that have a significant difference. The average failure times for these individuals together with the population average failure time are presented in Figure 5.10. We can see that individuals a, d and e have significantly shorter average inter-failure kilometres, whereas individuals b, c, f and g have significantly longer average inter-failure kilometres compared to the rest of the population. Individuals a, d and e are classified as "bad". Similarly, individuals b, c, f and g are classified as "good".

We introduce two new covariates, classes "bad" and "good". The effects of these covariates on the AFTM survival function are presented if Figures 5.11 and 5.12. We can see that the covariate indicating a "bad" individual accelerates the failure time, whereas the covariate indicating a "good" individual decelerates it.

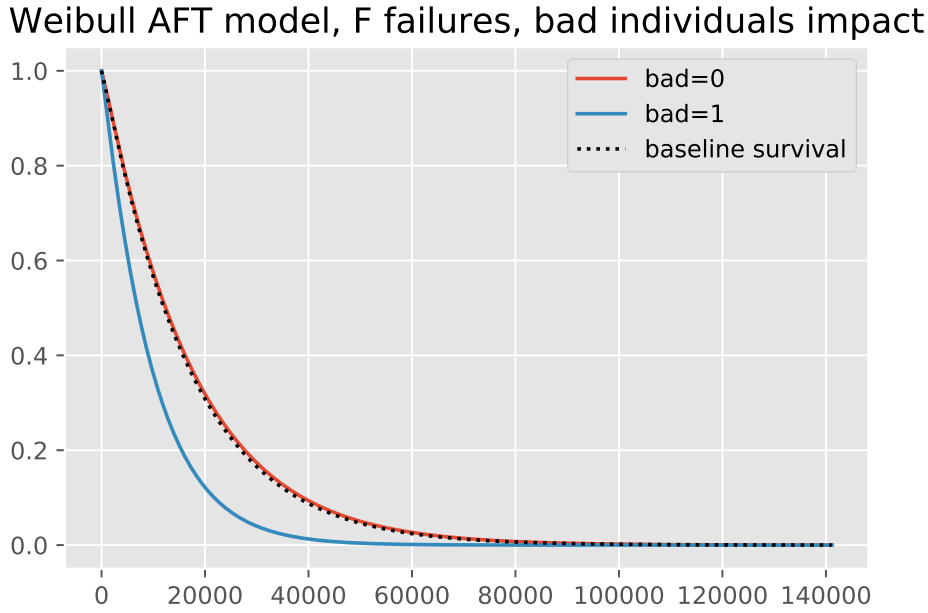


Figure 5.11: Effect on the survival function of the individuals classified as "bad" in the AFTM for F-failures for fleet X.

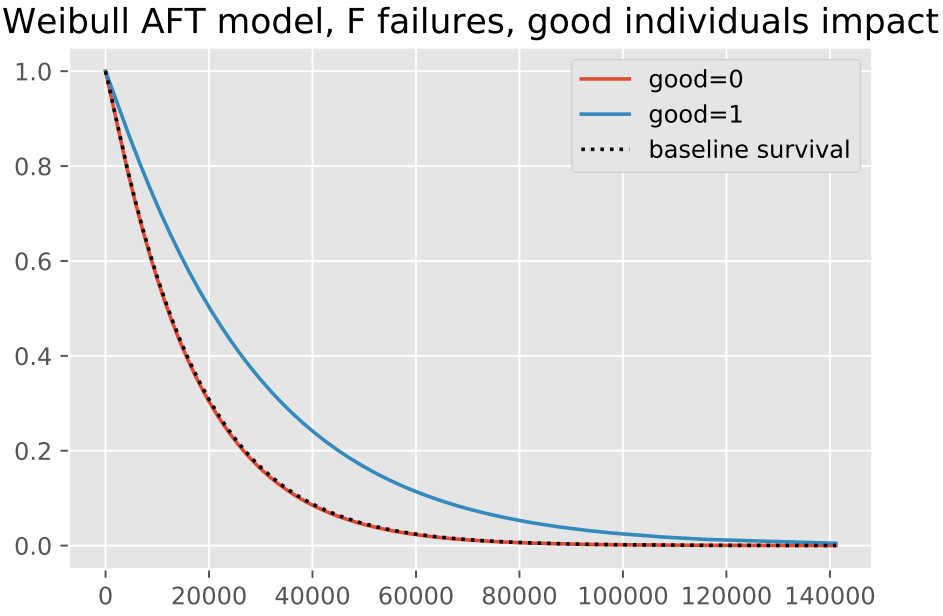


Figure 5.12: Effect on the survival function of the individuals classified as "good" in the AFTM for F-failures for fleet X.

5.4 Accelerated Failure Time Models

The analysis of the covariate effects on failures of type F presented in the previous section resulted in some covariates found significant, while others were not. Similar analyses were conducted for all failure types. Table 5.5 presents the covariates found significant for each failure type for fleet X. The significance is determined based on the significance statistic presented in Section 5.2, with a significance level $\alpha = 0.05$. There is a lack of data on failures of type P and U, hence these cannot be modelled with an AFTM and are left out of the model.

Table 5.5: The covariates included in the AFTMs for each failure type for fleet X, the average failure interval in days for the failure type, and the concordance index (c-index), representing the predictive accuracy for each model.

Failure type	Season	Speed	Km since PM	Age	Good	Bad	Average interval	C-index
B	x	x			x	x	111	0.67
C	x	x	x		x	x	70	0.59
D	x	x			x	x	86	0.57
E	x	x	x		x	x	50	0.62
F	x	x	x		x	x	40	0.61
G	x	x			x	x	76	0.58
H	x				x		105	0.57
J	x					x	46	0.57
K		x				x	98	0.59
L	x	x				x	91	0.61
M	x	x		x		x	114	0.63
N	x	x	x	x	x		116	0.61
P	NA	NA	NA	NA	NA	NA	157	NA
Q	x	x	x		x	x	65	0.62
R	x	x	x			x	48	0.61
S	x	x			x		127	0.59
T	x	x			x		97	0.59
U	NA	NA	NA	NA	NA	NA	NA	NA

For type F failures, modelled with five covariates, the logarithm of the failure time (equation (5.1)) will be of the form

$$\log T_i = 9.38 + 1.18x_{se,i} - 0.73x_{sp,i} + 0.22x_{PM,i} + 0.5x_{g,i} - 0.64x_{b,i} + 0.02\epsilon_i.$$

We note that the regression parameters representing the season, kilometres since previous preventive maintenance and "good" individuals are positive, indicating that these covariates have an decelerating effect on the failure time as the value of the covariates increase. The regression parameters representing the average speed and "bad" individuals are negative, indicating that these covariates have an accelerating effect on the failure time as the value of the covariates increase. The survival function (equation (5.4)) for failure category F is then

$$S_F(t|\mathbf{x}_i) = \exp\Big(- \exp\Big(\frac{1}{0.02}\big(9.38 + 1.18x_{se,i} - 0.73x_{sp,i} + 0.22x_{PM,i} + 0.5x_{g,i} - 0.64x_{b,i}\big)\Big)t^{1/0.02}\Big).$$

Figure 5.13 presents the survival function for failure category F with different covariate values; Table 5.6 presents the covariate values for the three different survival functions in the figure. The average case corresponds to the survival function in June/August, when the average speed since the previous failure has been 100 km/h, the kilometres since the previous preventive maintenance action are 10 000 km, and when the rolling stock individual in question does not perform significantly better or worse than the rest of the fleet. The good case corresponds to the survival function in July, when the average speed has been 80 km/h, the kilometres since the previous preventive maintenance actions are 20 000 km and the rolling stock individual performs significantly better than the rest of the fleet. The bad case corresponds to the survival function in February/March, when the average speed has been 120 km/h, the kilometres since the previous preventive maintenance actions are 5 000 km and the rolling stock individual performs significantly worse than the rest of the fleet.

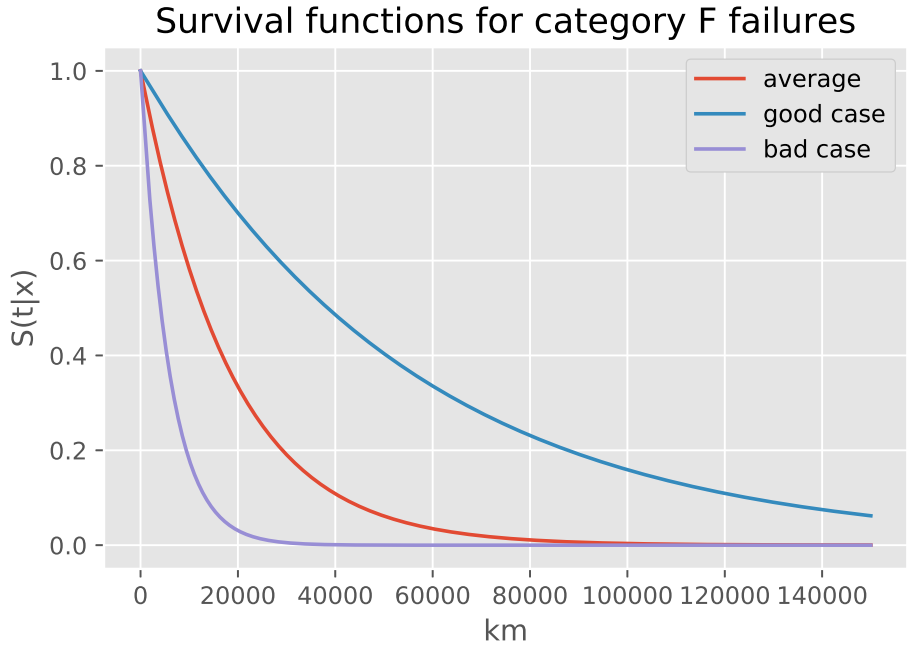


Figure 5.13: Three different survival functions for category F failures. The covariate values for each survival function are presented in Table 5.6.

Table 5.6: Covariate values for the three survival functions presented in Figure 5.13.

	Average	**Good case**	**Bad case**
Season	0.8	1	0.6
Speed	1	0.8	1.2
Km since PM	1	2	0.5
Good	0	1	0
Bad	0	0	1

5.4.1 Concordance Index

For survival models, traditional model accuracy measures, such as R^2, are not directly applicable. What can be used instead is, for example, the concordance index. The concordance index, also referred to as the c-index, is a measure for predictive accuracy of a survival prediction model (Steck et al., 2008). It is the probability that the model correctly predicts the order of the failures for a random pair of subjects. A model with concordance index of 0.5 indicates random predictions, whereas an index of 1 implies a perfect model. A concordance index of 0 implies perfect anti-concordance and can be transformed to a perfect model by multiplying the coefficients by minus one.

In this thesis we have data that is uncensored, meaning that the data only consists of data points where a failure already has occurred. When there is no censoring, the concordance index can be estimated by the proportion of the pairs that the model has ordered correctly. The predicted order of any two subjects i and j is determined by the predicted survival times $S(\hat{t}_i|\mathbf{x}_i)$ and $S(\hat{t}_j|\mathbf{x}_j)$. To calculate the concordance index for all pairs of observations, we compare the prediction and the observed failure time. The pair (i, j) is concordant if $t_i < t_j$ and $S(\hat{t}_i|\mathbf{x}_i) < S(\hat{t}_j|\mathbf{x}_j)$.

The formula for calculating the concordance index c for uncensored data is

$$c = \binom{n}{2}^{-1} \sum_{t_i<t_j} \sum \Big[\mathbb{1}(S(\hat{t}_i|\mathbf{x}_i) > S(\hat{t}_j|\mathbf{x}_j) \Big],$$

where $\mathbb{1}(\cdot)$ is an indicator function, taking the value 1 if the condition is met and value 0 otherwise.

The concordance indices for the AFTMs for each failure type for fleet X are presented in Table 5.5. The values range from 0.57 (D, H and J failures) to 0.67 (B failures). Due to noise in the data, the concordance index of a

fitted survival model generally ranges between 0.55 and 0.7 (Davidson-Pilon, 2019). All AFTMs have an index larger than 0.55, indicating at least some modelling accuracy. However, as the majority of the concordance indices are closer to 0.55 than to 0.7, we want to further examine the accuracy of the models. Also, the concordance index only illustrates how well the AFTM describes the data used for estimating the model, not considering the prediction accuracy of the AFTM. Hence, we use a validation dataset to test prediction accuracy. These results are presented and discussed in Section 6.1.

5.5 Modelling Failures Found During Repair

As concluded in Chapter 4, 9% of the data consists of inter-failure kilometres equal to zero. The inter-failure kilometres equal to zero cannot be modelled with a Weibull AFTM. Therefore, we have to develop another model for include these failures in the overall prediction of failures. For this we use a combination of Bernoulli distributions. First, we calculate the probability of a failure being found in conjunction with repair, i.e. at the same depot stop. This is simply the number of instances one or more failures are found during repair, divided by the number of failures with inter-failure kilometres larger than zero. Thereafter, we examine how many failures are found at each repair and calculate the probabilities for each amount of failures.

To illustrate this, we give an example using the data for one failure category. The probability of finding a failure in conjunction with a repair is

$$p = \frac{N_{\text{One or several failures found during repair}}}{N_{\text{Failure with inter-failure km }>0}} = \frac{75}{802} \approx 0.0935.$$

This indicates a probability of finding a failure in conjunction with repair to be 9.35%. For this failure type, number of failures found during repair has been either 1, 2 or 3. The probability of finding each of the number of failures is simply the occurrence times for the number of failures, divided by the number of times failures are found during repair. For the failure category, the probability of finding $i\,(i \in \{1, 2, 3\})$ failures during a depot stop indicates

$$p_i = \frac{N_{i \text{ failures found during repair}}}{N_{\text{One or several failures found during repair}}}.$$

For $i = 2$, this is

$$p_2 = \frac{5}{75} \approx 0.0667.$$

That is, if failures are found in conjunction with repair, the number of failures found is 2 with a probability of 6.67%. Combining the two calculated probabilities gives the probability of finding two failures in conjunction with repair, namely $0.0935 \cdot 0.0667 \approx 0.0062 = 0.62\%$.

Based on the calculated probabilities of finding failures at the depot, one or several failures may be added to the predicted number of failures in the prediction model. Given the above example, on average 0.62% of the simulated failures result in adding two additional failures to the prediction of total number of failures in this failure category.

5.6 Simulation Model

The flowchart in Figure 5.14 presents the model used for simulating failures. The model takes user inputs as specified in Figure 5.1. For each failure category the model starts by checking if the data consists of at least 70 failures in that failure category. We assume that at least 70 data points are needed for estimating a failure model. We found 70 data points to be a suitable limit when analysing the data and fitting AFTMs to the inter-failure kilometres. If the historical data consists of less than 70 data points, that failure category is not considered at all in the prediction of failures. This does not affect the result significantly, since if a failure category has had fewer than 70 failures during the past two years, there will probably not be that many failures of that category during the prediction period.

If the number of historical failures is larger than 70 an AFTM is fitted to the data. If the fitting was successful, the log-likelihood of the AFTM is compared to the log-likelihoods of fitting regular exponential, Weibull and log-normal distributions to the data. If an AFTM could not be fitted, or if a regular distribution fits the data better (based on the log-likelihood values), a regular distribution is used for the simulation of failures in this failure category. The simulation procedure in this case is presented in Figure 5.16. If an AFTM is the best fit to the data, the AFTM is used in the simulation of failures. This simulation procedure is presented in Figure 5.15.

After the failures have been simulated, the model saves the results for all simulation rounds for this failure category and moves on to the next category. If all failure categories have already been simulated, the model moves on to calculating the final results. First, the sum of the failures in all failure categories is calculated for each simulation round. The prediction of the number of failures is the average number of failures for the simulation rounds.

The 95% confidence intervals are calculated by sorting the simulation round results and choosing the rounds corresponding to 2,5% and 97,5% of the iterations. That is, as we simulate failures 100 times and sort the values from smallest to largest, the lower confidence interval is the 3rd result and the upper confidence interval is the 98th result. The prediction for critical failures and repair times are calculated similarly.

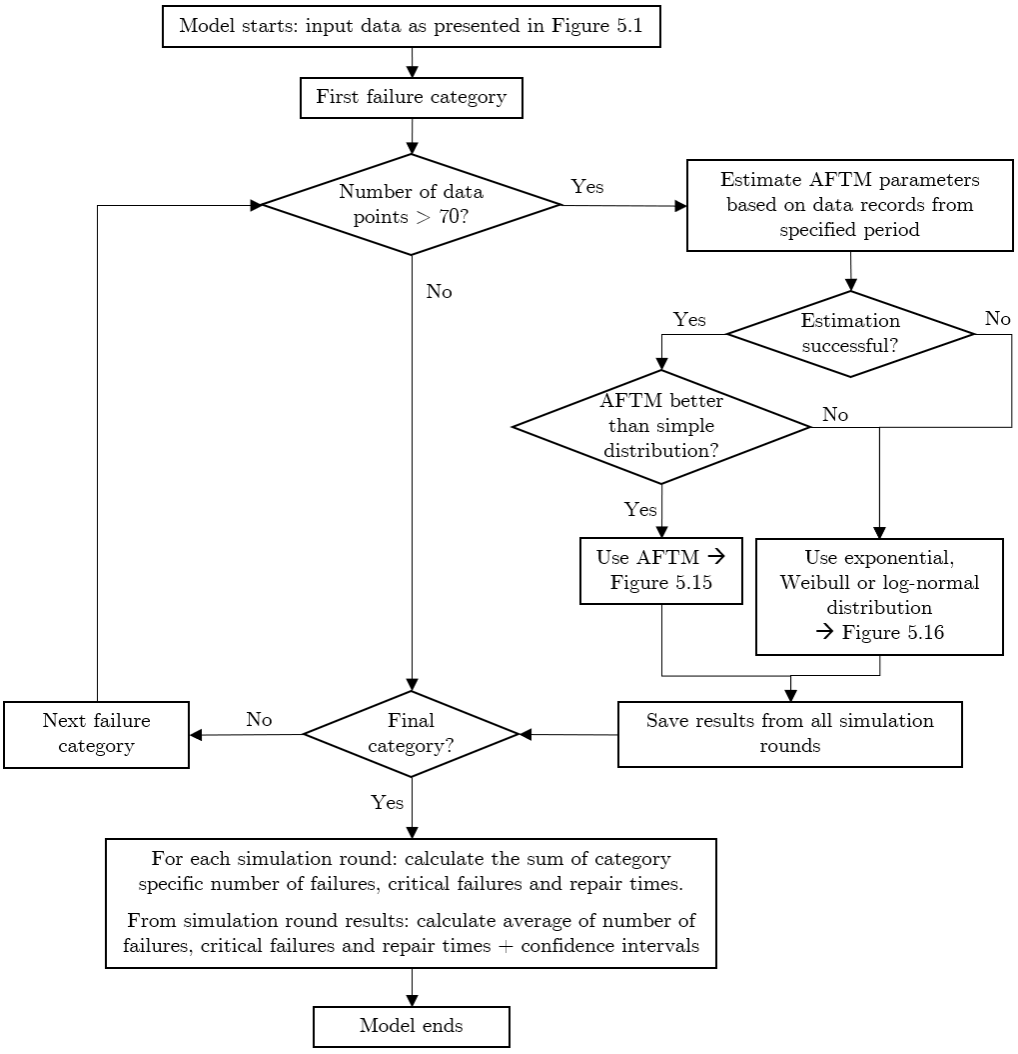


Figure 5.14: An overview of the simulation model. For a more detailed presentation of the simulation of the actual failures, see Figure 5.15 or Figure 5.16, depending on if an AFTM could be fitted to the failure category data or not and if the AFTM is a better fit to the data than a regular distribution.

The flowchart in Figure 5.15 illustrates the simulation procedure for the failures of a failure category to which an AFTM is a good fit. In the flowchart, N corresponds to the number of failures, H is the prediction horizon in kilometres and S is the number of simulation rounds to be performed.

In the beginning of each simulation round the number of failures is set to zero. Failures are simulated for each rolling stock individual and the number of failures from the simulation round is the combined number of failures for all individuals. First, the AFTM covariate values are set for the individual based on the current state, namely the month, the kilometres since the previous preventive maintenance action, the average speed since the previous failure, the total kilometres of the rolling stock individual, and values 0/1 indicating if the individual performs significantly better or significantly worse than the rest of the fleet. Depending on the user's input the model either assumes that a failure has just occurred, i.e. the kilometres since the previous failure are zero, or the model takes into account the timing of the previous failure. If the previous failure is not considered the season covariate value of the month the prediction period begins is used. In the case that the previous failure is taken into account, the season covariate used for generating the first failure is that of the month the previous failure occurred in. The total simulated kilometres (Km_ tot) are set to zero.

A failure kilometre (Km) is generated from the Weibull AFTM distribution with the specified covariate values. The failure time follows a distribution with the survival function in equation (5.4). If the previous total simulation kilometres plus the simulated failure kilometre are shorter than the specified prediction horizon, a failure is added to the number of failures. In the case that the timing of the previous failure is considered, the model checks that the first generated failure kilometre exceeds the kilometres since the previous failure and generates new failure kilometres until this is the case. This corresponds to the conditional probability of failure, given that the rolling stock has survived until the beginning of the prediction period without failing.

Thereafter, a value between 0 and 1 is generated from a uniform distribution. If the number (Rand) is below the probability of finding additional failures in conjunction with repair, a number of failures is generated and added to the total number of failures. The calculation of the probability of finding failures during repair and how the number of failures found is generated are presented in Section 5.5. This way the failures with inter-failure kilometres equal to zero are included in the total number of predicted failures.

After the failures have been added to the total number of failures, another value between 0 and 1 is generated from a uniform distribution. This number

(Rand) illustrates the probability of the next failure being found before the previous failure has been repaired. If Rand falls below the probability, the total simulated kilometres are only updated by adding the simulated failure kilometre. Otherwise, the total simulated kilometres are updated by adding the simulated failure kilometre as well as the average kilometres from failure to repair for the specific failure category. This way the failure process illustrated in Figure 4.4 is accurately modelled.

When the simulated total kilometres have been updated, the covariate values are updated as well. If the month has changed, the season covariate value is updated. The age of the rolling stock is updated in the same way as the total simulated kilometres, as are the kilometres since previous preventive maintenance action. We however assume that a preventive maintenance action is performed every 25 000 km, which means that the covariate value for kilometres since previous preventive maintenance actions is always under 25 000 km.

Next, the model checks if the total simulated kilometres are longer than the prediction horizon. If not, a new failure kilometre is generated from the Weibull AFTM distribution with the updated covariate values and the model checks if the generated kilometre falls within the horizon. If the total simulated kilometres become longer than the horizon, the model moves on to simulating failures for the next individual.

After simulating the failures for all individuals, the model calculates the number of critical failures as well as the total repair times. The number of critical failures is simply the number of failures times the historical share of critical failures. The total repair time is the number of failures times the average repair time for the failure category. These values are saved and the model moves on to the next simulation round. If this was the final simulation round, the results of this failure category are saved and the model moves on to the next failure category (Figure 5.14).

In the case that an AFTM is not the best fit to the category failures, the simulation of failures follow the process presented in Figure 5.16. The process is the same as when simulating failures using an AFTM, but now no covariate values have to be considered or updated. The model simply generates failure times from an exponential, Weibull or log-normal distribution, depending on which is the best fit to the historical inter-failure kilometres. The goodness of fit is determined with the KS-test presented in Section 4.2, and the distribution with the highest p-value in the KS-test is chosen. Given that the Weibull distribution is the best fit, the failure kilometres are generated from a distribution with a survival function as presented in equation (3.2).

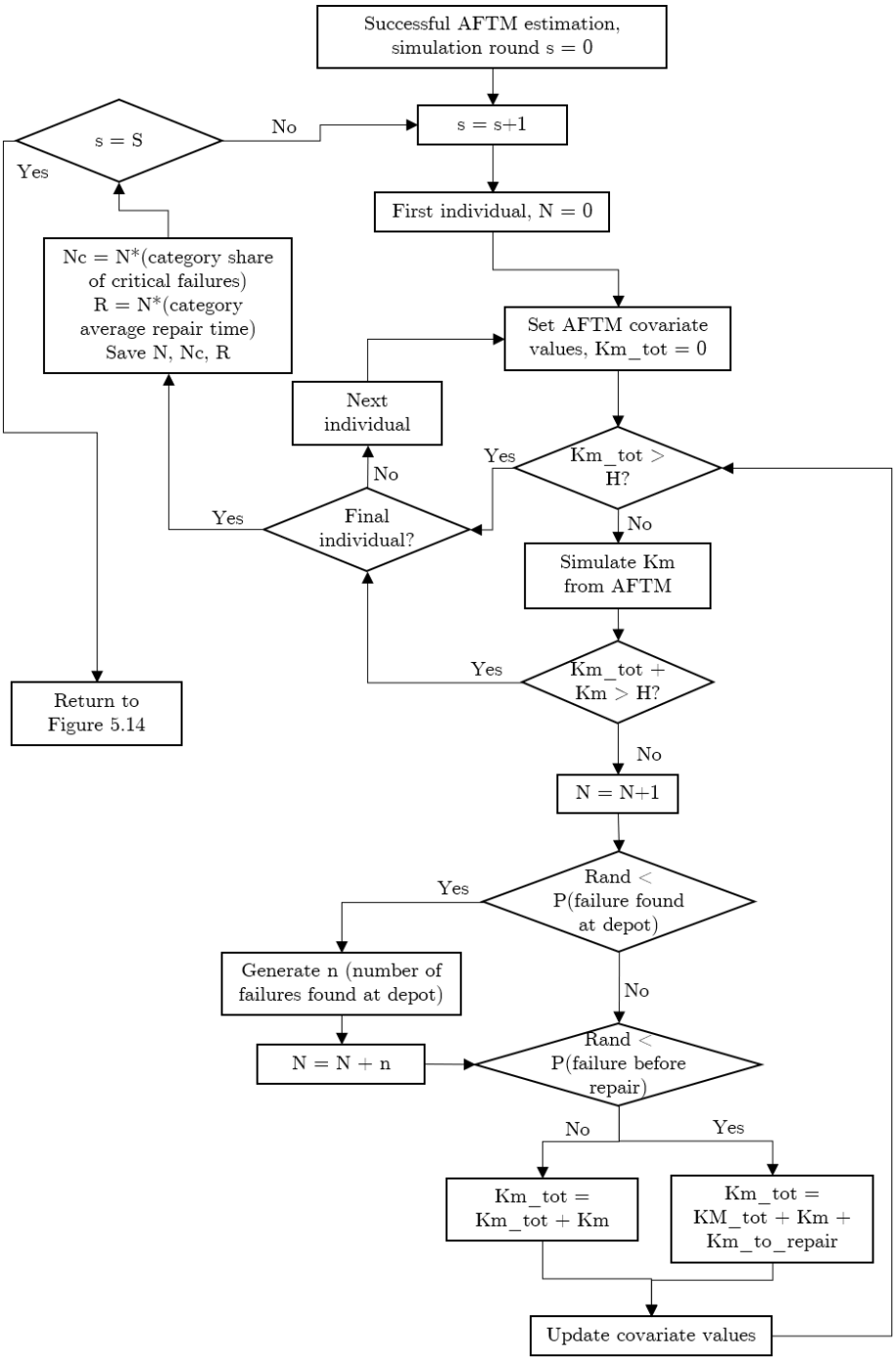


Figure 5.15: A flowchart of the simulation of failures for a failure category to which an AFTM is a good fit. The flowchart starts and continues in Figure 5.14. N corresponds to the number of failures, H is the prediction horizon in kilometres and S is the number of simulation rounds to be performed.

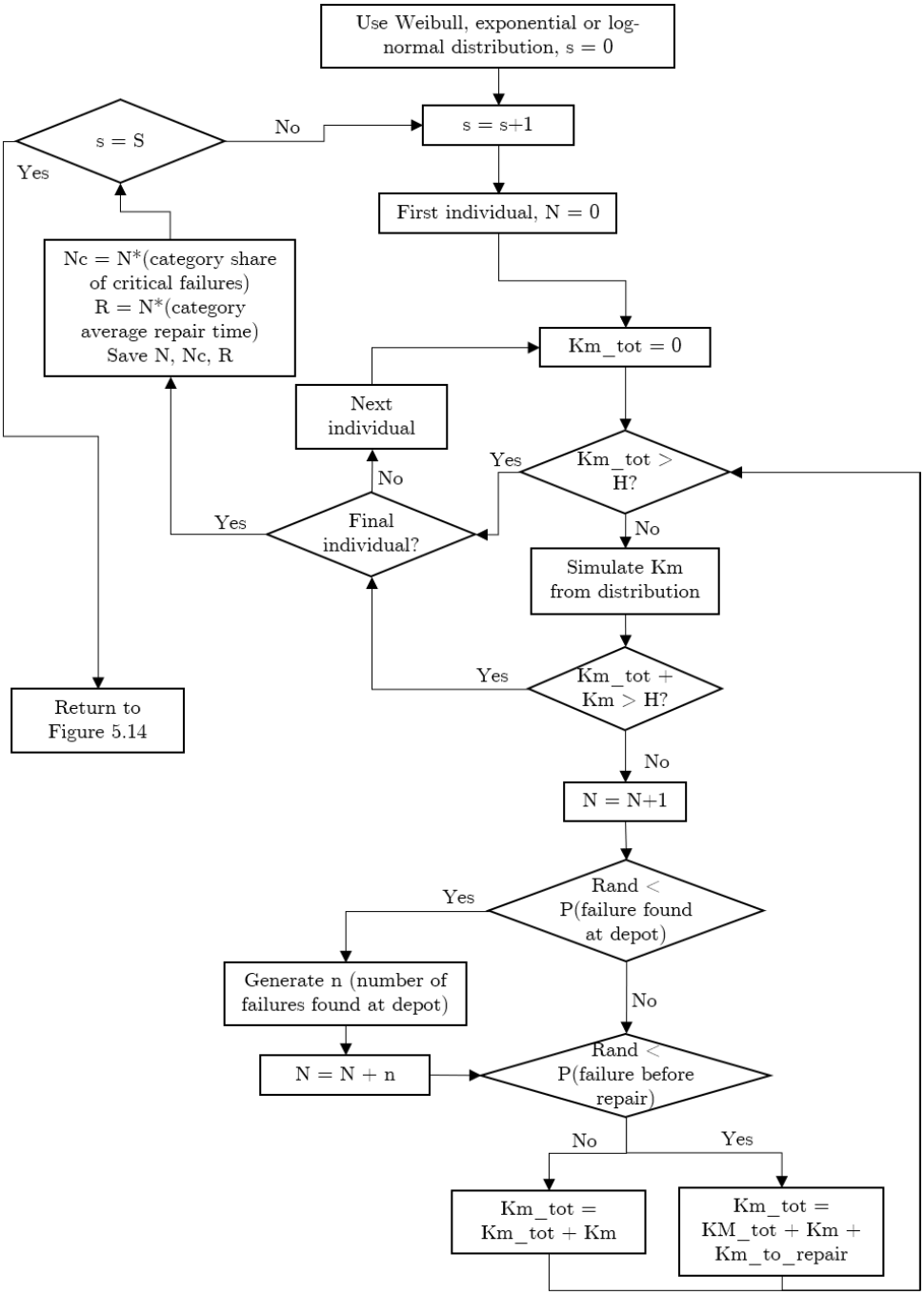


Figure 5.16: A flowchart of the simulation of failures for a failure category to which a regular distribution is the best fit. The flowchart starts and continues in Figure 5.14. N corresponds to the number of failures, H is the prediction horizon in kilometres and S is the number of simulation rounds to be performed.

Chapter 6

Results and Discussion

This chapter presents the results of a validation test, in which the model's prediction accuracy is evaluated by comparing the predicted number of failures to the realised number of failures in a test dataset. Reasons for problems with the prediction accuracy are reviewed and ideas for further model improvement are discussed. This chapter also summarises the identified applications for the developed failure prediction model in the maintenance operations at VR and discusses limitations in the use of the model in these application areas.

6.1 Prediction Accuracy

To test the prediction accuracy of the failure prediction model we construct two datasets: one training dataset for model parameter estimation and one test dataset for comparing the realised number of failures with the predicted number. We want to test the accuracy of short-term predictions as well as long-term predictions. Hence, we predict the number of failures for an entire year, on a quarterly basis for four quarters, on a monthly basis for twelve months, and also on a weekly basis for eight weeks. For the weekly predictions we also look at the predicted repair times and the number of critical failures. After presenting the results, we discuss problems with the prediction accuracy and what they may result from.

6.1.1 Yearly Prediction

Failure category specific results for a yearly perdiction are presented in Table 6.1. Both the predicted number of failures as well as the realised number of failures are presented. The Prediction/Reality-column shows the percentage of overestimation or underestimation of the number of failures. The confidence interval (CI) of the prediction of the combined number of failures in all categories is also presented. The kilometres since the previous failure are not considered.

Table 6.1: The predicted number of failures for a year for each failure category, as well as the realised number of failures during the prediction period. The Prediction/Reality-column shows the percentage of model over- or underestimation. The confidence interval for the total predicted number of failures is given.

Failure category	**Predicted failures**	**Lower CI**	**Upper CI**	**Realised failures**	**Prediction/ Reality**
B	290			251	15%
C	547			487	12%
D	493			375	17%
E	755			1 137	-34%
F	1 037			1 169	-11%
G	530			422	26%
H	344			276	25%
J	840			904	-7%
K	410			322	27%
L	432			321	34%
M	413			226	83%
N	236			193	22%
P	0			26	
Q	704			682	3%
R	828			986	-16%
S	263			309	-15%
T	416			305	36%
U	0			4	
Total	**8 481**	**8 322**	**8 665**	**8 395**	**1%**

For the one-year prediction period the model predicts the total number of failures to be 8 481. The confidence interval for this prediction is [8322, 8665]. In the test dataset the realised number of failures for the year is 8 395, which is very close to the predicted number of failures.

We can see that failure categories F, J and Q are quite accurately predicted. The number of failures in categories M, T and L are greatly overestimated, whereas the number of failures in category E are greatly underestimated. However, the combined number of failures is accurately predicted with only a 1% overestimation, and the realised number of failures falls within the prediction's confidence interval. The accurate prediction cannot however be trusted, since the prediction accuracies for all failure categories are not good. The overestimation of some category failures combined with the underestimation of failures in other categories happen to result in a number of failures close to the realised number of failures.

The results would suggest that the AFTMs for categories F, J and Q succeed in modelling the failure process well, whereas the models for other categories are not as good. Categories F, J and Q are large failure categories, hence the datasets for these failures are large. A larger dataset can be assumed to provide a more accurate model, as possible outliers in the data do not affect the model significantly. Failure category E is also large, but the AFTM for this category does not provide accurate predictions. This might result from the failure process for this category being more random, which prevents the finding of relevant factors affecting the time to failure.

As some of the failure category models are clearly lacking in prediction accuracy, we should not consider these in the predicting of failures before improving them. These models should be reviewed and other methods for predicting the number of failures in these categories can be sought.

We choose to have a closer look at the prediction result of the failure categories which have been over- or underestimated by a percentage in the range [-20%, 20%]. This means considering failure categories B, C, D, F, J, Q, R and S. The combined predicted number of failures for these categories is 4 947, whereas the realised failures are 5 163. This indicates an underestimation of 4%, which can be considered an accurate prediction.

6.1.2 Quarterly Prediction

The quarterly realised and predicted number of failures with confidence intervals are presented in Table 6.2. A percentage indicating the prediction's over- or underestimation is also presented. The predictions are carried out separately for each quarter and the data used for model estimation is updated for each prediction round so that it consists of data from two years backwards from the starting date of the prediction period. Also, the kilometres since the previous failure are not considered.

Table 6.2: The predicted number of failures per quarter with confidence intervals, as well as the realised number of failures during the prediction period. The Prediction/Reality-column shows the percentage of model over- or underestimation.

Quarter	Predicted failures	Lower CI	Upper CI	Realised failures	Prediction/ Reality
Q1	2 326	2 255	2 446	2 513	-7%
Q2	2 513	2 373	2 577	1 714	47%
Q3	2 094	1 998	2 202	2 026	3%
Q4	2 014	1 922	2 142	2 142	-6%
Total	**8 946**	**8 548**	**9 367**	**8 395**	**7%**

Predicting the number of failures per quarter gives an overestimation of 7% on the number of failures for the whole year. The predictions for the third and the fourth quarters are quite accurate and the realised number of failures for these quarters fall within the confidence interval of the predictions. The prediction for the first quarter is also pretty good, even though the number of failures are slightly underestimated. For the second quarter the prediction overestimates the number of failures greatly. It seems as if the model does not succeed in accurately capturing the seasonal variation in the number of failures between quarters.

Due to the problem with varying prediction accuracy between failure categories, the quarterly predictions presented here should not be trusted without examining the category specific predictions. We can see that the sum of the quarterly predictions result in a larger number of predicted failures for the whole year compared to when predicting the whole year at once. The quarterly predictions are conducted separately, starting over in the beginning of each quarter and not considering the generated failure kilometres that go beyond the prediction period of three months. This means that failure kilometres are generated for all rolling stock individuals using the season covariate values of January, April, July and October, as these months start each quarter. In the case of simulating the whole year at once, it might be that these months are skipped for most individuals. If a failure kilometre corresponding to three months is generated in March, the next failure is generated in June. Hence, no failure is generated using April's covariate values. If the three month failure kilometre is generated in March in the case of a quarterly prediction, a new failure is still generated in April as the next quarter begins. This shows that the length of the prediction period and the start date of the period have an effect on the predicted number of failures.

6.1.3 Monthly Prediction

Table 6.3 presents the monthly predicted number of failures with confidence intervals and the monthly realised number of failures. The table also presents the percentage of the prediction's over- or underestimation. The predictions are carried out separately for each month and the data used for model estimation is updated for each prediction round so that it consists of data from two years backwards from the starting date of the prediction period. For example, the prediction of the number of failures in July (2018) is based on a model estimated using data from July 2016 to June 2018. The kilometres since the previous failure are not considered.

Table 6.3: The predicted number of failures per month with confidence intervals, as well as the realised number of failures during the prediction period. The Prediction/Reality-column shows the percentage of model over- or underestimation.

Month	**Predicted failures**	**Lower CI**	**Upper CI**	**Realised failures**	**Prediction/ Reality**
January	844	783	897	1 031	-18%
February	838	792	898	813	3%
March	918	873	987	669	37%
April	934	903	978	583	60%
May	866	815	916	597	45%
June	760	670	822	534	42%
July	764	696	827	692	10%
August	881	828	935	681	29%
September	751	702	831	653	15%
October	737	672	786	718	3%
November	752	701	814	654	15%
December	704	674	734	770	-9%
Total	**9 750**	**9 109**	**10 425**	**8 395**	**16%**

We notice that even if February, September and December are quite accurately predicted, there are large difference between the prediction and the realised number of failures for the other months. This is especially clear in Figure 6.1, which graphically presents the values of Table 6.3. The figure also shows the number of failures per month in the dataset used for estimating the model. It seems that the prediction model cannot capture the seasonal variations in the number of failures to an extent large enough, which also was noticed for the quarterly predictions. Taking the average of the historical number of failures would be a better prediction for the number of failures

for most months.

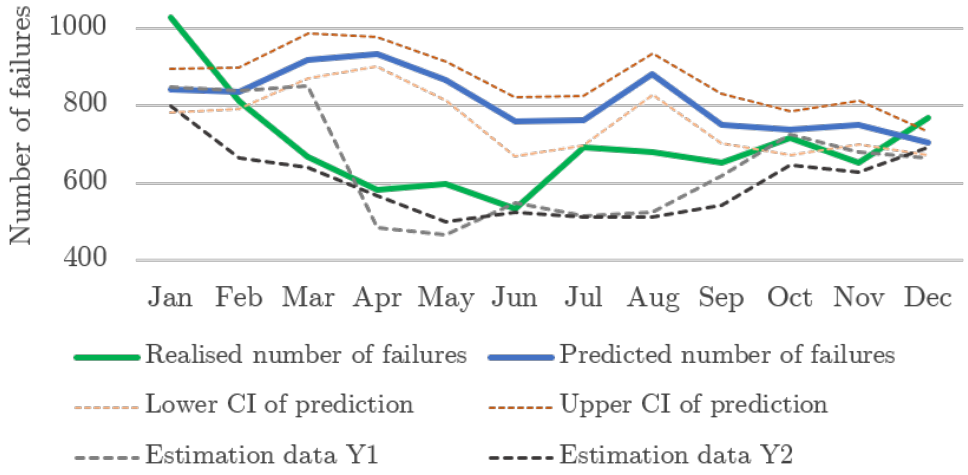


Figure 6.1: The number of realised and predicted failures for 12 months. The prediction's confidence interval (CI) and the historical number of failures from both years (Y1 and Y2) in the estimation data are presented as well.

As the distribution of the failure times in an AFTM depends on the covariate values, the distribution changes over time. This causes some problems with combining the predicted number of failures for each month into a prediction of the total number of failures for the whole year, as discussed with the quarterly predictions. In the case of the failure times following a regular exponential distribution, the combined monthly predictions would result in the same yearly prediction as when predicting the yearly failures at once. This is due to the exponential function being memoryless. With the AFTMs in this thesis this is not the case as the failure times follow different Weibull distributions at different times.

To further illustrate the problem with the seasonal variation we present the monthly predictions for failure categories F, M and R in Figures 6.2, 6.3 and 6.4. Failure category F is very accurately predicted, and the combined yearly difference between the total realised number of failures and the total predicted number of failures is only -3%. The model predicts the number of failures better than using the average number of historical failures as the prediction. Failure category M on the other hand is extremely poorly predicted by the model. The difference in the yearly realised failures and predicted failures is +203%. The seasonal variations are not captured at all for the M-failures, and the peaks in April and August are probably due to problems

with the AFTM's season covariate. For this category the historical monthly averages would be a more accurate prediction on the number of failures.

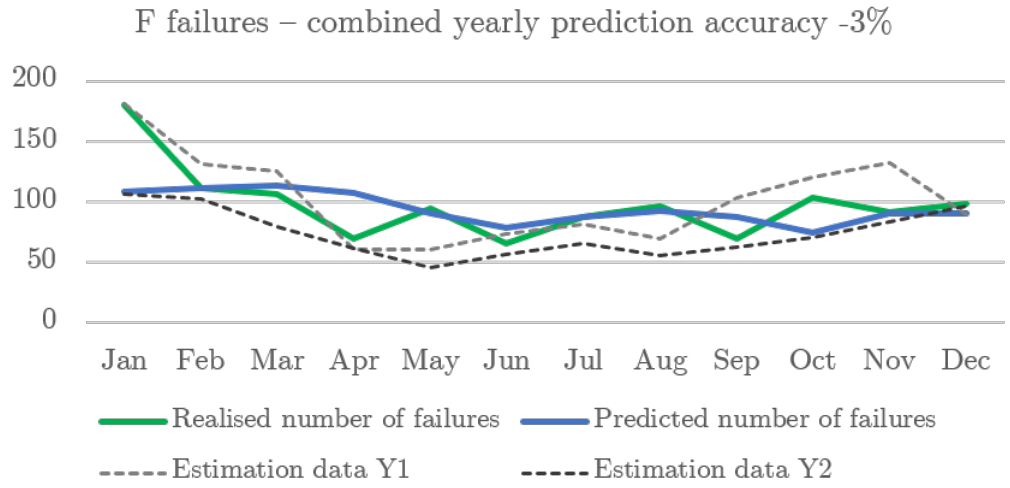


Figure 6.2: Monthly predicted and realised number of failures for failure category F. The number of failures in the estimation data are presented as well.

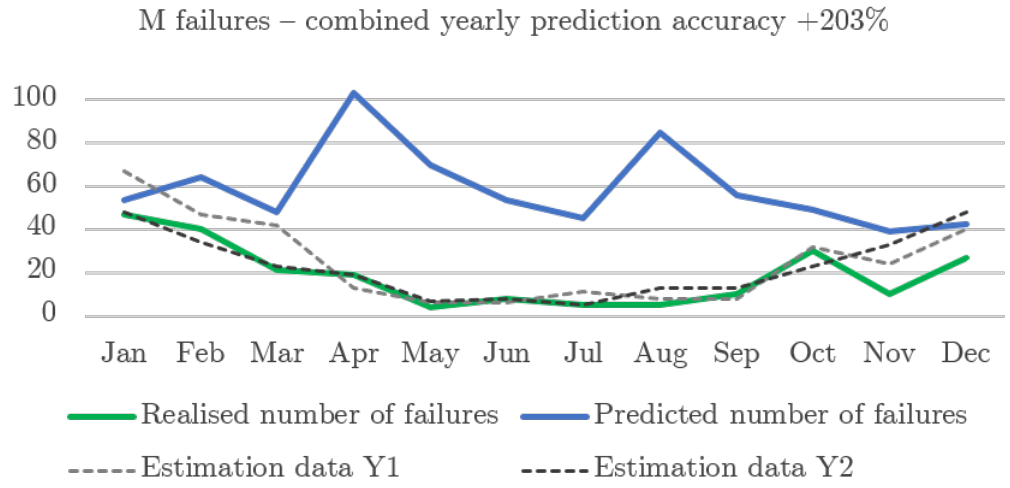


Figure 6.3: Monthly predicted and realised number of failures for failure category M. The number of failures in the estimation data are presented as well.

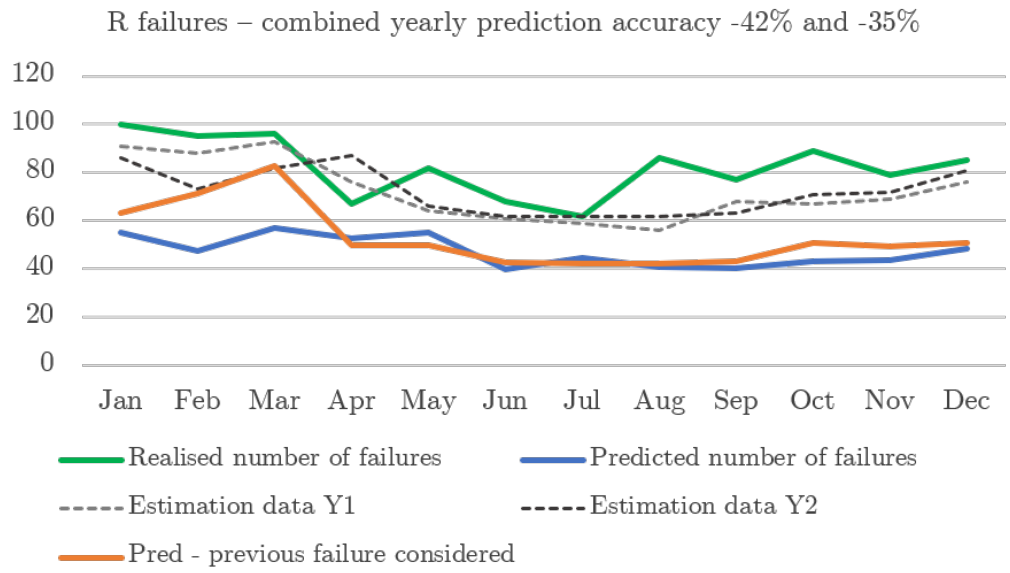


Figure 6.4: Monthly predicted and realised number of failures for failure category R. The predicted number of failures given that the previous failure is taken into account in the simulation are also presented. The number of failures in the estimation data are presented as well.

In contrast to the large overestimation for category M, the number of failures for category R are consistently underestimated and results in a combined yearly underestimation of 42%. For this category, using historical monthly average as the predicted number of failures would result in a better prediction, even though it would also underestimate the number of failures.

Based on the histogram of category R failures presented in Figure 4.8b, the average inter-failure kilometres for failure category R are long. This combined with the information that R failures occur on average every 48 days (see Table 5.5) would suggest that the reason for the poor prediction accuracy of the monthly failure predictions for category R is due to the short prediction period. As the model starts simulating failures with the assumption that a failure has just occurred, enough R category failures do not have time to occur in a prediction period of only 30 days. Hence, Figure 6.4 also presents the predicted failures for category R given that the simulation model takes the kilometres since the previous failure into account when generating the first failure kilometre. In this case the first failure kilometre is simulated with a season covariate value based on when the previous failure was repaired. The model generates failure kilometres until the generated kilometre exceeds the number of kilometres since the previous repair. This corresponds to the

conditional probability of failure given that the rolling stock has survived until this point without a failure occurring. We can see that for category R this simulation approach results in an improved prediction accuracy for January, February and March. This yields a combined yearly prediction accuracy of -35%. However, we would expect the combined results to be better when the kilometres since the previous failure are taken into account in the simulation. The yearly underestimation when simulating the whole year at once was only -16% (see Table 6.1). It might be that a prediction period of one month is too short to accurately predict R category failures, even if the history is taken into account.

6.1.4 Weekly Prediction

Table 6.4 presents the prediction and realisation of the number of failures on a weekly basis for weeks 1-8 in 2019. The predicted and realised number of failures are presented in Figure 6.5 as well. The predictions are carried out separately for each week and the data used for model estimation is updated for each prediction round so that it consists of data from two years backwards from the starting date of the prediction period. The kilometres since the previous failure are not considered. The predictions for weeks 1, 2, 7 and 8 are very accurate. For the combined eight-week period, the total number of failures is however underestimated by 12%.

Table 6.4: The predicted number of failures per week with confidence intervals, as well as the realised number of failures during the prediction period. The Prediction/Reality-column shows the percentage of model over- or underestimation.

Week	Predicted failures	Lower CI	Upper CI	Realised failures	Prediction/ Reality
1	185	161	227	185	2%
2	181	161	206	190	-5%
3	180	144	208	237	-24%
4	197	174	237	271	-27%
5	197	166	219	241	-18%
6	203	173	228	249	-19%
7	197	174	217	190	4%
8	200	171	232	194	3%
Total	**1 543**	**1 324**	**1 774**	**1 757**	**-12%**

The graph in Figure 6.5 shows that the realised number of failures fluctuate

largely between weeks, whereas the predicted number of failures is more stable over time. This illustrates a limitation with the prediction model: it does not succeed in capturing variations in shorter time intervals. Again, it should also be noted that due to the varying failure category prediction accuracies, the predicted number of failures should not be considered accurate without closer examining the failure category specific failures. Using the historical weekly average as the predicted number of failures would yield approximately the same result as the prediction model.

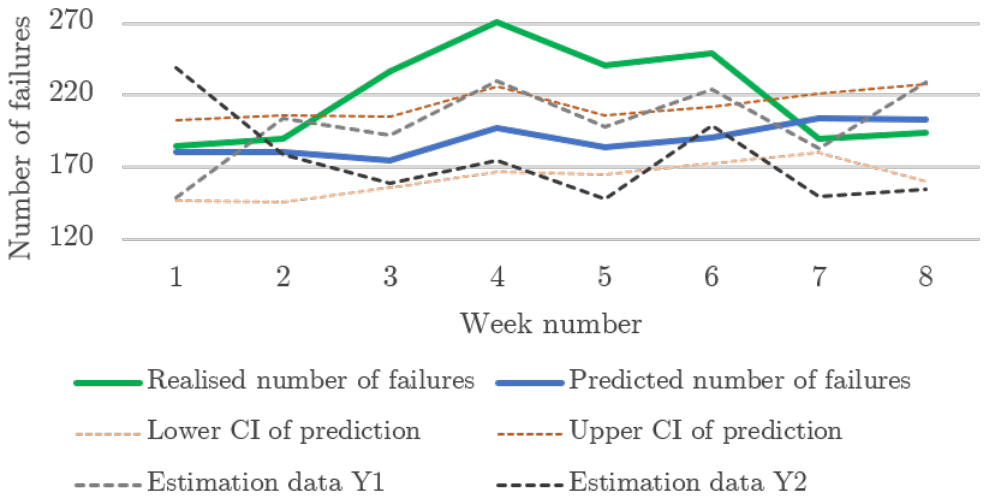


Figure 6.5: Graphical presentation of the number of realised failure and the number of predicted failures for weeks 1-8 in 2019. The prediction's confidence intervals (CI) and the historical number of failures in the estimation data are presented as well.

For the weekly predictions we also look at the accuracy of the predicted number of critical failures, as well as the accuracy of the total repair times. These results are presented in Tables 6.5 and 6.6, respectively. Due to the way the critical failures are calculated in the prediction model, the number of critical failures is quite constant for all prediction weeks. The number of critical failures is just a percentage of the total number of predicted failures, which is based on the historical share of critical failures. The prediction model does not succeed in capturing the large variations between weeks in the number of critical failures. This might be a limiting factor for the short-term maintenance scheduling utilising the model.

The total repair time for the eight weeks combined is underestimated by 38%. The realised fluctuations in the repair times between weeks are large

and are not captured by the model. The weekly under- or overestimation of the repair times vary between -57% and 5%.

We must however note that the predicted repair times and number of critical failures suffer from the same lack of accuracy as the predicted number of failures, due to the repair times and critical failures depending on the predicted number of failures. Hence, if the number of failures is not accurately predicted this affects the accuracy of the repair time and critical failure predictions as well.

Table 6.5: The predicted number of critical failures per week with confidence intervals, as well as the realised number of failures during the prediction period. The Prediction/Reality-column shows the percentage of model over- or underestimation.

Week	**Predicted critical failures**	**Lower CI**	**Upper CI**	**Realised critical failures**	**Prediction/ Reality**
1	2.4	1.9	2.9	2	18%
2	1.6	1.3	1.9	3	-46%
3	1.6	1.1	2.0	1	57%
4	1.8	1.6	2.3	3	-39%
5	1.9	1.5	3.4	7	-72%
6	1.9	1.6	2.3	6	-68%
7	1.8	1.5	2.2	2	-9%
8	1.9	1.5	2.6	0	100%
Total	**15.0**	**11.9**	**19.5**	**24**	**-38%**

Table 6.6: Weekly predictions of the total repair times expressed in hours for repairing all predicted failures (presented in Table 6.4).

Week	**Predicted repair time** (h)	**Lower CI**	**Upper CI**	**Realised repair time**	**Prediction/ Reality**
1	1 213	1 016	1 455	1 892	-36%
2	1 159	1 046	1 320	1 935	-40%
3	1 154	917	1 351	1 880	-39%
4	1 274	1 153	1 518	2 968	-57%
5	1 268	1 058	1 429	2 638	-52%
6	1 300	1 099	1 472	1 989	-35%
7	1 265	1 120	1 406	1 343	-6%
8	1 287	1 115	1 520	1 230	5%
Total	**9 920**	**8 524**	**11 471**	**15 875**	**-38%**

A prediction period of only seven days is short. Hence, we would expect that the timing of the previous failure would have a great effect on the weekly failures. We want to examine this by predicting the weekly failures so that the situation in the beginning of the prediction period is considered, i.e. the kilometres since the previous failure are taken into account in the simulation of failures. These results for weeks 1-8 2019 are presented in Figure 6.6. We can see that the prediction accuracy is lower and that the predictions are more stable between weeks, compared to the predictions presented in Figure 6.5 where the previous failures have not been considered in the simulation. This is a surprising result, as we would expect the considering of the current situation to improve the prediction accuracy. The short prediction period might be the reason for the inaccurate prediction. Also, the failure category specific results should be reviewed to see if any specific failure categories are significantly over- or underestimated.

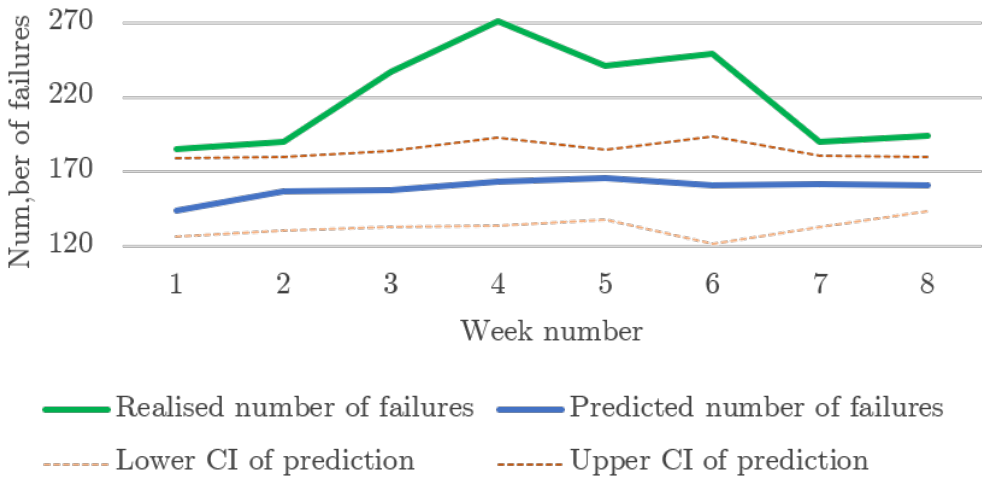


Figure 6.6: Weekly predictions on the number of failures when kilometres since the previous failures have been considered.

6.1.5 Comments on the Prediction Accuracy

We can conclude that there are problems with the prediction accuracy and we notice that the predictions vary depending on the length of the prediction period. The prediction accuracy also differs between periods, e.g. between months. Especially for some failure categories the lack of prediction accuracy is significant. For example, the number of category R failures are consistently

underestimated, but the yearly prediction is better than the monthly predictions. In contrast, category F failures are accurately predicted on both a yearly and a quarterly basis. We identify a number of possible reasons for the lack of prediction accuracy in some failure categories:

- The failure process is partly random and hence cannot be accurately modelled.
- Covariates identified as relevant for the failure process do not describe the failure process well enough.
- The failure categories are too broad.
- The quality of the data is not good enough.
- The data used for model parameter estimation is from a too short time period.

A key question in the evaluation of the model's prediction power is the randomness of the failure processes of the rolling stock. If the failures to a large extent occur randomly, they cannot be assumed to be altogether accurately modelled. The concordance indices presented in Table 5.5 portray the AFTMs' descriptive accuracy of the historical data; the low index value of 0.57 for failure categories D, H and J indicates that these models do not describe the data very well, which may be due to randomness of the failure processes. We would hope for concordance indices of 0.7, as presented in Section 5.4.1.

The low concordance indices can also be due to including the wrong covariates in the AFTMs. In this thesis we examine the effect of five factors on the failure process. There might however be additional factors better describing the failure process and the potential lack of relevant covariates might very much affect the prediction accuracy of the model. For example a weather covariate could have a significant impact on the time to failure. The ability to consider different types of weather in the failure prediction would be especially useful for budgeting purposes, where scenario analysis is relevant. Comparing the difference in number of predicted failures for a bad year weatherwise to that of a good year weatherwise, would allow for the comparing of worst case scenario versus best case scenario for the next year's total number of predicted failures. A weather covariate might also better capture the seasonal fluctuations in the number of failures.

The lack of seasonal variations in the predictions is a notable issue with the current model. The results imply that the seasonal covariate does not succeed

in capturing the variations. Hence, investigating alternative covariates for including the seasonal variations is encouraged. The main issue with the current seasonal covariate is that it is specified for the month the failure time is generated in. However, the effect of the seasonal variate in the data is based on the when the failure occurs. For example, if a failure is noticed in February, it contributes to the season covariate value in February. However, as the average inter-failure times are longer than a month for all failure categories, the effect of the season is delayed in the simulation model. Instead of basing the season variate on the historical failure intensity of the month, the season covariate should consider the failure intensity of upcoming months. This could be done with a combined probability distribution which considers the conditional probabilities of failure in upcoming months, given that the individual has survived until then. This way, simulating a failure time in February would not solely be based on the historical average of February's failure intensities, but rather include the failure intensities of all other months as well.

The model's broadness might be another explanation for the inaccurate failure predictions. As the model gives a prediction for all failure types, the failure type specific models have not necessarily been calibrated as well as they could have. This was clearly visible when comparing the yearly failure category specific predictions. The failure categories lacking in prediction accuracy should be developed further. This could be done by including other covariates in the AFTM to better describe the failure process, or using a completely different approach to the modelling of these failures. Also, developing component specific prediction models would presumably improve the prediction accuracy, but this requires component specific failure data. A problem in the current category specific models is that the occurrence of one failure might not affect the occurrence of the next failure in the same category. With component specific failure models this issue would be eliminated. In addition to the need for large enough component specific failure datasets, the number of models to be developed would be huge as a rolling stock consists of a lot of components.

We noticed that the length of the prediction period affects the prediction accuracy of some failure categories, which is probably due to the seasonal covariate value. The monthly predictions for category R were worse than when predicting the whole year at once. We tested if taking into account the kilometres since the previous failure would improve the monthly prediction accuracies for type R failures, and some improvement could be seen. By testing prediction periods of different lengths and with different starting dates, we could identify the months when the seasonal covariate values yield good

results and what length on the prediction period gives the most accurate prediction. This information could be used in the further improvement of the prediction model. Also, for all failure categories we should test what effect the considering of the current situation in the simulation of failures has on the prediction.

As discussed in Section 4.1, there are some limitations with the failure and maintenance records. These limitations might affect the prediction accuracy of the model. Especially the inconsistency with reporting of failures results in the model output not necessarily matching the reality. As the failures are not always reported as they are found, the model predicts the reporting of a failure rather than the actual occurrence of a failure. The unspecific repair lead times result in the prediction of the total failure repair time being uncertain. The failure and maintenance records should be improved and checked for accuracy to guarantee a more reliable model.

Another limiting factor might be the length of the period the failure and maintenance records are from. The data used for estimating the model parameters is only from a two-year period. The two-year period might be too short for developing an accurate failure prediction model, as variations between years might be large and are not necessarily evened out with data from only two years. The weather might be a factor significantly affecting the number of failures, which adds to the variation of the number of failures between years. Including a longer time period in the data for estimating the failure prediction model parameters would probably even out this type of yearly variations.

6.2 Model Applications

The developed prediction model presented in Chapter 5 produces a prediction on the number of failures and the total repair times for all failures during a given time period. The number of critical failures and their repair times are estimated as well. The critical failures are specified, as these affect the short-term maintenance scheduling process significantly.

Based on interviews with key maintenance personnel at VR, we have identified several application areas in the maintenance operations for the failure prediction model. These application areas are presented in Figure 2.1:

- Short-term maintenance planning tool to give a prediction on the need for repair during the next week.

- Monthly and yearly prediction of repair needs with confidence intervals are helpful in the budgeting process.
- The model facilitates the understanding of the failure process for a rolling stock fleet, which is helpful for the fleet engineers, especially when developing the preventive maintenance programme.
- The model helps to identify rolling stock individuals performing significantly worse than the rest of the fleet, providing an opportunity to more closely examine those to improve their performance.
- In the optimising of opportunistic maintenance, a model for failure prediction serves as a building block for the optimisation model.

For the short-term maintenance planning process, more specifically for the scheduling of next week's maintenance actions, the failure prediction model provides a prediction of the number of failures for the next week. An accurate failure prediction allows for the scheduler to take these into account when scheduling next week's preventive maintenance actions. For the maintenance scheduler, the predicted number of critical failures are especially interesting. The critical failures normally cause changes in the planned maintenance schedule, as these prevent the use of the rolling stock and need to be repaired as soon as possible. Hence, some preventive maintenance actions might be rescheduled due to the need for repair of the critical failures. If the maintenance scheduler can schedule an accurate amount of time for repairing predicted critical failures, the need for rescheduling might be reduced. The total repair times for the predicted failures might also be of interest for the maintenance scheduler.

Even though some category specific predictions were found lacking in Section 6.1, we would still suggest that the prediction model is included in the weekly maintenance scheduling for a test period to further test the accuracy of the information it provides and to determine its usefulness. The predicted number of failures was found more accurate than the predicted repair times and number of critical failures. The schedulers can also use the model to predict those categories which have a better model, such as category F failures. However, the schedulers are rather interested in the total number of failures than failures of a specific type.

A limiting factor for the utilisation of the prediction model for predicting the number of critical failures is the manner in which the model calculates them. The number of critical failures is just a percentage of all predicted

failures based on the historical share of critical failures. To improve the accuracy of the predicted number of critical failures, a failure prediction model focused solely on critical failures could be developed in the same manner as the model presented in this thesis. However, only approximately 1% of the overall failures are critical, which correspond to just over 200 critical failures during the two-year period the data covers. This is too small a dataset for estimating a failure prediction model per failure category. Hence, the developed model's critical failure prediction is currently the best prediction at hand. In Section 6.1 we found that the number of realised critical failures has a large variation between weeks, but the prediction model does not succeed in catching this variation. The predicted number of critical failures is expected to be more accurate over a longer prediction period, given that the total number of failures are accurately predicted.

The budgeting for the following year is done in autumn. The current budgeting process relies on historical number of failures, hence an accurate failure prediction would presumably improve the budgeted number of failures. By altering the number of kilometres to be driven during the following year or by altering the covariate values, the failure prediction model can be used as a tool for scenario analysis. The opportunity to examine different scenarios in the budgeting process is probably the greatest benefit the prediction model offers for the process. As concluded in Section 6.1, some of the failure category specific models are not that accurate on a monthly basis. Hence, relying on the monthly failure predictions in the budgeting of the total number of failures is not to recommend. However, the number of failures in categories with a better prediction accuracy, e.g. F failures, can be predicted on a monthly basis and used in the budgeting process, whereas the budgeted number of failures in other categories can be based on historical values. This approach with combining a prediction and historical values is likely to result in a more accurate budget than relying solely on historical data.

For the fleet engineers working with improving the reliability and availability of a specific fleet, the prediction model's structure gives insight into the failure process for a specific fleet and failure category. For them the predicted number of failures is not necessarily of interest, but rather the effects the covariates have on the failure process. They might be interested in what type of failures occur more often during the winter or if the rolling stock is driving at a higher speed. Identifying fleet individuals performing worse than average is also of interest for the fleet engineer. The identified "bad" individuals can be examined more closely to improve their performance. For starters, the fleet engineers can utilise those category specific models which were found to have a good prediction accuracy. The engineers might be able

to provide additional insight into the failure processes of the failure categories for which accurate enough models have not yet been developed. The insight can be used in the further improvement of these models.

For developing an optimal preventive maintenance programme, a key is to be able to predict failures. Also, the prediction model must be able to model the preventive maintenance actions' effect on the failure process, which can be done with AFTM. Hence, the developed failure prediction model can be utilised in the development of an optimal preventive maintenance programme, should the fleet engineers want to thoroughly examine and optimise the current programme. However, as the model does not portray the actual failure process very well due to the lack in prediction accuracy, utilising the model in its current form in the optimising of the preventive maintenance programme, might result in a programme that is not optimal for the actual failure process.

The model can be integrated to a model for finding an optimal opportunistic maintenance strategy. The strategy would specify what maintenance actions are optimal to perform together. A model for predicting upcoming failures is needed in the development of a model for opportunistic maintenance optimisation. Hence, the failure prediction model serves as a building block for the improvement of opportunistic maintenance at VR. However, as with optimising the preventive maintenance programme the limited prediction accuracy of the failure prediction model might lead to problems in the optimisation of the opportunistic maintenance strategy. If the failure prediction model cannot capture the real failure process, the found strategy might not be optimal.

The developed prediction model serves as a good basis to build upon and develop further. Some failure categories predict the number of failures accurately, while the models for other categories should still be improved. As mentioned, improvements can be made e.g. by adding relevant covariates to the model, or by considering a dataset from a longer time period in the estimation of the model parameters. The model can, for example be developed to include censor data, hence simultaneously contributing to the further automation of condition based maintenance. A completely different approach than the AFTM might also be better suited for the modelling of some categories. As the application areas of a failure prediction model in maintenance operations at VR are many, the further improvement of the prediction model is strongly encouraged. However, if the suggested improvement actions do not increase the prediction accuracy, it might be that some failure category failures happen too randomly to be accurately modelled.

Chapter 7

Conclusions

This thesis developed a failure prediction model for the Finnish rolling stock maintenance company VR Maintenance Ltd. First, information on the current maintenance operations was collected and utilisation possibilities for a failure prediction model were investigated. Then, a literature review on methods for failure modelling was conducted. Based on the available maintenance and failure records and the findings of the literature review, an accelerated failure time model (AFTM) was selected as the method for modelling failures at VR. The failure and maintenance records from a two-year period were analysed and five factors were identified to have an effect on the failure process: the month, the average speed, the kilometres since previous preventive maintenance action, the age of the rolling stock expressed in total kilometres driven, and the split between "good" and "bad" rolling stock individuals. An AFTM with relevant factors was developed separately for each failure category for each rolling stock fleet. The developed prediction model provides a prediction on the number of failures and the total repair time for all failures during a specified prediction period. Additionally, the model specifies the predicted number of critical failures.

Based on interviews with key maintenance personnel at VR, the utilisation of a failure prediction model was found to enhance several elements of the maintenance operations at VR. For short-term maintenance scheduling, the failure prediction model provides a prediction on the number of failures for the next week, which can be utilised in the scheduling of preventive maintenance actions. Especially the number of critical failures was found to be of interest. In long-term maintenance planning the prediction model can be utilised in the process of developing the preventive maintenance programme. Also, it helps to identify rolling stock individuals performing significantly

worse than the rest of the fleet. These "bad" individuals can then be examined more closely to improve their performance. The budgeting process was also found to benefit from a failure prediction model. Rather than basing next year's budget solely on the historical number of failures, the model's prediction on the number of failures can be used. As the covariate values in the prediction model can be altered, the model can also be used for scenario analysis in the budgeting process.

Additionally, the model was found to help further future maintenance operation improvement projects, such as the development of an opportunistic maintenance model or the further automation of the condition based maintenance. In the development of these, the failure prediction model was found to serve as an important building block.

Some of the failure categories could be modelled successfully with an AFTM and provided accurate predictions, whereas there was a lack of prediction accuracy for other categories. The accuracy of the predicted total number of failures was however found inadequate. Even though the combined prediction of the number of failures in all failure categories gave a yearly prediction result very close to the number of realised failures, the category specific predictions were not all accurate enough. The number of failures in some categories were overestimated, whereas the number of failures was underestimated in other. This resulted in the over- and underestimations cancelling out each other's effect, which lead to a false accurate number of total failures. This finding indicates that the model cannot be used in its current form to predict the overall number of failures. Some of the failure categories were accurately predicted and the failures in those categories can be predicted using the developed models. The models for categories performing worse in the prediction of failures should be further investigated and improved.

The models can be improved by including additional covariates in the AFTM, for example by adding a covariate describing the weather condition. A problem with the AFTMs performing poorly was that the seasonal variations were not accurately captured. The season covariate used is indeed questionable, since it considers the month of the previous failure in the generating of the next failure. Instead, the seasonal effect of future months should be considered, since a failure is probable to occur in a later month due to the average time to failure being longer than a month in all failure categories. As we cannot know beforehand when the failure will occur a combination of conditional probabilities could be used to capture the seasonal effect of future months. Updating the season covariate in all AFTMs to consider future

failure probabilities would most likely improve the prediction accuracy.

Another improvement action would be to consider data from a longer time period in the model estimation. The data was also found partly lacking in quality and improving the quality of the data might improve the model's prediction accuracy. Additionally, component specific models could give more accurate predictions, given that there is enough component specific data for the estimation of a model. If these actions do not improve the prediction accuracy of the AFTM, a completely different approach to the modelling of that category's failures might be needed. The literature study in Chapter 3 presented different methods for failure modelling. A basic exponential, Weibull or log-normal distribution might be a suitable option, and would be easily implemented. Another easy approach is to use the average historical number of failures from previous years during the prediction period. This method was found quite accurate for monthly category M failures, which were greatly overestimated by the AFTM. A model based on Bayesian theory could be considered if some external effect is to be included in the prediction. In that case a proportional hazard model, which is a survival regression model similar to the AFTM, could also be considered.

We can conclude that some failure categories could be accurately predicted with an AFTM, whereas other categories could not. The further improvement of the category specific prediction models is encouraged, since succeeding in developing an accurate failure prediction model for all failures would enhance several elements of the maintenance operations at VR. This would improve reliability and availability of the rolling stock and reduce maintenance costs. We suggest that the model is included in the weekly maintenance scheduling as a pilot study, as this area of the maintenance operations is expected to profit most from the prediction model. The category specific predictions should be examined separately, so that the actual prediction accuracies can be evaluated without over- and underestimations cancelling each other out. The pilot study would provide information on the usefulness of the model and its overall ability to provide accurate insights. Improvements to the model could be done based on these pilots, ultimately making it part of standard maintenance operations at VR.

Bibliography

Ahmad, R. and Kamaruddin, S. (2012). An overview of time-based and condition-based maintenance in industrial application. *Computers & Industrial Engineering*, 63(1):135–149.

Alaswad, S. and Xiang, Y. (2017). A review on condition-based maintenance optimization models for stochastically deteriorating system. *Reliability Engineering & System Safety*, 157:54–63.

Annala, J. (2018). VR Group, Maintenance development manager. Personal interview (19.11.2018).

Bayes, T. (1763). An essay towards solving a problem in the doctrine of chances. 1763. *Philosophical Transactions of the Royal Society*.

Beiser, J. A. and Rigdon, S. E. (1997). Bayes' theorem in the 21st century. *IEEE Transactions on Reliability*, 46(2):291–295.

Block, H. W., Langberg, N. A., and Savits, T. H. (1993). Repair replacement policies. *Journal of Applied Probability*, 30(1):194–206.

Bowles, J. B. and Peláez, C. E. (1995). Fuzzy logic prioritization of failures in a system failure mode, effects and criticality analysis. *Reliability Engineering & System Safety*, 50(2):203–213.

Brown, M. and Proschan, F. (1983). Imperfect repair. *Journal of Applied probability*, 20(4):851–859.

Buzacott, J. A. (1970). Markov approach to finding failure times of repairable systems. *IEEE Transactions on Reliability*, 19(4):128–134.

Cheng, Y. and Tsao, H. (2010). Rolling stock maintenance strategy selection, spare parts' estimation, and replacements' interval calculation. *International Journal of Production Economics*, 128(1):404–412.

Davidson-Pilon, C. (2019). Lifelines Survival regression. (Accessed: 26/03/2019) https://lifelines.readthedocs.io/en/latest/Survival%20Regression.html.

De Souza, G. F. M. (2012). *Thermal Power Plant Performance Analysis*. Springer.

Dekker, R. (1996). Applications of maintenance optimization models: a review and analysis. *Reliability engineering & system safety*, 51(3):229–240.

Dinmohammadi, F., Alkali, B., Shafiee, M., Bérenguer, C., and Labib, A. (2016). Risk evaluation of railway rolling stock failures using fmeca technique: a case study of passenger door system. *Urban Rail Transit*, 2(3-4):128–145.

Distefano, S. and Puliafito, A. (2007). Dynamic reliability block diagrams vs dynamic fault trees. In *2007 Annual Reliability and Maintainability Symposium*, pages 71–76. IEEE.

Doyen, L. and Gaudoin, O. (2004). Classes of imperfect repair models based on reduction of failure intensity or virtual age. *Reliability engineering & system safety*, 84(1):45–56.

Duffuaa, S. O., Raouf, A., and Campbell, J. D. (2015). *Planning and Control of Maintenance Systems*. Springer.

Kaarela, J. (2018). VR Group, Diesel locomotive maintenance planner. Phone interview (22.11.2018).

Kalbfleisch, J. D. and Prentice, R. L. (2002). *The statistical analysis of failure time data*. John Wiley & Sons.

Kaufmann, A., Grouchko, D., and Cruon, R. (1977). *Mathematical models for the study of the reliability of systems*. Elsevier.

Kehä, P. (2018). VR Group, On-site maintenance foreman. Personal interview (10.12.2018).

Kijima, M. (1989). Some results for repairable systems with general repair. *Journal of Applied probability*, 26(1):89–102.

Le Gat, Y. and Eisenbeis, P. (2000). Using maintenance records to forecast failures in water networks. *Urban Water*, 2(3):173–181.

Lee, E. T. and Wang, J. (2003). *Statistical methods for survival data analysis*. John Wiley & Sons.

Lehtola, T. (2018). VR Group, Electic locomotive maintenance planner. Personal interview (20.11.2018).

Levo, A. (2018). VR Group, Multiple unit train maintenance planner. Personal interview (22.11.2018).

Lindqvist, B. (2006). On the statistical modeling and analysis of repairable systems. *Statistical science*, 21(4):532–551.

Mancuso, A., Compare, M., Salo, A., and Zio, E. (2017). Portfolio optimization of safety measures for reducing risks in nuclear systems. *Reliability Engineering & System Safety*, 167:20–29.

Massey Jr, F. J. (1951). The Kolmogorov-Smirnov test for goodness of fit. *Journal of the American statistical Association*, 46(253):68–78.

Parta, R. (2018). VR Group, Electric locomotive engineer. Phone interview (10.12.2018).

Peng, W., Shen, L., Shen, Y., and Sun, Q. (2018). Reliability analysis of repairable systems with recurrent misuse-induced failures and normal-operation failures. *Reliability engineering & system safety*, 171:87–98.

Pham, H. and Wang, H. (1996). Imperfect maintenance. *European journal of operational research*, 94(3):425–438.

Røstum, J. (2000). *Statistical modelling of pipe failures in water networks*. PhD thesis, Department of Hydraulic and Environmental Engineering, Norwegian University of Science and Technology.

Scarf, P. A. (1997). On the application of mathematical models in maintenance. *European journal of operational research*, 99(3):493–506.

Sharma, A., Yadava, G. S., and Deshmukh, S. G. (2011). A literature review and future perspectives on maintenance optimization. *Journal of Quality in Maintenance Engineering*, 17(1):5–25.

Sheu, S.-H., Griffith, W. S., and Nakagawa, T. (1995). Extended optimal replacement model with random minimal repair costs. *European Journal of Operational Research*, 85(3):636–649.

Steck, H., Krishnapuram, B., Dehing-oberije, C., Lambin, P., and Raykar, V. C. (2008). On ranking in survival analysis: Bounds on the concordance index. In *Advances in neural information processing systems*, pages 1209–1216.

Stern, S., Behrendt, A., Eisenschmidt, E., Reimig, S., Schirmers, L., and Schwerdt, I. (2017). The rail sector's changing maintenance game. Technical report, McKinsey & Company.

Sun, Y. (2006). *Reliability prediction of complex repairable systems: an engineering approach*. PhD thesis, Faculty of Built Environment and Engineering, School of Engineering Systems, Queensland University of Technology, Australia.

The SciPy community (2019). Statistical functions (scipy.stats). (Accessed: 26/03/2019) https://docs.scipy.org/doc/scipy/reference/stats.html.

Torpo, N. (2019). Opportunistic maintenance modeling. Bachelor's thesis. Department of mathematics and systems analysis, Aalto University, Finland.

Urbani, M. (2017). Opportunistic predictive maintenance modeling: a simulation approach. Master's thesis, Department of Industrial Engineering, Universita degli Studi di Trento, Italy.

VR Group (2019a). Pendolino. (Accessed: 22/03/2019) https://www.vr.fi/cs/vr/fi/pendolino-juna.

VR Group (2019b). Vectron -sähköveturi. (Accessed: 04/04/2019) https://www.vr.fi/cs/vr/fi/sahkoveturi.

VR Group (2019c). Veturityypit. (Accessed: 04/04/2019) https://www.junablogi.fi/fi/vrgroup/vr-group-yrityksena/liiketoiminnot/juna-liikennointi/veturityypit/.

Wang, H. (2002). A survey of maintenance policies of deteriorating systems. *European journal of operational research*, 139(3):469–489.

Watson, H. A. et al. (1961). Launch control safety study. *Bell Telephone Laboratories*.

Weckmann, G. R., Shell, R. L., and Marvel, J. H. (2001). Modeling the reliability of repairable systems in the aviation industry. *Computers & industrial engineering*, 40(1-2):51–63.

Yang, Q., Hong, Y., Chen, Y., and Shi, J. (2012). Failure profile analysis of complex repairable systems with multiple failure modes. *IEEE Transactions on Reliability*, 61(1):180–191.