MASTER

Software tool for predicting component lifetimes with an application to trucks

Vermeulen, S.J.

Award date:
2016

Link to publication
Software tool for predicting component lifetimes with an application to trucks

by
S. J. Vermeulen

BEng of Engineering in Logistics Management - NHTV 2013
Student identity number 0860862

in partial fulfilment of the requirements for the degree of

Master of Science
in Operations Management and Logistics

Supervisors:
MSc. M. Swaving, DAF Trucks N.V., Costing & Analysis
Dr. E. Topan, Technische Universiteit Eindhoven, OPAC
Ir. Dr. S.D.P. Flapper, Technische Universiteit Eindhoven, OPAC
TUE. School of Industrial Engineering.
Series Master Theses Operations Management and Logistics

Subject headings: statistical inference, component lifetime prediction, parametric inference, predictive analytics, automotive industry, censored data
It should be noted that all costs and component lifetimes are fictitious due to confidentiality reasons.
I. Abstract
This master thesis describes a graduation project carried out at DAF Trucks N.V. in Eindhoven. A software tool was constructed that predicts the component lifetimes of trucks. DAF did not have a validated model for predicting component lifetimes for longer periods. A literature study was conducted focusing on comparing different prediction methods and choosing the best method for predicting component lifetimes. Based on the literature study, an empirically tested method has been proposed to predict the component lifetimes via parametric inference. Also, more insight was required in knowing when enough knowledge about the component lifetime is available such that a prediction of several years is realistically possible. An empirically tested method has been proposed that gives an indication to the stability of the prediction. Finally, a software tool has been programmed that executes both proposed methods for all components.
II. Executive summary
This report is the result of a master thesis project carried out at DAF Trucks N.V. (DAF) in Eindhoven. DAF Trucks N.V. (hereafter: DAF) is a company whose core activities are developing, manufacturing, marketing, selling and servicing medium and heavy trucks. DAF has been founded in 1928 and became a subsidiary company of PACCAR Inc. since 1996. This project was executed at the Costing & Analysis department, part of the aftersales department.

Introduction
DAF offers a range of repair & maintenance services for customers that purchase a new truck. The repair & maintenance services are offered by DAF MultiSupport. The costs for these services are determined in the Costing & Analysis department. Maintenance costs and repair costs are the two main cost drivers for the repair & maintenance contracts. The cost of a repair and maintenance contract is roughly divided into 40% for maintenance and 60% for component repairs. DAF does not aim to make a profit from selling repair & maintenance contracts, but does not want to make a loss on them either. Maintenance costs are easy to determine as they are planned beforehand. Repairs are not planned and are therefore difficult to predict. DAF makes predictions of the component lifetimes to determine the expected repair cost of the repair & maintenance contracts. There are two main issues with the prediction of the component lifetimes. First of all, the predictions of the component lifetimes are made manually via discussions with product engineers and after sales analysts. They would like to have a validated mathematical model to aid them in predicting component lifetimes more accurately. Secondly, when a prediction is made, there is no indication whether the made prediction will have to be adjusted much afterwards (when the component lifetime prediction is mature).

To address the two issues relating the prediction of the component lifetimes, DAF has set up the following business objective:

Construct a mathematical model that predicts component lifetime and gives an indication of component lifetime maturity.

This research focuses on predicting component lifetime for components which lifetime is known enough such that a prediction of several years is realistically possible. Since there is no well-defined approach that identifies whether a component is in the mature phase, the mathematical model will need to recognize when a component lifetime is sufficiently known. An indication of component lifetime maturity will be given together with the component lifetime prediction.

Data analysis
In total there are 51,337 trucks (EURO 5 and EURO 6) included in the research. The mathematical model uses data from customers that purchased a three year “Warranty Plus – Vehicle” contract. It is very likely that a defect will not be observed in the first three years, which means that the data is censored. On average, most of the components are censored. During the preparation of the data, it has been found that for certain components there are significantly more defects observed just before the end of the contract. This has been investigated separately.
Method for predicting component lifetimes
Three prediction methods have been compared in a literature study on the requirements for being able to use them and for which purpose they are most effective. Between lifetime distributions, stochastic processes and artificial neural networks, lifetime distributions have been chosen as the best method to predict component lifetimes as it offered the most flexibility, is not stuck to certain dangerous assumptions and has less chance of overfitting. Four non-negative distributions have been chosen for use in the mathematical model to approximate the true component lifetime: the Weibull distribution, the exponential distribution, the gamma distribution and the lognormal distribution. Two fitting techniques have been compared: the Maximum Likelihood Estimation (MLE) and the least squares or rank regression (RR) method. The MLE outperforms RR when censoring is not evenly distributed among observed defects. In this research, the majority of the censored data is at exactly 1096 when the truck warranty ends, resulting in the MLE being the superior method for fitting the distributions on the data. The predictions are validated empirically by sub setting eight years of data and giving the model increasingly more data. Each data subset it compared to an empirical distribution function to evaluate whether the mathematical model works and how it performs over time. The performance is measured by comparing the cumulative and the individual failure probabilities with the empirical failure probabilities.

The mathematical model has been expanded to handle component improvements. The model evaluates whether the data of the old component and the new component by comparing the Kaplan-Meier empirical distribution functions. If they show similar failure behavior, the old data is used together with the new data to predict component lifetimes. If the failure behavior is not similar, the component lifetime is changed and the old data is not used for new predictions.

Determining component lifetime maturity
A component is mature when the true component lifetime is known enough such that a prediction done now is the same as a prediction done in the future. A method has been developed that uses coefficient of variation for comparing the changes in the fitted distribution parameters of the past time.

\[
\text{Component maturity (reject H0)} = \frac{s(FphV_{b})}{\bar{x}(FphV_{b})} < 0.3, \quad \text{for } b = 1, \ldots, 12 \quad \text{Eq. 1}
\]

where \( s \) is the standard deviation of the estimated FphV of the fourth year (1461 days) for all 12 subsets \( b \), \( \bar{x} \) is the mean estimated FphV value of the fourth year (1461 days) for all 12 subsets \( b \). If the result of the division, called coefficient of variation, is smaller than the cut-off value \( \alpha \) then H0 will be rejected, meaning that the component lifetime prediction is in the mature phase. The 12 subsets \( b \) are subsets of the complete dataset, starting from the start of the observation and ending relatively fixed to the end of the observation. The subsets are spread equally over 2 years with the smallest subset ending 22 months before the end of the observation, and the largest subsets ends at the end of the observation (= whole dataset).
Software implementation of the mathematical model
The mathematical model has been programmed in R-Studio and a SQL query has been written to extract data from the server. Having the mathematical model operational as an independent program allows for easy implementation and complete evaluation of the mathematical model. The programming code is available in this report such that parties within or outside DAF can easily read and use it.

Conclusions
The main assignment has been successfully completed. The mathematical model is able to predict the component lifetime and a method has been developed that gives an indication of component lifetime maturity. This section will describe all important findings and conclusions that were made throughout the report.

An empirical evaluation of the mathematical model has been made. The component lifetime predictions are overestimating the number of failures per hundred vehicles compared to the empirical distribution function. It is difficult to draw conclusions why this happens. The overestimation seems to be equal for most of the individual component predictions. The prediction errors are normally distributed with a small spread, so the overestimation is there but it is very constant over all individual component lifetime predictions. When taking into account this overestimation, the component lifetime predictions are useful for DAF to determine the cost price of the repair & maintenance contracts.

DAF wanted to have an indication of component lifetime maturity for the component lifetime predictions. A method has been developed that gives this maturity indication for every component lifetime prediction. The component lifetime maturity method is based on the coefficient of variation of a series of 12 estimated failures per hundred vehicles (of the fourth year) over the past two years. If the maturity coefficient is below 0.3 then the component lifetime prediction is in the mature phase. Via empirical evaluation it was shown that component lifetime predictions with a lower coefficient have a higher chance of remaining stable, this is very useful for DAF to know how much it can depend on the estimated failure per hundred vehicles.

Recommendations
With the rollout of the new claim handling system called Dealer Claim Entry (DCE), there will be more possibilities regarding predicting component lifetimes. Currently it is very hard to get failure data beyond an observation period of three years because the longer period repair & maintenance defect claims do not have an associated component number. With DCE, this component actually is stored. The first DCE rollout was three years ago in Poland, and Germany followed soon after. This means that in a few years there will be a lot of data available for trucks that have an observation period that is longer than three years. DAF should change the current prediction methods to include this data. It is also recommended to re-run the evaluation of the mathematical model to investigate whether choosing different distributions to predict the component lifetimes is helping in making better predictions.

Each truck has a different configuration of installed components, which leads to that not all components are installed on each truck. It is recommended to include this information in the analysis. However, at this moment it is not possible to extract this information.
Predictions are done for component numbers. There are different configurations possible within the same component number, because the underlying part numbers are not always exactly the same. The underlying parts could be of a different brand or a different version number. Different underlying parts could have an influence on the overall lifetime of the component number. It is recommended to include this information in the analysis as well. However, at this moment it is not possible to extract this information.
III. Preface
This report is the result of my master thesis project at DAF Trucks N.V. which I conducted in partial fulfilment for the degree of Master in Operations Management and Logistics at the Eindhoven University of Technology. I would like to express my gratitude towards some people that have helped me during the graduation project.

First, I would like to thank all my colleagues at DAF for helping during my graduation project. In particular I would like to thank Marijke Swaving and Ellen Koreman for giving me the opportunity to do my graduation project at DAF, supporting me during the project and offering me a job at DAF afterwards.

I would also like to thank my university supervisors, Engin Topan and Simme Douwe Flapper for their support throughout this project. Engin, thanks for your continuous enthusiasm for my graduation project, very usable and timely feedback (even in the evenings and weekends). I hope you will have a good time in Twente. Simme Douwe, thank you for your extensive feedback making me see things from a different perspective and your willingness to join as a second supervisor. I am sure that it has improved the quality of my thesis.

The conclusion of the graduation project ends my time as student. I am very proud to have made it this far. I would like to thank the friends I met during my time as student who make the whole experience as enjoyable as it was.

Furthermore, I would like to express my gratitude towards my family for their unconditional love and support throughout my life. Last but not least I would like to thank Ine for always being there and supporting me.
### IV. Definitions and abbreviations

<table>
<thead>
<tr>
<th>Definition or abbreviation</th>
<th>Explanation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Component lifetime</td>
<td>The time of operation in calendar days of a truck component between the first use by the customer and the breakdown of the component.</td>
</tr>
<tr>
<td>Component maturity</td>
<td>The extent to which the predictions of a component have been stable in the last two years. Divided in three phases: Pre-launch, initial and mature.</td>
</tr>
<tr>
<td>DAF</td>
<td>DAF Trucks N.V. A company whose core activities are developing, manufacturing, marketing, selling and servicing of medium and heavy trucks.</td>
</tr>
<tr>
<td>Defect</td>
<td>A defect is when a component on a truck has failed and requires a repair or replacement to work again. A defect is observed in this thesis when it has been repaired by a DAF dealer and the claim accepted by DAF factory.</td>
</tr>
<tr>
<td>Driveline components</td>
<td>All components directly related to generating power and delivering it to the road. Including engine, transmission, drive shaft and the rear axle.</td>
</tr>
<tr>
<td>Extended warranty</td>
<td>Additional warranty on top of the standard warranty. Covers for up to three years.</td>
</tr>
<tr>
<td>Failure mode</td>
<td>A failure mode is the specific characteristics of materials that result in the failure. It may generally describe the way the failure occurs. Examples are corrosion, wear or fractures.</td>
</tr>
<tr>
<td>FphV</td>
<td>Failure per hundred vehicles. A number of failures per hundred vehicles for a component.</td>
</tr>
<tr>
<td>Kaplan-Meier estimator</td>
<td>Also known as the product limit estimator, is a non-parametric statistic used to estimate the survival function from lifetime data.</td>
</tr>
<tr>
<td>Lifetime distribution</td>
<td>A statistical distribution for modelling the (component) lifetime.</td>
</tr>
<tr>
<td>Likelihood function</td>
<td>Function that calculates the likelihood (probability) of generating exactly the given data set from a chosen statistical distribution with chosen parameters.</td>
</tr>
<tr>
<td>MLE</td>
<td>Maximum Likelihood Estimation. Iterative method for estimating the parameters of a statistical distribution to maximize the likelihood function.</td>
</tr>
<tr>
<td>Observation period</td>
<td>The observation period is the number of days between the delivery of a truck to the customer and the last day of the service contract (warranty or R&amp;M) or 31-12-2015, whichever comes first.</td>
</tr>
<tr>
<td>Query (SQL query)</td>
<td>A SQL query is a request for information from a database table or combination of tables</td>
</tr>
<tr>
<td>R&amp;M, Repair &amp; maintenance</td>
<td>Long term service offered to customer to take care of repairs and maintenance of trucks. Up to eight years running times.</td>
</tr>
<tr>
<td>RDB Global</td>
<td>Reliability Database. Database in which data about failures for each component is stored.</td>
</tr>
</tbody>
</table>
| Right censored data        | A defect occurs after the observed period, but the exact time
<table>
<thead>
<tr>
<th><strong>Reliability</strong></th>
<th>The probability of failure-free performance over a components lifetime in a period.</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>RR</strong></td>
<td>Rank Regression. Robust estimation of regression parameters.</td>
</tr>
<tr>
<td><strong>SQL</strong></td>
<td>Structured Query Language is a special-purpose programming language designed for managing data held in a database.</td>
</tr>
<tr>
<td><strong>Standard warranty</strong></td>
<td>A guarantee given to the purchaser of a truck, specifying that the manufacturer will make any repairs or replace defective parts free of charge for a predefined period of time. The standard warranty period is one year, with two years for the driveline.</td>
</tr>
<tr>
<td><strong>Time to defect</strong></td>
<td>The time of operation in calendar days of a truck component between the first use by the customer and the observed breakdown of the component. Also called months in service.</td>
</tr>
<tr>
<td><strong>Truck component</strong></td>
<td>A component in a truck with a specific function. For example a door-handle, a window, dashboard or an air compressor.</td>
</tr>
<tr>
<td><strong>Truck component group</strong></td>
<td>A collection of components that exist to perform a specific function of a truck. There are many component groups.</td>
</tr>
<tr>
<td><strong>Truck main group</strong></td>
<td>A collection of component groups that exist to perform a major function of a truck. There are ten main groups in a truck.</td>
</tr>
</tbody>
</table>
Contents

I. Abstract................................................................................................................................. iv

II. Executive summary ................................................................................................................. v

III. Preface ................................................................................................................................. ix

IV. Definitions and abbreviations ............................................................................................ x

V. List of figures .......................................................................................................................... xv

VI. List of tables .......................................................................................................................... xvii

1 Introduction ............................................................................................................................. 1

1.1 DAF Trucks N.V. .................................................................................................................. 1

1.1.1 Products .............................................................................................................................. 1

1.1.2 Services and after sales ..................................................................................................... 3

1.1.3 Truck components ............................................................................................................. 4

1.2 Research project ................................................................................................................... 7

1.2.1 Assignment and Deliverables .......................................................................................... 8

1.2.2 Research questions .......................................................................................................... 8

1.2.3 Scope ................................................................................................................................ 9

1.3 Report outline ........................................................................................................................ 10

2 Data analysis .......................................................................................................................... 12

2.1 Analysis of the available data ............................................................................................. 12

2.1.1 Available data on trucks ................................................................................................. 12

2.1.2 Available data on defects .................................................................................................. 12

2.1.3 Available data on component improvements .................................................................. 13

2.2 Retrieving the data using SQL ............................................................................................ 13

2.3 Overview of the data used in this research ......................................................................... 14

2.3.1 Overview of the trucks in this research .......................................................................... 14

2.3.2 Overview of the defects in this research .......................................................................... 16

3 Method for predicting component lifetimes ........................................................................... 18

3.1 How to predict component lifetime ..................................................................................... 18

3.2 Predicting component lifetime method ............................................................................... 20

3.2.1 Choosing appropriate distributions to fit on the available data ...................................... 20

3.2.2 Estimating the values for the parameters of the distributions .......................................... 22

3.2.3 Comparing goodness-of-fit of each distribution ............................................................... 23
3.3 Extension of the method to handle predictions for improved components ........................................... 27
  3.3.1 Comparing data similarity with the Kaplan-Meier estimator ............................................................... 28
3.4 Evaluating the performance of the predictions .................................................................................................. 32
  3.4.1 General evaluation setup ........................................................................................................................... 32
  3.4.2 Evaluating the prediction accuracies of the mathematical model ............................................................. 39
  3.4.3 Evaluating the distribution of the chosen lifetime distributions ............................................................... 41
4 Determining component lifetime maturity ......................................................................................................... 45
  4.1 Defining operational definition for component lifetime maturity ................................................................. 45
    4.1.1 Determine the number of subsets and time between subsets ................................................................. 48
    4.1.2 Determining the cut-off value for maturity ............................................................................................... 50
    4.1.3 Operational definition for component lifetime maturity .......................................................................... 52
    4.1.4 Procedure flowchart for determining component lifetime maturity ................................................. 53
  4.2 Numerical example for determining component lifetime maturity ............................................................. 53
  4.3 Evaluating the performance of the predictions for mature components ....................................................... 55
5 Software implementation of the mathematical model (R code) ............................................................................ 58
6 Research conclusions and recommendations ......................................................................................................... 60
  6.1 Research conclusions ..................................................................................................................................... 60
  6.2 Recommendations ......................................................................................................................................... 61
  6.3 Academic relevance ....................................................................................................................................... 61
  6.4 Further research ............................................................................................................................................. 62
7 Bibliography ....................................................................................................................................................... 64
APPENDIX A: SQL query for retrieving the data ...................................................................................................... 68
  SQL query for the claim data ............................................................................................................................... 68
  SQL query for the truck data ............................................................................................................................... 71
APPENDIX B: Delayed reporting of defects ................................................................................................................ 72
APPENDIX C: Many claims at end of warranty or repair & maintenance contract ...................................................... 75
  Detecting the peak in claims near the end of a contract ...................................................................................... 76
  Evaluating the predictions for components with many defects at the end of the contract .............................. 78
  Detailed overview of the components with many defects at the end of the contract ....................................... 80
APPENDIX D: Basic statistics .................................................................................................................................... 81
APPENDIX E: List of non-negative distributions .................................................................................................... 83
APPENDIX F: Why Rank Regression is not suitable for heavily censored data ...................................................... 85
V. List of figures

Figure 1: The three DAF Truck series XF, CF and LF. ................................................................. 2
Figure 2: Truck axle configurations (DAF Costing & Analysis, 2016). .................................................... 2
Figure 3: DAF MultiSupport Repair & Maintenance packages (DAF MultiSupport, 2015). ....................... 3
Figure 4: Contract lengths where “plus” and “care” contracts are called extended warranty and R&M respectively (DAF MultiSupport, 2015) ................................................................. 3
Figure 5: Structure of the report ........................................................................................................... 11
Figure 6: Graphical representation of the SQL query used to extract data from the database for use in the mathematical model. ............................................................................................................. 14
Figure 7: Number of observed trucks as a function of days observed since delivery .................................... 16
Figure 8: Number of concurrent observed trucks over time .................................................................... 16
Figure 9: Distribution choice flow chart (DAMODARAN, NA). ................................................................. 21
Figure 10: Maximum likelihood estimation for fitting a Weibull distribution. ........................................... 23
Figure 11: Q-Q plots show whether a fitted distribution is a good fit for the underlying data points. These are four plots for four different component numbers. The straight line is the fitted distribution; the points are the underlying data. ................................................................................................................................. 26
Figure 12: Graphical representation of the four distributions to assess goodness of fit for component number 40250 (“Pulleys near the engine region”). The thick black crosses represent the Kaplan-Meier empirical distribution (the data), the red line is the fitted distribution. The dashed lines are the 95% confidence intervals for the Kaplan-Meier empirical distribution. ......................................................................................... 27
Figure 13: Comparing two empirical Kaplan-Meier survival plots of the same component before and after the improvement (component number 15455 “Dosing Module EAS”). ................................................................. 30
Figure 14: Comparing two empirical Kaplan-Meier survival plots of the same component before and after the improvement. Notice the very small scale on the y-axis ........................................................................ 31
Figure 15: Each subset of the complete dataset contains an additional year of data. ................................. 32
Figure 16: Histograms of the prediction errors per component for subset 3, FphV after 2 years. Left histogram shows every component, right histogram is without the one outlier at around -15 to improve readability. .............. 40
Figure 17: Histogram of the prediction errors for all components for subset 5, FphV at year 1 .................... 40
Figure 18: Distribution of lifetime distributions over the different subsets .................................................. 42
Figure 19: Chosen distributions within subset 8, categorized from components with few observed defects to components with many defects ................................................................. 42
Figure 20: Subsets are fixed between time 0 and relatively fixed to time T (end observation time) ............ 47
Figure 21: Spread of the previous predictions for determining maturity ................................................... 50
Figure 22: Performance of the predictions for different maturity cut-off values, lower is better ................ 51
Figure 23: Number of components that have H0 rejected per cut-off value ............................................ 52
Figure 24: Flowchart for determining component lifetime maturity ....................................................... 53
Figure 25: Histograms of the prediction errors per component for subset 3, FphV after 2 years. Left histogram shows every component, right histogram shows only components with H0 rejected .............................................. 56
Figure 26: Histograms of the prediction errors per component for subset 5, FphV after 1 year. Left histogram shows every component, right histogram shows only components with H0 rejected .............................................. 56
Figure 27: Flowchart of the global processes within mathematical model ................................................. 58
Figure 28: Graphical representation of the query shown in this appendix ................................................. 68
Figure 29: Causes of unclean data ......................................................................................................... 72
Figure 30: Component number 10306 has a strong increase in observed defects after every 12 months ......... 73
Figure 31: Example histogram showing many defects claimed in the very last period of the contract .......... 76
FIGURE 32: FORMS OF SKEWNESS (SAFARI).........................................................................................................................81
FIGURE 33: FORMS OF KURTOSIS..................................................................................................................................................81
FIGURE 34: THE PROBABILITY DENSITY FUNCTION OF A NORMAL DISTRIBUTION. (DAMODARAN, STATISTICAL DISTRIBUTIONS)...........82
FIGURE 35: RANK REGRESSION CAN BE DONE ON THE Y AXIS OR THE X AXIS........................................................................85
VI. List of tables

TABLE 1: TRUCK MAIN GROUPS WITH ASSOCIATED COMPONENT NUMBERS ..........................................................4
TABLE 2: OVERVIEW OF THE DIFFERENT TRUCK TYPES INCLUDED IN THIS RESEARCH ...............................15
TABLE 3: ADVANTAGES AND DISADVANTAGES OF THE DIFFERENT PREDICTION TECHNIQUES ................20
TABLE 4: EXAMPLE FOR CALCULATING THE KAPLAN-MEIER ESTIMATE .........................................................29
TABLE 5: THE NUMBER OF OBSERVED TRUCKS IN EACH SUBSET ...................................................................33
TABLE 6: DESCRIPTION OF EACH ROW AND COLUMN IN TABLE 7 ....................................................................33
TABLE 7: SUMMARY SCREEN SHOWING THE MOST IMPORTANT RESULTS OF THE MATHEMATICAL MODEL FOR COMPONENT NUMBER 30508 (= PINION BEARING CAGE SEAL, REAR AXLE). .................................................................34
TABLE 8: AFphV VALUES FOR EACH I, J (LEFT) AND THE NUMBER OF COMPONENTS IN EACH SUBSET (RIGHT) ..........................................................................................................................36
TABLE 9: AFphVKM VALUES FOR EACH I, J ..............................................................................................................37
TABLE 10: SWITCH TO DCE PER COUNTRY FOR THE COUNTRIES THAT ARE CURRENTLY INCLUDED IN THE RESEARCH. N/A MEANS THAT DCE HAS NOT BEEN IMPLEMENTED YET. .................................................................38
TABLE 11: PREDICTION ACCURACIES (EQ. 15) FOR ALL I, J (LEFT) AND THE NUMBER OF COMPONENTS K IN EACH SUBSET (RIGHT) .................................................................39
TABLE 12: STANDARD DEVIATION OF THE PREDICTION ERRORS FOR ALL SUBSETS J AND ALL FphV TIMES I (EQ. 17) ........................................................................................................41
TABLE 13: PREDICTION ACCURACIES FOR ALL I, J (EQ. 15) WHEN CHOOSING ONLY WEIBULL DISTRIBUTION ..........................................................................................................................43
TABLE 14: STANDARD DEVIATION OF THE PREDICTION ERRORS FOR ALL SUBSETS J AND ALL FphV TIMES I (EQ. 17) WHEN CHOOSING ONLY WEIBULL DISTRIBUTION ..........................................44
TABLE 15: SUBSETS FOR COMPONENT 63402 USED AS DATA FOR DETERMINING COMPONENT LIFETIME MATURITY ..................................................................................................................54
TABLE 16: PREDICTED SHAPE AND SCALE PARAMETERS OF THE WEIBULL DISTRIBUTION AFTER FITTING IT ON THE SUBSETS ..................................................................................54
TABLE 17: PREDICTED FAILURES PER HUNDRED VEHICLES (FphV) FOR THE FIRST FOUR YEARS FOR EACH SUBSET b ..................................................................................................................54
TABLE 18: PREDICTION ACCURACIES (EQ. 15) FOR ALL I, J FOR ONLY COMPONENTS K THAT HAVE H0 REJECTED (EQ. 24) (LEFT) AND THE NUMBER OF COMPONENTS COMPARED (RIGHT) ........................................................................................................55
TABLE 19: STANDARD DEVIATIONS OF THE PREDICTION ERRORS (EQ. 17) FOR ALL I, J FOR ALL K THAT HAVE H0 REJECTED (ARE MATURE) ........................................................................57
TABLE 20: ADVANTAGES AND DISADVANTAGES OF THE DIFFERENT PREDICTION METHODS .................................................................62
TABLE 21: CHANCE FOR DELAYS IN REPORTING DEFECTS TO DEALERS OR SERVICE POINTS ........................................73
TABLE 22: PREDICTION ACCURACIES (OF ALL COMPONENTS K) FOR ALL I, J (LEFT) AND THE NUMBER OF COMPared COMPONENTS (RIGHT) ........................................................................................................78
TABLE 23: STANDARD DEVIATION (OF ALL COMPONENTS K) OF THE PREDICTION ERRORS FOR ALL I, J .............79
TABLE 24: PREDICTION ACCURACIES (OF ONLY COMPONENTS WITH MANY DEFECTS AT THE END OF THE CONTRACT) FOR THE DIFFERENT SUBSETS (LEFT) AND THE NUMBER OF COMPARED COMPONENTS (RIGHT) ........................................................................................................79
TABLE 25: STANDARD DEVIATION (OF ONLY COMPONENTS WITH MANY DEFECTS AT THE END OF THE CONTRACT) OF THE PREDICTION ERRORS (EQ. 17) FOR ALL I, J ........................................80
TABLE 26: LIST OF NON-NEGATIVE DISTRIBUTIONS .............................................................................................83
TABLE 27: CASE EXAMPLE FROM RELIASSOFT SHOWING SHORTFALLS WHEN RANK REGRESSION IS DEALING WITH CENSORED DATA. .................................................................86
1 Introduction

This is the master thesis executed at DAF Trucks N.V (Hereafter: DAF) in Eindhoven, the Netherlands. The project will be performed at the Costing & Analysis department. The first chapter gives an introduction to the research environment, the research project and the report outline.

1.1 DAF Trucks N.V.

DAF Trucks N.V. (hereafter: DAF) is a company whose core activities are developing, manufacturing, marketing, selling and servicing medium and heavy trucks. DAF has been founded in 1928 by the brothers Hub and Wim van Doorne who developed it from a simple engineering business and blacksmith workshop to the fastest growing truck manufacturer in Europe. DAF became a subsidiary company of PACCAR Inc. since 1996.

DAF currently occupies a leading position in the European truck market with a market share of 13.8% in the heavy-duty segment (16 tonnes and over). It holds an 8.8% market share for the light segment (6-16 tonnes). DAF is market leader in the Netherlands, the UK and Hungary and the number one supplier in the truck market in Belgium and Poland (DAF Facts and figures, 2016). In 2014, the total market for heavy-duty commercial vehicles in the European Union was almost 227,000 units.

DAF has four production facilities. Eindhoven, the headquarters of DAF, holds the main production facility. The design centre is also located there. Other production facilities exist in Westerlo (Belgium), Leyland (United Kingdom), and Ponta Grossa (Brazil). These production facilities have a combined area of 2,200,000 m². DAF has a total workforce of 9420 FTE (DAF Facts and figures, 2016).

1.1.1 Products

DAF offers a complete range of trucks from 7.5 tonnes Gross Vehicle Weight up to 50 tonnes Gross Combination Weight and above. Every transport application has unique requirements – so every DAF vehicle is specialized, built to order from a range of cab, chassis, driveline and electronic configurations, providing the lowest operating cost per kilometre in the industry, excellent transport efficiency and optimum comfort for the driver. DAF also supplies engines, axles and cabs to manufacturers of buses, coaches, off road vehicles and special vehicles for agriculture and industry. The customers of DAF are companies that want to use trucks for own transport, third party logistics companies or rental companies.

DAF has three series of trucks, XF, CF and LF (Figure 1). XFs are designed for long distance transport. It offers high fuel efficiency, low operating cost and the highest driver comfort. CFs are standard trucks that excel in their versatility. Customers that order a CF have the most options available to configure their trucks to suit their needs. LFs are designed for short distance distribution work. It offers the highest manoeuvrability and payload efficiency.
There are two chassis types built at DAF: Rigid chassis (FA) and Tractor chassis (FT). For each chassis type a multitude of configurations are possible, shown in Figure 2.

The FA and FT are available for all three series and are also by far the most commonly sold axle configuration. For this reason, the project to create a prediction model will focus on doing predictions for the FA and FT chassis.
1.1.2 Services and after sales

DAF offers a range of services. They can help the customer in choosing the right vehicle, arrange financing and keep their fleet in perfect condition (DAF Services, 2016). Since this research is about warranty, repairs and maintenance, a detailed explanation on the repair & maintenance service is given.

DAF MultiSupport **repair & maintenance** program offers a tailor made service proposal to maintain vehicles for customers. A truck always comes with a standard one year factory warranty, and a two year standard warranty for driveline components. Driveline components are directly related to generating power and delivering it to the road, including engine, transmission, drive shaft and the rear axle. Customers can purchase extra insurances for their vehicle via different packages in the program. These packages range from maintenance (preventive maintenance in literature) or repairs (corrective maintenance in literature) all the way up to complete packages including breakdown support and legal inspections. Figure 3 and Figure 4 show the coverage of each package and the maximum length of each package. The packages are normally chosen together with the salesperson during the purchase of a new truck. About 40% of the customers buys a “Warranty Plus” package, 30% buys a “Care” package and the rest does not use the program.

![Figures 3 and 4](DAF MultiSupport, 2015)

DAF uses a network of franchise DAF dealers to perform all the required repairs. Franchise means that the dealers are independent organizations responsible for their own profits, but have an agreement with DAF Trucks N.V. to allow the use of their name. The dealers maintain the trucks for customers that...
have a repair & maintenance package bought from DAF. The dealer sends a claim to DAF whenever they help one of the customers of DAF. DAF then compensates the dealer for the labor and parts used to help their customer.

1.1.3 Truck components

Every truck has a comprehensive parts list. This list is structured in a hierarchy with four layers:

1. Main group (e.g. Cab or chassis)
2. Component group (e.g. Interior or exterior of the cab)
3. Component number (e.g. Driver’s chair, curtains or dashboard in the interior of a cab)
4. Parts (e.g. backrest, armrest and the suspension of the driver’s chair)

There are ten main groups that perform a major function of the truck, e.g. the cab, the chassis, front axle, etc. The ten main groups are divided into multiple groups. For the “Cab” main group, this would be the interior, exterior, body or the suspension of the cab. Each of the component groups contains several components that each performs a basic function in the truck. For the cab interior this would be the driver’s seat, climate control, dashboard, beds, curtains, and more. A component number is always a 5-digit number. The number relates to the main group that the component belongs to. See Table 1 for the number relations. A customer can usually choose between multiple variants of a component, for example a luxurious wooden or a normal plastic dashboard. These variants have unique article numbers that tell exactly what parts will be installed on the truck.

Table 1: Truck main groups with associated component numbers

<table>
<thead>
<tr>
<th>Main group</th>
<th>Component group</th>
<th>Component numbers used per main group</th>
</tr>
</thead>
<tbody>
<tr>
<td>Chassis</td>
<td>(multiple groups)</td>
<td>10000-19999</td>
</tr>
<tr>
<td>Front axle</td>
<td>...</td>
<td>20000-29999</td>
</tr>
<tr>
<td>Rear axle</td>
<td>...</td>
<td>30000-39999</td>
</tr>
<tr>
<td>Motor general</td>
<td>...</td>
<td>40000-44999</td>
</tr>
<tr>
<td>Motor installation</td>
<td>...</td>
<td>45000-49999</td>
</tr>
<tr>
<td>Transmission</td>
<td>...</td>
<td>50000-54999</td>
</tr>
<tr>
<td>Steering system</td>
<td>...</td>
<td>55000-59999</td>
</tr>
<tr>
<td>Cabs</td>
<td>...</td>
<td>60000-69999</td>
</tr>
<tr>
<td>Brake installation</td>
<td>...</td>
<td>70000-79999</td>
</tr>
<tr>
<td>Electrical installation</td>
<td>...</td>
<td>80000-89999</td>
</tr>
</tbody>
</table>

The word “component” will be used many times throughout this research project. The word always refers to the third layer of the hierarchy structure (explained above), an object that performs a basic function in the truck. Any numbers associated refer to the 5-digit component number. Since components are important for this project, more details on this subject will be described in the following sub-sections.
1.1.3.1 Component defects
DAF stores data from each claim that franchise dealers send to DAF to receive compensation for helping DAF customers. Every claim that DAF stores for defective components contains information on the following points:

- The truck chassis number
- Date for when the defect was reported to the dealer by the customer
- The component number of the defective component
- Monetary compensation given to the dealer

Using the truck chassis number that is attached to the claim, it is possible to get details on the truck specifications and details about the customer using the truck. The component number tells which component has failed.

1.1.3.2 Truck component maturity
DAF uses different component lifetime prediction techniques for newly developed components and components that have been on the market for a long period. DAF classifies the certainty of the component lifetime knowledge in three ways: “Pre-launch phase”, “initial phase”, and “mature phase”.

Pre-launch phase - In the pre-launch phase, knowledge on the component lifetime distribution comes from accelerated life tests, predictions and physical product tests from the component developer or the supplier of the component.

Initial phase - A component goes to the initial phase after it has been introduced in the truck. During this phase, warranty claims (data points) will start coming in. The predicted component lifetime will be adjusted by experts as time passes and more warranty claims enter the database.

Mature phase - After a while, the required adjustments to the predicted component lifetime will get smaller and smaller. No new adjustments to the component lifetime are being introduced. When this happens, enough knowledge is available about the component lifetime to call it mature.

DAF knows the switch from the Pre-launch phase to the Initial phase. It is when the first new component gets installed on a truck and delivered to the customer. The switch from the initial phase to the mature phase means that component lifetime predictions made for mature components should not change when more data is added, but there is no operational measure.

1.1.3.3 Truck component improvements
Truck components get improved to meet new requirements or increase its reliability. Some examples: rattling noise coming from a specific roof spoiler type; introduction of a new driver seat; NOx sensor premature failure in certain conditions. Whenever a significant change is done to a component, a record is saved by DAF in a database called “Datalab”. This record contains the following data:

- Truck series, chassis type and axle configuration
- Improved component number
- Short description of the improvement reason
• Chassis number that had the first improved product installed
• Date for when the improved component is introduced

When a component is improved over multiple truck types, a separate record is saved for each different truck type. An improved component will have a different lifetime compared to the old component, even though it still performs the same function. This means that the knowledge on the component lifetime decreases after an improvement. It is possible that old lifetime data becomes unusable when a component improvement changes the component lifetime too much. The mathematical model that will be constructed during this research will need to take this into account.
1.2 Research project
Each truck comes with a standard warranty and purchasable extended warranties. DAF does not aim to make a profit from selling extended warranties and repair & maintenance contracts, but does not want to make a loss on them either. Determining the costs on maintenance is easy since all the maintenance plans are predetermined. Determining the costs for defects is much more difficult, since they are not anticipated. To predict the number of defects, DAF needs to predict the lifetime of the warranted components. Component lifetime is the time of operation in calendar days of a truck component between the first use by the customer and the breakdown of the component. Currently, the prediction of component lifetime is made with expert knowledge from product engineers and aftersales analysts. The predictions are adjusted after failure data becomes available. DAF classifies the knowledge on the component lifetimes in three phases: pre-launch, initial, and mature. During the first two stages they are uncertain of the component lifetime, which could result in unfair prices for the customer or making losses on the contracts. Experts say that it takes four or five years before the knowledge of the component lifetime switches from the initial phase to the mature phase. In the mature phase, DAF is more certain of the component lifetimes and can offer a fairer price for the customer. For this research project, DAF would like to investigate whether there is an improvement possible on two subjects:

- The predictions of component lifetimes are made via discussions with product engineers and after sales analysts. They would like to have a mathematical model that has been validated to help DAF in predicting component lifetimes more accurately.
- It takes on average four or five years before a component switches to the mature phase according to the current situation. DAF would like to know whether this could be shorter. There is also no mathematical determination for component maturity. This mathematical determination should be developed.

To know whether a study is really necessary, it is important to research whether the initial project stated by DAF is necessary. Then a formal definition for the project will be given. After this, the scope of the project is given and finally all the deliverables needed to finish the project.

Very recently a new statistical prediction model called “Automated Weibulls” was developed by the Costing & Analysis department in an effort to help DAF in predicting component lifetimes. The “Automated Weibulls” model fits a Weibull distribution using Maximum Likelihood Estimation (MLE) (Aldrich, 1997) and Rank Regression (RR, also known as least squares) (Marquardt, 1963) on the lifetime of a component. The model is currently only used to signal big deviations in predicted component lifetimes. The model is not able to recognize when a component is mature, and assumes that the components follow a Weibull distribution. Finally, the person who created the model has left the company without explaining how the programming works. This means that the model in its current state is a black box which works for the latest truck series, but cannot be changed to predict future series. Following the completion of the prediction model, several questions remained unanswered. It is not completely known whether the MLE or RR technique is better for this situation. There is also no indication to the maturity of a predicted component lifetime which makes it hard to trust the prediction. On a higher level, the other prediction techniques such as stochastic processes or neural networks were
not thoroughly investigated. DAF would like to have a thorough investigation on how to predict component lifetimes and how to set up a good evaluation for a prediction model.

There is a lot of money involved in the extended warranty and repair & maintenance contracts. Due to confidentiality reasons, the costs involved cannot be displayed.

The goal of the project is to construct a mathematical model that predicts component lifetimes. The model will give its users an indication to the extent a component has reached maturity. This model will ultimately help DAF in helping the product engineers and aftersales analysts in doing more accurate predictions on component lifetimes in a less time consuming effort.

1.2.1 Assignment and Deliverables

Based on the previous section, the assignment of this project:

Construct a mathematical model that predicts component lifetime and gives an indication of component lifetime maturity.

This research focuses on predicting component lifetime for components which lifetime is known enough such that a prediction of several years is realistically possible. Since there is no well-defined approach that identifies whether a component is in the mature phase, the mathematical model will need to recognize when a component lifetime is sufficiently known. An indication of component lifetime maturity will be given together with the component lifetime prediction.

The mathematical model is the combination of the three deliverables listed below:

1. Method for predicting component lifetimes
   a. Choosing the prediction technique
   b. Extension of the method to handle predictions for improved components
   c. Evaluation of the performance of the predictions
2. A definition for when a component is mature
   a. Method for detecting maturity
3. Automate the methods of the above deliverables for all components

1.2.2 Research questions

Based on the deliverables listed in section 1.2.1, research questions are created that need to be answered before the deliverables can be realized.

For the first deliverable, “Method for predicting component lifetimes”, the following research questions are created:

- What methods are available for predicting component lifetime?
  - What are the requirements before being able to use each method?
  - For which purpose is each method most effective?
- Which prediction technique is best used based on the available data?
- What are the best indicators to measure the performance of the mathematical mode?
• What are acceptable values for the performance indicators?
• How to do predictions when components have been improved?

The second deliverable, “A definition for when a component is mature”, has the following research questions:

• What is maturity?
• What factors influence maturity?
• Which technique is the best for evaluating maturity for this research?
• What is the mathematical method to determine when a component reaches the mature phase?

The third deliverable, “Automate the methods of the above deliverables for all components”, has the following research questions:

• Which software program is best used?
• What programming code lines need to be written to automate the methods of the above deliverables?

1.2.3 Scope
Some aspects within the research are being bounded or pre-defined. The primary reason for doing this is because there is only limited time to complete the research.

• Predictions are best when each failure mode can be isolated and predicted independently. Truck parts have the least possible failure modes, but there is no truck part lifetime data available. Therefore, predictions for the component lifetime will be executed on the component level.
• There will be a distinction between the different truck series (LF, CF, and XF) and truck chassis (FT and FA) as it is believed that there is a difference in component lifetime between these categories.
• The project will focus on doing predictions for the FA and FT axle configurations. Of all the axle configurations used in Figure 2, the FA and FT is by far the most used configuration. Predicting these configurations correctly is DAF’s highest priority since these predictions have the highest effect on total costs.
• Replacements due to component recalls (immediate or at-next-service) are not taken into account for determining component lifetime. Recalls are very rare and are not defined as a normal breakdown of a component for this research. DAF wants to know the component lifetime for components that break down during normal operation (manufacturing variation, wear out of components or variation in customer usage).
• Technology never stands still and trucks keep changing. To ensure a homogenous database with comparable trucks and components, only EURO5 and EURO6 trucks delivered between 01-01-2008 and 31-12-2015 are included in the data.
• For trucks that have a Repair & Maintenance contract, no component numbers are stored for observed defects. This makes linking defects to a component impossible. Only trucks from the
United Kingdom have the component number stored because they use a different database for storing claim information. Therefore, R&M data comes only from UK trucks.

- Only finalized and accepted repair claims are included in the data in order to filter out false defects. Because finalizing repair claims can take a long time, all observations of component defects stop at the first of January, 2016. This also ensures a stable non-changing database during the research project.

- In order to get components that have a similar normal-operation, the dataset will contain only of trucks delivered to the following countries: Austria, Belgium, Czech Republic, Spain, Denmark, Finland, France, Germany, Greece, Hungary, Ireland, Italy, Luxembourg, Netherlands, Norway, Poland, Portugal, Romania, Spain, Sweden and the United Kingdom. This is based on an infrastructure quality index (Statista, 2016) and discussion with DAF experts.

- It is unclear which components are covered in which warranty contract and exploring this would cost too much time. Therefore, only trucks with a full extended warranty (Warranty plus – vehicle) that cover all components are added to the dataset for analysis.

- A repeating defect for the same component is more a function of repair quality compared to the design and manufacturing quality (Rai & Singh, 2009). Therefore, in order to separate manufacturer quality from service quality, use of time to first defect is recommended.

- The mathematical model should be able to predict the lifetime of all components within one night, between 20:00 and 6:00 (10 hours). This is to make sure that there is enough server capacity for ad-hoc analysis by DAF engineers during the day.

1.3 Report outline
The structure of the report is shown in Figure 5. The deliverables from the project are included in the figure.
Chapter 2 “Data analysis” will analyse the available data for use in the research project, and gives an overview of the used data. Any interesting effects seen during analysis will also be described. Chapter 3 “Method for predicting component lifetimes” discusses the different prediction techniques, explains the used prediction technique, and evaluates the performance of the predictions. Chapter 4 “Determining component lifetime maturity” describes how component lifetime maturity is defined and measured. The performance of the model will be evaluated again for mature components. Chapter 5 “Automation of the mathematical model (R code)” describes the R code used to run all the parts of the mathematical model automatically. Finally, chapter 6 “Research conclusions and recommendations” describes the overall research conclusions, limitations and recommendations.
2 Data analysis

DAF stores a lot of data on various subjects that could be useful for this project. The available data will be analyzed for usefulness in this project. After analyzing, a query has to be constructed to retrieve the data. The retrieved data will be overviewed. Then, any phenomena in the data that could influence the quality of the predictions will be analyzed and corrected if necessary.

2.1 Analysis of the available data

The main goal of the mathematical model is to predict the component lifetime of components that have reached the mature phase. The model should also be extended to handle component improvements. To construct this mathematical model and make it work, data is needed. Most importantly, data about the trucks, claims, and components are needed. In general more data is better, but the data does have to be homogenous in order to do good analysis on them. Therefore, for data to be useful it is subject to several constraints to ensure homogeneity. The constraints are discussed with experts from DAF, as well as literature, in the scope. DAF stores its data in an electronic database. The following subsections analyzes the available data for usefulness in this project.

2.1.1 Available data on trucks

The following available data on trucks could be useful for this research project, collected and filtered according to the scope of the project:

- Chassis number
- Truck chassis (FA and FT)
  - Only FA and FT trucks are used.
- Truck series (LF, CF and XF)
- Truck product range (14-16T, XF95, XF105, XF MX-11, etc)
- Engine type (EURO 5, EURO 6)
  - Only EURO 5 and EURO 6 trucks are used.
- Truck delivery date
  - Only trucks delivered between January 1 2008 and December 31 2015.
- Truck delivery country
  - Only trucks that are delivered to the following countries: Germany, France, Italy, Spain, Poland, The Netherlands, Greece, Czech Republic, Portugal, Hungary, Austria, Denmark, Luxembourg and the United Kingdom
- Truck service contract type
  - Trucks with a three year “Warranty – Plus” contract only, with the addition of trucks with a “Full care” contract of any length that are delivered to the U.K.
- Truck repair & maintenance contract duration length

2.1.2 Available data on defects

The following available data on defects could be useful for this research project, collected and filtered according to the scope of the project:

- Chassis number
Only claims with a matching chassis number from the remaining trucks in section 2.1.1

- Component defect date
- Component number of the defective component
- Time to defect (Date difference between truck delivery date and defect date)
- Claimed repair costs
- Paid repair costs
- Claim sort (Warranty , R&M repair, maintenance, field action)
  - Only claims from defects in warranty, or warranty and R&M for trucks from the UK
- Claim country (country of the dealer where the repair was claimed)
- Claim status (active or finalized)
  - Only claims that have been finalized and accepted

2.1.3 Available data on component improvements
The following available data for component improvements could be useful for this research project, collected and filtered according to the scope of the project:

- Component number
- Improved truck series (LF, CF, XF)
- Improvement date
- Improvement description

2.2 Retrieving the data using SQL
Accessing data in the database requires sending a query to the server. A query is a request for information from a database. Queries have to be written in a “Structured Query Language”, called SQL, for the database to correctly return the requested data. The available data has already been analyzed in section 2.1, and any constraints from the scope have been added there as well. To retrieve the required data, a query was written that can be used as input for the mathematical model. Since the query is quite complex and very long, it can be read in “APPENDIX A: SQL query for retrieving the data”. A graphical representation of the data used in the database is shown in Figure 6
The SQL server called “Datalab” contains multiple data tables with raw data. These are shown at the right side of Figure 6. It is possible to extract information from these data tables via a query. For readability, performance and aggregation purposes, sub-queries are defined within a query. These sub-queries are shown in the middle part of the figure. The names of the subqueries are purely for the programmer to know what the purpose of the sub-query is. For each query or subquery it is possible to add filters or groupings to them to retrieve only the required data and not more. Each arrow in the figure is a specific data transfer from a data table. It selects specific columns from the data table as specified by the programmer.

As said, the repair & maintenance data from the United Kingdom is stored in a different database. The UK is the only country with an independent database for their claims. This database is very useful as it does store the component numbers for any repair & maintenance “care” package claims, as opposed to the other database which does not. A download was made with the required data, which was in turn imported as a data table (dbo.Stefan_ukclaims) in the currently used database in order to be able to query the data as required.

2.3 Overview of the data used in this research
A query has been written that extracts all the data matching the criteria listed in section 2.1. The characteristics of the data will be described in the following sub-sections.

2.3.1 Overview of the trucks in this research
All the required data from section 2.1.1 is available, so it can be used in this research. In total there are 51.337 trucks included in the research of which 6688 are LF series, 4701 are CF series and 39.948 are XF series. A detailed overview is shown in Table 2.
Table 2: Overview of the different truck types included in this research

<table>
<thead>
<tr>
<th>Product range</th>
<th>Engine type</th>
<th>Number of trucks</th>
</tr>
</thead>
<tbody>
<tr>
<td>CF65</td>
<td>EURO 5</td>
<td>513</td>
</tr>
<tr>
<td>CF66</td>
<td>EURO 6</td>
<td>95</td>
</tr>
<tr>
<td>CF75</td>
<td>EURO 5</td>
<td>471</td>
</tr>
<tr>
<td>CF76</td>
<td>EURO 6</td>
<td>270</td>
</tr>
<tr>
<td>CF85</td>
<td>EURO 5</td>
<td>2344</td>
</tr>
<tr>
<td>CF86</td>
<td>EURO 6</td>
<td>1008</td>
</tr>
<tr>
<td>LF36</td>
<td>EURO 6</td>
<td>1042</td>
</tr>
<tr>
<td>LF45</td>
<td>EURO 5</td>
<td>2645</td>
</tr>
<tr>
<td>LF46</td>
<td>EURO 6</td>
<td>476</td>
</tr>
<tr>
<td>LF55</td>
<td>EURO 5</td>
<td>1770</td>
</tr>
<tr>
<td>LF56</td>
<td>EURO 6</td>
<td>755</td>
</tr>
<tr>
<td>XF105</td>
<td>EURO 5</td>
<td>25413</td>
</tr>
<tr>
<td>XF106</td>
<td>EURO 6</td>
<td>14469</td>
</tr>
<tr>
<td>XF116</td>
<td>EURO 6</td>
<td>65</td>
</tr>
<tr>
<td>XF95</td>
<td>EURO 5</td>
<td>1</td>
</tr>
</tbody>
</table>

Truck series will be separated by series (LF, CF and XF) and chassis (FA and FT) in the mathematical model, providing a separate prediction for each series. The separation is made because it is thought by DAF that the trucks are used differently, causing different component lifetimes.

The observation period is the period (in days) in which a defect will be observed when it happens. This period is the number of days between the delivery of a truck to the customer and the last day of the service contract (warranty or R&M) or 31-12-2015, whichever comes first. Nearly all of the trucks in the research have a “Warranty Plus – Vehicle” contract of three years, except for trucks in the UK that can also have a longer repair & maintenance contract (up to eight years). The concurrent number of observed trucks is plotted as a function of days observed in Figure 7. This shows that the number of trucks observed reduces drastically after 1096 days (three years).
The research includes trucks delivered to customers from 01-01-2008 until 31-12-2015. Figure 8 shows the number of concurrent trucks being observed at any time within the interval. In early 2010 there were about 5000 trucks simultaneously being observed, in 2014 there were about 15000 trucks being observed simultaneously. The increasing line can be explained since trucks delivered before 2008 are not in the dataset.

All the data that was required from section 2.1.2 could be successfully downloaded, so it can be used in this research. For the 51,337 observed trucks, many defects have been observed over 2023 types of defects (component numbers). Each of the 2023 types of defects has at least 1 observed defect. By dividing the number of defects by the number of components you get the average number of defects.
per component. Dividing this by the number of trucks shows the number of components that have an observed defect during the observation period. The other components will fail somewhere after the observation period. This is also called type I censoring, or right censoring. The distribution of defects over the different types of defects is very similar to the Pareto distribution: The top 20% of the component numbers account for 87% of the observed defects. This means that the remaining 13% of the observed defects are divided among 80% of the component numbers, resulting that the censoring effect gets even stronger for most of the component numbers. This can be classified as very heavy censoring for most of the component numbers. The methods used in this research must be capable of handling right censored data. It also means that anything done with this data amounts to extreme extrapolation because there are so few actual failures (or none at all), which will make predictions very difficult. Because DAF is not interested in the really well performing components, no predictions are necessary for components that have less than 10 observed defects.

A limitation in the data is that there is no list of installed components available for each truck that can be related to the component numbers for each component. Therefore, the assumption is made that all trucks have every component installed. This will be true for most components, but definitely not for all of them. Optional components such as fog lights or a second fuel tank might not be installed on every truck. What is not installed cannot break down, so currently the component will be seen as a censored observation while it actually should not have been an observation at all. The mathematical model will then overestimate the component lifetime for these optional components.

Another limitation on the data is that it is only possible to see component numbers for received claims, not the actual component article number. Component number 41140 is the starter motor, but there is no information on the brand, the series or version number. The mathematical model therefore cannot make a distinction between the different article numbers and thus predicts the overall component lifetime.

Several interesting phenomena were found during the extraction and exploration of the data. There is a recurring pattern found when making a histogram of the number of claims over time. There seems to be a peak in claims at month 24, 36, 48 and 60. This yearly recurring pattern will be discussed in “APPENDIX B: Delayed reporting of defects”. Also, for several components there are large peaks in the number of claims just before the end of the warranty period or the repair & maintenance contract period. These end-contract outliers will be discussed in “APPENDIX C: Many claims at end of warranty or repair & maintenance contract”
3 Method for predicting component lifetimes
This chapter describes the method that is used in the mathematical model to predict the component lifetimes. First, a literature study is performed to determine the best method for predicting the component lifetimes in this situation. Then, it will be explained how the chosen method works and how this method is applied to the data. Afterwards, the method will be extended to handle recorded component improvements. Finally, the method will be empirically evaluated.

3.1 How to predict component lifetime
There are many ways to do predictions. One will be better than the other. For DAF, a good prediction is one that shows the probabilities of failure for a set time period such as the length of a repair & maintenance contract. This section shows a number of prediction techniques found from literature. The advantages and disadvantages of these techniques will be discussed here. Finally, a choice will be made which technique will be used in this project to predict component lifetimes.

**Lifetime distributions:** This method tries to capture the lifetime probability distribution of a component by estimating the parameters of a selected probability distribution to match the data over time. This distribution can then be used to predict the chance to failure for a given period. Kleyner & Sandborn (Kleyner & Sandborn, 2005) applied the exponential distribution and the Weibull distribution to capture the lifetime distribution of their components. In accordance with the bathtub curve (different failure rates during early life/useful life/wear out), Kleyner & Sandborn used the exponential distribution to simulate early life defects and the Weibull Distribution for random failures during the useful-life period. J. Wu et al. (Wu, McHenry, & Quandt, 2013) applied a two and three parameter Weibull distribution to determine failure rates in automotive components.

**Stochastic processes:** This method predicts warranty claims using stochastic processes. Kalbfleisch et al. (Kalbfleisch & Lawless, 1991) uses a log-linear Poisson model whereby reports of warranty claims can be estimated to obtain the expected number of warranty claims per unit in service as a function of the time in service. Kaminskiy & Krivtsov (Kaminskiy & Krivtsov, 2000) develop warranty prediction models based on three stochastic processes: the G-renewal process that is a generalized renewal process introduced by Kijima and Sumita (Kijima & Sumita, 1986), the ordinary renewal process and the non-homogeneous Poisson process. K.D. Majeske (Majeske, 2007) developed a non-homogeneous Poisson predictive model for automobile warranty claims.


Each category has many successful studies and implementations for predicting component lifetimes. The advantages and disadvantages of each method will be discussed below.

Neural networks have several advantages and disadvantages (Tu, 1996). They have a good track record for finding complex non-linear relationships between independent and dependent variables. They also
require less formal statistical training by the user. It is possible to use multiple different training algorithms, one of which might perform better. Neural networks viewed by many as being magic hammers which can solve any machine learning problem. Therefore, people tend to apply them indiscriminately to problems for which they are not well suited. It is often better to study the situation and data more in-depth and then find a technique with a strong theoretical foundation. Also, neural networks are a black box which makes training them difficult. The training outcome depends mostly on the number of initial parameters, where more parameters will always result in a better fit. This makes the method prone to overfitting. Overfitting means that the model represents the training data very well, but performs badly when it is evaluated with other data of the same process. It is hard to troubleshoot the model when it is not working as expected, as it is unknown how the model is solving the problem. When it returns a good fit, it is hard to feel confident as it might have been over fitted to the training data. Unlike the statistical methods that give a confidence interval based on the theoretical foundation of the method, neural networks are unable to do so as they lack the theoretical foundation. Concluding, neural networks seem to be a good method for exploratory analysis on large datasets where there is little knowledge on the relationships of variables. As knowledge about the data increases, specific techniques with a good theoretical foundation outperform neural networks. For this study, there is a good understanding on the relationships within the data. There are also plenty techniques available with a good theoretical foundation to get the required results.

The differences in usability between lifetime distributions and stochastic processes are quite small. Several advantages and disadvantages are discussed (Wu & Akbarov, 2011) (Mitzenmacher, 2004). Stochastic processes use failure rates instead of using the exact failure date of each component. This may cause information loss as they are obtained as a ratio of repair claims to the number of products in service (i.e. they integrate two observations into one). Stochastic processes are often built on the assumption that the claim rates follow a specific form such as the power law for non-homogeneous Poisson process (Kaminskiy & Krivtsov, 2000) (Majeske, 2007). This assumption might not hold. Also is it hard to evaluate this assumption as it looks very similar to a log-normal distribution, but return different end results. Finally, as power laws are very heavy tailed distributions, misrepresenting the tail of the distribution could have severe consequences (Mitzenmacher, 2004). Lifetime distributions offer more flexibility as there are multiple non-parametric distributions available to fit on the data. One danger to fitting probability distributions is that one might assume that the data represents the exact number of failed products, for example in (Rai & Singh, 2005). This might not hold due to the effects described in “APPENDIX B: Delayed reporting of defects” where component failures are not reported or not detected immediately upon failure. For this case, the effect of non-reporting will be smaller than normal. This is because trucks undergo rigorous inspections during maintenance and mandatory MOT tests (APK in Dutch). It could still mean that the report is delayed until the next inspection. Furthermore, none of the techniques have a mechanism to value recent received repair claims over older repair claims in the dataset. Recent failures could be more important as manufacturing faults are gradually removed over time as the production processes improve.

A summary of the advantages and disadvantages is given in Table 3.
Table 3: Advantages and disadvantages of the different prediction techniques

<table>
<thead>
<tr>
<th>Prediction technique</th>
<th>Advantages</th>
<th>Disadvantages</th>
</tr>
</thead>
</table>
| Lifetime distributions     | Strong theoretical foundation  
Many different statistical distributions available  
Method is known to DAF | Assumption that data represents the exact number of failures might not hold |
| Stochastic processes       | Strong theoretical foundation                                                | Information loss due to use of failure rates  
Power law assumption might not hold |
| Artificial neural networks | Finds complex relationships in large databases  
Require less formal statistical training  
Good for exploratory analysis in large databases | Only exploratory uses  
No theoretical foundation  
Black box makes training and troubleshooting difficult  
No confidence intervals on results  
Prone to overfitting |

Concluding, lifetime distributions seem to be a good choice for predicting component lifetimes in this research project. Stochastic processes could still be interesting to investigate further, but are less likely to provide good results compared to lifetime distributions as they require rigid assumptions. DAF already has a basic prediction model using the Weibull distribution to predict component lifetimes. So besides the probable technical advantages, it will be easier to introduce a new method based on lifetime distributions as they are already familiar with it. Neural networks will not be used for the various reasons discussed above.

3.2 Predicting component lifetime method

Predicting component lifetime using lifetime distributions is done by taking a statistical distribution and configuring the parameters of the function in such a way that it represents the component lifetime as closely as possible. After a representable distribution has been set, it is possible to calculate the chance to failure within an interval. This section explains precisely how the data is transformed into an expected component lifetime usable for DAF.

3.2.1 Choosing appropriate distributions to fit on the available data

Many statistical distributions have been developed over the course of history. Each distribution was once constructed for a specific function. As time passed, it could happen that the same distribution proved useful for multiple scenarios. Distributions are given a name when they are really important for theory or applications. This section investigates which distributions are useful for simulating the lifetime of truck components. “APPENDIX D: Basic statistics” explains the statistical terms required to understand this and the following sections.

There are many distributions, at least over one hundred, to choose from. Every distribution belongs to a family. Every unique family can be differentiated by the following characteristics:
• Continuous or discrete
• Finite, infinite, our bounded support
• Skewness behavior

The easiest division is whether a continuous or discrete distribution is required. As a truck component is able to fail at any random time, a continuous distribution is required.

For components in normal operation, it is not possible to fail before use. This means that the distribution representing the component lifetime should not have support on negative values. Theoretically there is no hard value after which all components have failed. This means that a non-negative distribution is required.

The skewness behavior does not necessarily have to be the same for each component. Lifetime data is often not normally distributed for several reasons. First, the wear-out phase causes positive outliers in the data. Secondly, multiple failure mechanics could be active within the same data set, effectively combining two distributions into one. The effect of both varies per component, meaning that there is not one best failure distribution for all components.

Figure 9: Distribution choice flow chart (Damodaran, NA).

Figure 9 shows a flow chart which can help choosing a representative distribution. Based on the characteristics of lifetime data, only continuous non-negative distributions are required. Still, there are many distributions fitting these requirements, one more known than others. “APPENDIX E: List of non-negative distributions” shows an extensive list of the most popular distributions. Based upon expert knowledge at DAF and uses in scientific studies for lifetime studies, the following distributions will be fitted on the available data:
• Exponential distribution (1 parameter)
• Lognormal distribution (2 parameters)
• Gamma distribution (2 parameters)
• Weibull distribution (2 parameters)

The minimum extreme distribution from Figure 9 is not chosen to model component lifetimes because it is not bounded below by zero (it is not a non-negative distribution).

3.2.2 Estimating the values for the parameters of the distributions

Fitting distributions is possible using several methods. Popular methods are Maximum Likelihood Estimation (MLE) and Rank Regression (RR). The “Automated Weibulls” that DAF uses to signal big deviations in predicted component lifetime makes two predictions, one with MLE and the other with RR. However, RR is not able to handle the heavily censored data that DAF has. A detailed explanation is given in “APPENDIX F: Why Rank Regression is not suitable for heavily censored data”. For the reasons shown in the appendix, it is recommended that DAF does not use the RR method anymore. For this research project, MLE will be used to fit the parameters of the distribution to the data.

Maximum Likelihood Estimation (MLE), a method for estimating the values of the parameters of a statistical distribution given data (NIST/Sematech, 2003). MLE can be split into two parts, calculating the likelihood and maximizing the likelihood.

The first part is calculating the likelihood \( L \) using the likelihood function. The likelihood function indicates “given a data set, what is the likelihood of obtaining exactly that data set following the chosen probability distribution function”. This likelihood changes as the parameter values of the distribution are changed. This likelihood must be maximized for getting the function that mimics the data as closely as possible. The likelihood function is shown below:

\[
L = \prod_{i=1}^{n} f(x_i; \theta_1, \theta_2, \ldots, \theta_k) \tag{2}
\]

\( L \) is the likelihood function, \( n \) is the number of observed defects, \( f(x_i; \theta_1, \theta_2, \ldots, \theta_k) \) is the probability density function (PDF) with \( x_i \) as the \( i \)th observed (time to defect) data, and \( \theta_1, \theta_2, \ldots, \theta_k \) as the parameters to be estimated. For a two parameter Weibull distribution, the parameters would be beta (\( \beta \)) and eta (\( \eta \)). For complete data, the likelihood function is the product of the PDF functions, with one element for each data point in the data set. For this project, it is known that the data is right censored. Data is right censored when an observation of a component ended while there was no observed defect, i.e. the component still works at the end of the warranty period. This makes the likelihood function slightly more difficult. The likelihood function for right censored data:

\[
L_{censored} = \prod_{i=1}^{n} f(x_i; \theta_1, \theta_2, \ldots, \theta_k) \times \prod_{j=1}^{m} [1 - F(y_j; \theta_1, \theta_2, \ldots, \theta_k)] \tag{3}
\]
where \( n \) is the number of observed defects, \( m \) is the number of censored data points, \( F \) is the cumulative distribution function with \( y_j \) as the \( j \)th observed censored data. Note that the MLE methodology takes into account the values of the time until censored, which is why the MLE parameters from the case example in Table 27 were different. A higher likelihood means that the distribution has a higher chance to mimic the underlying data. Therefore, the likelihood function will be maximized.

The maximization of the likelihood function is done using the software program R-Studio (R-Studio, 2016) with the flexsurv package (Jackson, 2016). The maximization method has been completely automated for the mathematical model. R code for the automation of this part can be found in chapter 5. Finding the parameters that result in the highest value for this function is commonly done by taking the partial derivative of the log-linear equation for each parameter and setting it equal to zero. This results in a number of equations with an equal number of unknown, which can be solved simultaneously. The values that maximize the likelihood are known as the Maximum Likelihood Estimates, or MLE’s.

\[
MLE = \max(L_{\text{censored}})
\]  
Eq. 4

A typical MLE output example using real data is shown in Figure 10 for a component that had 39310 observed components (of which 5 are observed defects and 39305 were censored (fictitious due to confidentiality reasons)):

Estimates:
\[
\begin{align*}
\text{est} & \\
\text{shape} & = 0.53 \\
\text{scale} & = 23564
\end{align*}
\]

\[
N = 39310, \quad \text{Events: 5, Censored: 39305} \\
\text{Total time at risk: 31968637} \\
\text{Log-likelihood = -275.9954, df = 2} \\
\text{AIC = 555.9908}
\]

Figure 10: Maximum Likelihood Estimation for fitting a Weibull distribution.

The function shows the beta (shape in figure) and the eta (scale in figure) of the distribution, which are 0.53 and 23.564. \( N \) is the total number of trucks observed. Events are the number of observed defects. The trucks that did not have a repair during the observation period are shown as censored. df=2 means that the fitted distribution function had two degrees of freedom, or two parameters. The AIC is the Akaike Information Criterion, which will be explained and used in the next section for choosing the best distribution.

3.2.3 Comparing goodness-of-fit of each distribution

Four distributions have been fitted to the data in section 3.2.2, but only one will be used for predicting the component lifetime. A choice has to be made which distribution has the best fit on the data. This will be done using a graphical estimation and the Akaike Information Criterion (AIC) (Akaike, 1974). The AIC is a relative measure of the quality of a statistical distribution given data. It returns a relative estimate of the information lost when using the statistical distribution to generate the data. Information
lost stands for the error between the chosen component lifetime distribution and the true component lifetime distribution. The AIC is similar to the likelihood-ratio test, except that the likelihood-ratio test is only valid for nested models\(^1\). The AIC has no such restriction.

The maximized value of the likelihood function is required to calculate the AIC. This has already been calculated in section 3.2.2. Let \( L \) be the maximized value of the likelihood function for the model as explained in section 3.2.2. Let \( k \) be the number of estimated parameters (1 for the exponential distribution, 2 for the others) in the model. The AIC value of the model is then the following (Akaike, 1974):

\[
AIC = 2k - 2\ln(L)
\]  
\text{Eq. 5}

To quickly verify that the software package is calculating the AIC correctly, Figure 10 can be used as an example. The log-likelihood (-275,9954) and the number of parameters (2) is given for the fitted distribution. The AIC is then calculated: \( AIC = 2 \times 2 - 2 \times -275,9954 = 555,9908 \). After the AIC is calculated for each model, they can be compared to each other. The preferred distribution is the one with the lowest AIC value (Eq. 6). Looking at the formula, it penalizes the model for having more parameters to prevent overfitting, and it rewards a higher goodness-of-fit as assessed by the likelihood function.

\[
\text{Chosen distribution} = \min(AIC_i), \quad \text{for all } i
\]  
\text{Eq. 6}

where \( i \) is the index for the distributions, 1 = Weibull, 2 = lognormal, 3 = gamma, 4 = exponential.

Next to a relative best fit, it is also desirable to confirm whether the data follows the fitted distribution on an absolute scale as well. For regression techniques such as rank regression this is most often done by using the correlation coefficient (\( R^2 \)) method. Maximum likelihood estimation is not a regression technique, so other techniques have to be used. Popular techniques are the Kolmogorov-Smirnov test for normality, the Anderson-Darling test, Shapiro-Wilks, and the Cramer-von Mises criterion. The tests are all quite similar as they all compare the chosen cumulative distribution function with an empirical distribution function. None of the tests are created for handling censored data. There are modifications made to the abovementioned tests that allow for censored data. Detailed explanations on how they test the goodness of fit are given in the following reference: (Nikulin, Lemeshko, Chimitova, & Tsivinskaya, 2011). A problem with goodness-of-fit tests occurs when sample sizes become large. A null hypothesis taken literally (which is the only way you can take in formal hypothesis testing) is always false in the real world. If it is false, even in the tiniest amount, then a large enough sample size will produce a significant result leading to the rejection of the hypothesis (Lin, Lucas Jr, & Shmueli, 2013) (Sullivan & Feinn, 2012)

\(^1\) Two statistical models are nested if the first model can be transformed into the second model by imposing constraints on the parameters of the first model. For example the Weibull distribution can be transformed into the exponential distribution by setting the shape (\( \beta \)) parameter equal to 1.
(Bjorn, 2013). With the tens of thousands observations from the DAF database for each component, a hypothesis test becomes extremely powerful. Any observation that is even slightly deviating from the fitted distribution will cause the formal goodness-of-fit test to reject the null hypothesis that the data follows the distribution. Steven P. Millard programmed an R package for the adjusted Shapiro-Francia goodness of fit test, usable for normal and lognormal distributions. Four random component numbers (component numbers 62901, 40250, 55410 and 60401) from the DAF XF FT series were tested where the lognormal distribution was chosen as the best distribution by the mathematical model (using AIC). Each component number has almost 40,000 components observed. All four tests concluded that the true data distribution was not similar to the fitted lognormal distribution with a rounded p-value of 0.00000. This further strengthens the point that the known hypothesis tests are not useful in determining the goodness of fit for this research project. A good alternative for an absolute goodness of fit test is to create Quantile-Quantile plots (Q-Q plots). The Q-Q plot is a graphical technique for determining if two data sets come from populations with a common distribution. A quantile is the percent of points below the given value. For example the 5% quantile is the point at which 5% of the data fall below and 95% fall above that value. Q-Q plots take your sample data, sort it in ascending order, and then plot them versus quantiles calculated from a theoretical distribution. If the data are similar, the points in the Q-Q plot will approximately be on the line $y = x$. Figure 11 shows the Q-Q plots for the same four component numbers that were used in the Shapiro-Francia hypothesis test earlier. The fits shown in the figure are also representable for most of the components in this research.
The Q-Q plots in Figure 11 show that the component numbers follow the fitted line quite good, unlike the results from the Shapiro-Francia hypothesis tests that rejects a good fit without doubt because of the very high number of observations.

To be able to compare different distributions graphically, a different approach is needed. Figure 12 shows for one component number each of the four fitted distributions (Weibull, lognormal, gamma, exponential). The data and the scales of the graph are identical, the fitted distribution changes per graph. This makes it possible to determine manually whether the fitted distribution is acceptable. The mathematical model chooses for the lognormal distribution (top right) based on the AIC method explained earlier.
Figure 12: Graphical representation of the four distributions to assess goodness of fit for component number 40250 (“Pulleys near the engine region”). The thick black crosses represent the Kaplan-Meier empirical distribution (the data), the red line is the fitted distribution. The dashed lines are the 95% confidence intervals for the Kaplan-Meier empirical distribution.

For each component that is automatically processed by the mathematical model, these four plots will be generated too for the user to evaluate the goodness of fit if the user wants to. Evaluating the individual plots for this component shows that all fits are acceptable, except for the exponential fit which seems to be underestimating the data for a long period. The lognormal (lnormal fit) function has the lowest AIC, and indeed looks to follow the data better than the other models. As explained earlier, the mathematical model will choose the distribution with the lowest AIC. The R code that is used to create the Q-Q plots and the plots above is described in chapter 5.

Concluding, the mathematical model will use the AIC method to choose the distribution that has the best relative fit on the data. The engineer that uses the predictions can look at the absolute goodness-of-fit via the graphical representation when he wants to know more.

3.3 Extension of the method to handle predictions for improved components

Technology never stands still, neither does the market. Truck components get improved over time as a result of several factors. This could be changing customer needs, increasing reliability, decreasing costs, design changes, adding features, etc. Section 1.1.3.3 already introduced the existence of component
improvements. Every time a component is improved, its lifetime will also change. This section extends the mathematical model to recognize and react to a component improvement.

Component improvements make doing predictions difficult because the lifetime of the improved component will differ from the old component. This means that the available historic data about the old component is less useful, depending on the improvement. One of the scenarios could be that the reliability of the component was improved: then, without separating the data from the old and the new component, it will be likely that the component lifetime prediction of the improved component will be too pessimistic. Another scenario could be that the reliability of the component was not changed after the component was changed: disregarding the available data from the old component is then basically throwing away valuable information.

The available data on component improvements was described in section 1.1.3.3. It is well specified when a component is improved, so it is easy to separate the trucks with the old components from the trucks with the new components. Unfortunately, the reason for the component improvement is often poorly written in the available database. There is also no information available that says anything about the change in component lifetime.

The only real information about component improvements comes from the time of improvement, and which series are affected. Deciding whether to use the old data must come from monitoring and comparing the new data with the old data. If the new data shows very similar behavior to the old data, the old data will not be deleted. Section 3.3.1 describes the method on how the data is compared.

### 3.3.1 Comparing data similarity with the Kaplan-Meier estimator

There are two sets of data on component lifetime, one set with data from before the component improvement, the other from after the component improvement. A component improvement does not necessarily mean that the lifetime of the component is different. In order to maximize the available information for fitting a distribution, the two datasets will be compared.

Comparing both datasets will be done using the Kaplan-Meier estimator, also known as the product limit estimator (Kaplan & Meier, 1958). The Kaplan-Meier estimator is a non-parametric statistic used to estimate the survival function from lifetime data. Survival means that a component has not had a defect during the observed time. It is often used to measure the time-to-failure of machine parts. Other uses are measuring the fraction of patients living for a certain amount of time after treatment, the length of time people remain unemployed after a job loss, etc.

The Kaplan-Meier estimator is as follows:

\[
\hat{S}(t_{days}) = \prod_{i=1}^{t} \left( \frac{n_i - d_i}{n_i} \right) 
\]

Eq. 7

\(\hat{S}(t)\) is the probability of survival of one component up to time t (t in days). \(n_i\) is the number of components still functioning (at risk) just prior to time \(i\), where \(i\) is the day counter. \(d_i\) is the number of
observed defects at time $i$. For this project time is measured in days. A “survival rate” is calculated for each day between the first and $t$. The Kaplan-Meier estimator is used to estimate the component survival chance for a period up to time $t$ by taking the product of each individual survival rate. Because a component observation can end without breaking down (censored), it is not “at risk” anymore to be observed when it breaks down. Dealing with right censored data is done by not counting them with $n_i$ “at risk” the day after the observation ends. Table 4 shows an example of how the Kaplan-Meier estimator works in practice. In this example, the probability of survival of a fictitious component with 100 observations is given for the first five days. The number of censorings and observed defects are given in the table. The result is $\hat{S}(5) = 0.1730$

Table 4: Example for calculating the Kaplan-Meier estimate

<table>
<thead>
<tr>
<th>Time period</th>
<th>At Risk</th>
<th>Became unavailable (Censored)</th>
<th>Observed defects</th>
<th>Survived</th>
<th>Kaplan-Meier Survival Probability Estimate</th>
</tr>
</thead>
<tbody>
<tr>
<td>Day 1</td>
<td>100</td>
<td>3</td>
<td>5</td>
<td>95</td>
<td>$(95/100)=0.95$</td>
</tr>
<tr>
<td>Day 2</td>
<td>92</td>
<td>3</td>
<td>10</td>
<td>82</td>
<td>$(95/100) \times (82/92)=0.8467$</td>
</tr>
<tr>
<td>Day 3</td>
<td>79</td>
<td>3</td>
<td>15</td>
<td>64</td>
<td>$(95/100) \times (82/92) \times (64/79)=0.7$</td>
</tr>
<tr>
<td>Day 4</td>
<td>61</td>
<td>1</td>
<td>20</td>
<td>41</td>
<td>$(95/100) \times (82/92) \times (64/79) \times (41/61)=0.4611$</td>
</tr>
<tr>
<td>Day 5</td>
<td>40</td>
<td>1</td>
<td>25</td>
<td>15</td>
<td>$(95/100) \times (82/92) \times (64/79) \times (41/61) \times (15/40)=0.1730$</td>
</tr>
</tbody>
</table>

A good visualization is to plot the Kaplan-Meier estimator over time. Figure 13 shows two Kaplan-Meier curves for component number 15455 (Dosing Module EAS) before and after the improvement. It can clearly be seen that the improved component survival behavior is much better than the old component. After 500 days the new component has a Kaplan-Meier survival probability which is high while the old component has a survival probability which is lower. For this situation, it would not be beneficial to use the data from the old and the new component to predict the component lifetime.
Figure 13: Comparing two empirical Kaplan-Meier survival plots of the same component before and after the improvement (component number 15455 “Dosing Module EAS”).

Figure 14 shows an example where the component improvement does not show a change in the component lifetime. After 800 days, the improved component has a high survival probability. The old component has a survival probability that is very similar. Both lines in that figure are very close and thus it should be beneficial to combine both datasets to get a better prediction for the component lifetime.
Figure 14: Comparing two empirical Kaplan-Meier survival plots of the same component before and after the improvement. Notice the very small scale on the y-axis.

It is also possible to compare two or more Kaplan-Meier estimates for similarities using statistical methods. This can be done with the log-rank test, a hypothesis test to compare the survival distribution of two samples. The null hypothesis $H_0$: there is no difference between the populations in the probability of an event (here a defect) at any time point. The log-rank test is well known and very easy to calculate, a good explanation can be found at (Bland & Altman, 2004). Shortly, an expectation of defects is calculated for the old component group and new component group similar to the Kaplan-Meier method. Then, the null hypothesis is tested using the Chi square $\chi^2$ distribution. The null hypothesis is tested using the test statistic calculated for each group as the sum of $\frac{(O - E)^2}{E}$ where $O$ is the number of observed events and $E$ is the number of expected events. The degrees of freedom are the number of groups minus one, so one for this situation. Significance is chosen to be at $P<0.05$ in consultation with DAF, meaning that at $P<0.05$ it is concluded that the component improvement did indeed change the component lifetime.

\[
P \text{ value (reject } H_0) = \chi^2 \left( \sum_{i=1}^{n} \frac{(O_i - E_i)^2}{E_i} \right) < 0.05, \quad df = n - 1
\]  

where $i$ is the group index (1 is the new group and 2 is the old group) and $n$ is the number of groups. The number of groups is always 2 for this research as there is always one old component which is compared to one new. $df$ is the degrees of freedom, which is always 1 because there are always 2 groups.

When both components show the same component lifetime before and after the improvement ($H_0$ not rejected), the data from both groups will be used to predict the component lifetime. When the old component has a significantly different component lifetime, only the data from the new component will be used to predict component lifetime.
3.4 Evaluating the performance of the predictions

There are two main aspects that are interesting to evaluate:

1. Evaluating the prediction accuracies
2. Evaluating the distribution of the chosen lifetime distributions

To be able to evaluate these aspects, a general evaluation setup will be developed in the next section. Then, the evaluation setup for both aspects will be described in separate subsections.

3.4.1 General evaluation setup

Making the predictions with the complete dataset gives just one snapshot. To be able to evaluate the performance of the predictions over time, more snapshots are needed. It would be interesting to see how the predictions were if they were done a year ago, or two years ago, as this says something about the quality of the predictions with less data. The available data is therefore split into eight cumulative subsets. The smallest subset simulates how the predictions would have been seven years ago, when there was just one year data available. This subset would only contain trucks that are delivered in 2008, and contain only observed defects that happened in 2008. The next subset would be as if it were one year later compared to the smallest subset. In this subset, data will be from trucks delivered in 2008 and 2009, with defects observed in 2008 and 2009. Continuing, each new subset will have one more year worth of data. Because this research project has data available from beginning 2008 until end 2015, there are eight subsets in total. A graphical representation of the subsets is shown in Figure 15.

![Figure 15: Each subset of the complete dataset contains an additional year of data.](image)

For the rest of this evaluation, the subsets will be named subset 1 for the smallest subset, to subset 8 for the largest subset. Table 5 shows the distribution of the truck series over each subset.
Table 5: The number of observed trucks in each subset

<table>
<thead>
<tr>
<th>Subset</th>
<th>Number of XF trucks observed</th>
<th>Number of CF trucks observed</th>
<th>Number of LF trucks observed</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 (1 yr data)</td>
<td>2632</td>
<td>446</td>
<td>358</td>
</tr>
<tr>
<td>2 (2 yrs data)</td>
<td>4731</td>
<td>840</td>
<td>630</td>
</tr>
<tr>
<td>3</td>
<td>8496</td>
<td>1506</td>
<td>1317</td>
</tr>
<tr>
<td>4</td>
<td>14178</td>
<td>2159</td>
<td>2143</td>
</tr>
<tr>
<td>5</td>
<td>19280</td>
<td>2664</td>
<td>3336</td>
</tr>
<tr>
<td>6</td>
<td>25141</td>
<td>3239</td>
<td>4147</td>
</tr>
<tr>
<td>7</td>
<td>32782</td>
<td>4068</td>
<td>5448</td>
</tr>
<tr>
<td>8</td>
<td>39948</td>
<td>4701</td>
<td>6688</td>
</tr>
</tbody>
</table>

All parts in the mathematical model have been programmed in the statistical software program R (R-Studio, 2016). All the test results and choices made are stored in a detailed table, shown in Table 7. A description of each row and column is shown in Table 6. This report screen contains every result necessary to do the evaluation. The evaluation is also programmed in R (code in chapter 5), allowing for evaluations on a component level that would otherwise not be possible. Doing the evaluations per component ensures maximum accuracy of the evaluation. Because it is not possible to show the results per component individually, they will be shown on an aggregated level. For DAF, having the mathematical model working completely autonomously is already very valuable. It means that implementation in the current business processes will be very easy. It also allows doing predictions for different datasets or other ad-hoc analysis purposes.

Table 6: Description of each row and column in Table 7

<table>
<thead>
<tr>
<th>X yrs data (columns)</th>
<th>Subset used as explained in Figure 15</th>
</tr>
</thead>
<tbody>
<tr>
<td>Component number</td>
<td>Component number</td>
</tr>
<tr>
<td>Improved date</td>
<td>Shows the latest improvement data, if any</td>
</tr>
<tr>
<td>Use old data</td>
<td>Results of the hypothesis test of Eq. 8. TRUE if H0 is not rejected, FALSE if H0 is rejected. TRUE meaning that old data will be used (section 3.3, Eq. 8)</td>
</tr>
<tr>
<td>Observed defects</td>
<td>Number of observed defects in each subset</td>
</tr>
<tr>
<td>Defect outlier peak size</td>
<td>Number of observed defects that occurred just before the end of the contract (Appendix C, Eq. 37)</td>
</tr>
<tr>
<td>Trucks</td>
<td>Number of trucks observed with this component per subset</td>
</tr>
<tr>
<td>Dist</td>
<td>Name of the statistical distribution that best fits the available data (section 3.2.3, Eq. 6)</td>
</tr>
<tr>
<td>Dist shape</td>
<td>Shape parameter of the chosen statistical distribution (Figure 10)</td>
</tr>
<tr>
<td>Dist scale</td>
<td>Scale parameter of the chosen statistical distribution (Figure 10)</td>
</tr>
<tr>
<td>Maturity coefficient</td>
<td>Shows the stability indicator of the prediction, explained in chapter 4</td>
</tr>
<tr>
<td>FphV after X years</td>
<td>Predicted failures per hundred vehicles between delivery to the customer and the end of year X</td>
</tr>
</tbody>
</table>
Table 7: Summary screen showing the most important results of the mathematical model for component number 30508 (pinion bearing cage seal, rear axle).

<table>
<thead>
<tr>
<th>component number</th>
<th>1 yr data</th>
<th>2 yrs data</th>
<th>3 yrs data</th>
<th>4 yrs data</th>
<th>5 yrs data</th>
<th>6 yrs data</th>
<th>7 yrs data</th>
<th>8 yrs data</th>
</tr>
</thead>
<tbody>
<tr>
<td>improved date</td>
<td>30508</td>
<td>30508</td>
<td>30508</td>
<td>30508</td>
<td>30508</td>
<td>30508</td>
<td>30508</td>
<td>30508</td>
</tr>
<tr>
<td>use old data</td>
<td>NA</td>
<td>NA</td>
<td>NA</td>
<td>NA</td>
<td>NA</td>
<td>NA</td>
<td>NA</td>
<td>NA</td>
</tr>
<tr>
<td>observed defects</td>
<td>7</td>
<td>17</td>
<td>25</td>
<td>40</td>
<td>54</td>
<td>71</td>
<td>75</td>
<td>76</td>
</tr>
<tr>
<td>defect outlier peak size</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>9</td>
<td>10</td>
<td>12</td>
<td>14</td>
<td>15</td>
</tr>
<tr>
<td>trucks</td>
<td>2608</td>
<td>4681</td>
<td>8433</td>
<td>14018</td>
<td>18933</td>
<td>24769</td>
<td>32350</td>
<td>39990</td>
</tr>
<tr>
<td>dist shape</td>
<td>0,00</td>
<td>1,51</td>
<td>1,54</td>
<td>1,63</td>
<td>1,63</td>
<td>1,64</td>
<td>1,63</td>
<td>1,62</td>
</tr>
<tr>
<td>dist scale</td>
<td>0</td>
<td>24584</td>
<td>24288</td>
<td>20916</td>
<td>20842</td>
<td>20350</td>
<td>21096</td>
<td>21792</td>
</tr>
<tr>
<td>maturity coefficient</td>
<td>NA</td>
<td>NA</td>
<td>NA</td>
<td>NA</td>
<td>NA</td>
<td>NA</td>
<td>NA</td>
<td>NA</td>
</tr>
<tr>
<td>FphV after 1 year</td>
<td>NA</td>
<td>0,18</td>
<td>0,15</td>
<td>0,14</td>
<td>0,14</td>
<td>0,14</td>
<td>0,13</td>
<td>0,13</td>
</tr>
<tr>
<td>FphV after 2 years</td>
<td>NA</td>
<td>0,50</td>
<td>0,45</td>
<td>0,42</td>
<td>0,43</td>
<td>0,42</td>
<td>0,42</td>
<td>0,40</td>
</tr>
<tr>
<td>FphV after 3 years</td>
<td>NA</td>
<td>0,91</td>
<td>0,84</td>
<td>0,82</td>
<td>0,83</td>
<td>0,82</td>
<td>0,80</td>
<td>0,78</td>
</tr>
<tr>
<td>FphV after 4 years</td>
<td>NA</td>
<td>1,41</td>
<td>1,30</td>
<td>1,31</td>
<td>1,32</td>
<td>1,32</td>
<td>1,28</td>
<td>1,24</td>
</tr>
<tr>
<td>FphV after 5 years</td>
<td>NA</td>
<td>1,97</td>
<td>1,84</td>
<td>1,87</td>
<td>1,89</td>
<td>1,89</td>
<td>1,84</td>
<td>1,78</td>
</tr>
<tr>
<td>FphV after 6 years</td>
<td>NA</td>
<td>2,58</td>
<td>2,42</td>
<td>2,51</td>
<td>2,54</td>
<td>2,55</td>
<td>2,47</td>
<td>2,38</td>
</tr>
<tr>
<td>FphV after 7 years</td>
<td>NA</td>
<td>3,24</td>
<td>3,06</td>
<td>3,22</td>
<td>3,25</td>
<td>3,27</td>
<td>3,16</td>
<td>3,04</td>
</tr>
<tr>
<td>FphV after 8 years</td>
<td>NA</td>
<td>3,95</td>
<td>3,75</td>
<td>3,98</td>
<td>4,02</td>
<td>4,05</td>
<td>3,91</td>
<td>3,77</td>
</tr>
</tbody>
</table>

Table 7 shows the summary screen (with fictitious values due to confidentiality reasons) for component number 30508, which is the pinion bearing cage seal on the rear axle. The columns represent the subsets as explained earlier. Notice that the first subset has only seven observed defects. The predictions are only done with a minimum of ten observed defects. Therefore, the first subset has several NA or “Not Available” values. As seen from the summary screen, there are no recorded component improvements. The defect outlier peak size indicates that there are a significant number of defects observed at the very end of the contract period. The distribution chosen to fit the available data is the Weibull distribution for each subset. It is also possible that a different distribution is chosen between the different subsets when the mathematical model sees that the fit is better. The maturity coefficient shows the stability of the predictions, this will be explained in detail in chapter 4. The failures per hundred vehicles (FphV) shows the expected failures between delivery to the customer and the end of year i (i = number of years). How the FphV values are calculated will be explained in this section.

The $F_{phV_{k,j,k}}$ values are the predicted failures per hundred vehicles (components) between delivery to the customer and the end of year i (i = number of years) while using the data from subset j (j yrs data from Figure 15) and for component number k. The FphV value is calculated in the mathematical model by calculating the cumulative distribution function (CDF) value at time i, where i is in days instead of years, and multiplying it by 100. Since there are four different distributions used in the mathematical model there are also four different CDFs, denoted by m: 1 = Weibull, 2 = lognormal, 3 = gamma and 4 = exponential.
The function for the cumulative distribution function is dependent on the chosen distribution. For the Weibull distribution \((m = 1)\) it is:

\[
CDF_1(i_{\text{days}}) = \left(1 - e^{-\left(\frac{i_{\text{days}}}{\eta}\right)^\beta}\right)
\]

where \(i\) is the number of years in days, \(\beta\) is the shape parameter, and \(\eta\) is the scale parameter.

An example can be given by looking at Table 7 and reproducing the FphV value from after 3 years for subset 3 (3 yrs data), noted as \(FphV_{3,3,30508}\). \(i_{\text{days}}\) is 1096, three years in days. The distribution can be read from the table as Weibull, so \(m = 1\). The shape parameter \(\beta\) can be read from the table as 1,54 and the scale parameter \(\eta\) is 24.288. The FphV value is then:

\[
FphV_{3,3,30508} = \left(1 - e^{-\left(\frac{1096}{24288}\right)^{1.54}}\right) \times 100 = 0.84
\]

With one year more data the FphV prediction, noted as \(FphV_{3,4,30508}\) has decreased to 0.82. The most accurate prediction is with subset eight, or eight years of data \(FphV_{3,8,30508}\), where the estimated FphV value after three years is 0.78 failures. The FphV values from Table 7 will be used for the evaluation of the predictions as they are easy to compare for every different chosen distribution.

The summary screen shown in Table 7 is for only one component. To get the performance of the whole mathematical model, hundreds of these components need to be compared to the Kaplan-Meier empirical distribution. Because the model has been programmed in R, the evaluation can also be programmed in R to compare each component individually. Showing the results individually however is not possible since it would take too much space. This means that the comparison values will be presented in an aggregated way. To do this, all the summary screens are added together to form a summary screen for the whole model. Only the FphV values are added since the others are not interesting for evaluating the accuracy of the predictions. An aggregated FphV (AFphV) can be created that sums \(FphV_{i,j,k}\) for all components \(k\).

\[
AFphV_{i,j} = \sum_k FphV_{i,j,k}, \hspace{1cm} \text{for all } k; \text{ for } i = 1,...,8; \text{ for } j = 1,...,8
\]
software program, so i are the columns, j are the rows. The $AFp\text{h}V_{i,j}$ values for all i,j are shown in Table 8 (fictitious values).

Table 8: $AFp\text{h}V$ values for each i,j (left) and the number of components in each subset (right)

<table>
<thead>
<tr>
<th>Subset</th>
<th>FphV after 1 year</th>
<th>FphV after 2 years</th>
<th>FphV after 3 years</th>
<th>FphV after 4 years</th>
<th>FphV after 5 years</th>
<th>FphV after 6 years</th>
<th>FphV after 7 years</th>
<th>FphV after 8 years</th>
</tr>
</thead>
<tbody>
<tr>
<td>Subset 1</td>
<td>3.62</td>
<td>7.09</td>
<td>10.49</td>
<td>13.81</td>
<td>17.06</td>
<td>20.23</td>
<td>23.34</td>
<td>26.37</td>
</tr>
<tr>
<td>Subset 2</td>
<td>8.80</td>
<td>23.60</td>
<td>37.40</td>
<td>55.03</td>
<td>71.08</td>
<td>84.66</td>
<td>96.84</td>
<td>108.58</td>
</tr>
<tr>
<td>Subset 3</td>
<td>11.29</td>
<td>29.28</td>
<td>55.09</td>
<td>87.34</td>
<td>124.26</td>
<td>165.25</td>
<td>138.06</td>
<td>260.60</td>
</tr>
<tr>
<td>Subset 4</td>
<td>10.65</td>
<td>28.80</td>
<td>56.96</td>
<td>96.89</td>
<td>151.35</td>
<td>221.92</td>
<td>300.45</td>
<td>378.58</td>
</tr>
<tr>
<td>Subset 5</td>
<td>9.91</td>
<td>26.54</td>
<td>53.02</td>
<td>93.11</td>
<td>157.86</td>
<td>255.06</td>
<td>358.03</td>
<td>456.94</td>
</tr>
<tr>
<td>Subset 6</td>
<td>9.57</td>
<td>25.76</td>
<td>52.40</td>
<td>94.81</td>
<td>163.90</td>
<td>274.25</td>
<td>406.96</td>
<td>541.26</td>
</tr>
<tr>
<td>Subset 7</td>
<td>9.65</td>
<td>25.67</td>
<td>52.66</td>
<td>97.68</td>
<td>171.52</td>
<td>285.57</td>
<td>430.07</td>
<td>578.27</td>
</tr>
<tr>
<td>Subset 8</td>
<td>9.97</td>
<td>26.25</td>
<td>53.19</td>
<td>99.04</td>
<td>174.88</td>
<td>286.82</td>
<td>429.20</td>
<td>580.40</td>
</tr>
</tbody>
</table>

In the first subset, only 15 component predictions have been made. All the others did not have more than the minimum observed defects at that time. Also, any components that had a recorded improvement during the measurement are not taken into account as it would be comparing two different component lifetimes.

The aggregated FphV values are then compared to the aggregate of the real FphV values. The real “failures per hundred vehicles (FphV)” can be calculated per component number by calculating the Kaplan-Meier FphV value:

$$FphVKM_{i,j,k} = (1 - \hat{S}(i_{days})) \times 100$$

where $\hat{S}(i_{days})$ is the Kaplan-Meier survival function as explained in Eq. 7, $i_{days}$ is the number of years in days (as the Kaplan-Meier is calculated per day), and $(1 - \hat{S}(i_{days}))$ is the Kaplan-Meier CDF. Similar to the $AFp\text{h}V_{i,j}$ used for the predictions, an aggregate Kaplan-Meier FphV value $AFp\text{h}VKM_{i,j}$ can be calculated for all components k. Since the data is most accurate with the most available data, only the Kaplan-Meier FphV values of the whole dataset (subset 8, or j = 8) have to be calculated.

$$AFp\text{h}VKM_{i,j} = \sum_k FphVKM_{i,j,k}, \text{ for all } k; \text{ for } i = 1, ..., 8; \text{ for } j = 8$$

The predicted FphV of each subset of data can then be evaluated with the actual FphV from the complete dataset. A problem is that comparing the aggregated predicted FphV values from the first subset (Table 8) against the aggregated Kaplan-Meier FphV values is that it would be comparing the sum of 15 FphV values from subset one against the 651 FphV values from subset 8 (complete dataset). To make a fair comparison between the different subsets, the aggregated Kaplan-Meier FphV values from the complete dataset are filtered to only contain the components that have a predicted FphV in each of
the subsets. So the first subset only has the Kaplan-Meier FphV of the same 15 components that had a predicted FphV value in the first subset. The FphV sums are shown in Table 9 (fictitious values).

Table 9: AFphVKM values for each i,j

<table>
<thead>
<tr>
<th></th>
<th>FphV after 1 year</th>
<th>FphV after 2 years</th>
<th>FphV after 3 years</th>
<th>FphV after 4 years</th>
<th>FphV after 5 years</th>
<th>FphV after 6 years</th>
<th>FphV after 7 years</th>
<th>FphV after 8 years</th>
</tr>
</thead>
<tbody>
<tr>
<td>subset 1</td>
<td>1,27</td>
<td>2,82</td>
<td>5,87</td>
<td>14,66</td>
<td>24,29</td>
<td>24,29</td>
<td>29,87</td>
<td>29,87</td>
</tr>
<tr>
<td>subset 2</td>
<td>4,34</td>
<td>10,47</td>
<td>23,20</td>
<td>51,90</td>
<td>86,95</td>
<td>92,87</td>
<td>104,29</td>
<td>104,29</td>
</tr>
<tr>
<td>subset 3</td>
<td>6,19</td>
<td>14,69</td>
<td>32,53</td>
<td>92,97</td>
<td>162,23</td>
<td>173,77</td>
<td>196,36</td>
<td>196,36</td>
</tr>
<tr>
<td>subset 4</td>
<td>7,34</td>
<td>17,42</td>
<td>39,77</td>
<td>113,06</td>
<td>201,12</td>
<td>233,13</td>
<td>268,76</td>
<td>268,76</td>
</tr>
<tr>
<td>subset 5</td>
<td>8,29</td>
<td>19,65</td>
<td>44,14</td>
<td>127,86</td>
<td>236,78</td>
<td>280,62</td>
<td>324,93</td>
<td>324,93</td>
</tr>
<tr>
<td>subset 6</td>
<td>8,92</td>
<td>21,28</td>
<td>48,47</td>
<td>140,73</td>
<td>262,30</td>
<td>305,91</td>
<td>371,78</td>
<td>371,78</td>
</tr>
<tr>
<td>subset 7</td>
<td>10,32</td>
<td>23,66</td>
<td>51,87</td>
<td>146,74</td>
<td>271,01</td>
<td>322,51</td>
<td>396,29</td>
<td>396,29</td>
</tr>
<tr>
<td>subset 8</td>
<td>10,58</td>
<td>24,10</td>
<td>52,53</td>
<td>148,11</td>
<td>273,21</td>
<td>324,72</td>
<td>398,49</td>
<td>398,49</td>
</tr>
</tbody>
</table>

To summarize, the first row (subset) had 15 components with more than 10 observed defects and no recorded improvements at any point in the dataset. For these components a FphV value has been predicted using only the data from the first subset. These FphV values are then aggregated and shown in Table 8. To evaluate the accuracy, a non-parametric Kaplan-Meier FphV is calculated for the same 15 components, only now with the complete dataset (subset j = 8) to get the best accuracy. These 15 FphV values are also aggregated and shown in Table 9. Comparing these two tables gives a fair comparison of the accuracy of the predictions. To make the comparison easier, the Kaplan-Meier FphV values are divided by the predicted FphV values:

\[
\text{Prediction accuracy}_{i,j} = \frac{AF\text{phVKM}_{i,j}}{F\text{phV}_{i,j}} - 1, \quad \text{for all } i,j
\]

where the nominator are the values shown in Table 9 and the denominator are the values shown in Table 8. The results are shown in Table 11.

The method in Eq. 13 works good to evaluate the performance of the predictions for the whole model. However, it does not show how the underlying components are behaving individually. The error is a measure of the difference between the predicted FphV and the Kaplan-Meier for all component k, for each subset j, for each “FphV after i years”:

\[
\text{Error}_{i,j,k} = F\text{phVKM}_{i,j,k} - F\text{phV}_{i,j,k}, \quad \text{for all } i,j,k
\]

These errors can be shown in a histogram by fixing i and j and showing all errors k. Since i and j both run from 1,…,8 it would mean that 64 (8*8=64) histograms need to be created to show the behavior of the underlying components. This is still too much to print on paper so the errors are aggregated by taking the sample standard deviation s of the errors for i,j,k.

\[
\text{SDError}_{i,j} = s(\text{Error}_{i,j,k}), \quad \text{for all } k
\]

The standard deviation values of the individual differences can be shown in a table just like the other tables, as shown in Table 12. Only interesting histograms will be shown and discussed.
Because nearly all of the trucks have been observed for three years but not longer, the evaluation power decreases dramatically after these three years. Therefore, the evaluation will focus on the performance of the predictions of the FphV after one, two, and three years. The rest of the comparisons will still be shown for completeness, but will not be used to evaluate the performance of the model itself. Fortunately, a longer empirical evaluation might be possible in a couple of years. Currently, the trucks that are observed longer than three years have one of the “Full care” of “Flex care” repair & maintenance contract. As explained in the scope, component numbers are not stored in claims that come from trucks with a “care” repair & maintenance contracts, except trucks from the UK as they use a different claim handling system. DAF is currently switching to a different claim handling system for the rest of the countries. The old system was called Service Claim Handling (SCH) and the new system is called Dealer Claim Entry (DCE). The new system, DCE, does in fact store component numbers from claims from longer term care packages. DCE is slowly being implemented in increasingly more countries. The switch dates for every country that is in this research are listed in Table 10.

Table 10: Switch to DCE per country for the countries that are currently included in the research. N/A means that DCE has not been implemented yet.

<table>
<thead>
<tr>
<th>Country</th>
<th>Switch to DCE date</th>
</tr>
</thead>
<tbody>
<tr>
<td>Germany</td>
<td>1-11-2013</td>
</tr>
<tr>
<td>France</td>
<td>29-3-2015</td>
</tr>
<tr>
<td>UK</td>
<td>N/A</td>
</tr>
<tr>
<td>Italy</td>
<td>18-8-2014</td>
</tr>
<tr>
<td>Spain</td>
<td>28-5-2014</td>
</tr>
<tr>
<td>Poland</td>
<td>22-1-2013</td>
</tr>
<tr>
<td>Romania</td>
<td>14-10-2014</td>
</tr>
<tr>
<td>Netherlands</td>
<td>25-2-2016</td>
</tr>
<tr>
<td>Belgium</td>
<td>N/A</td>
</tr>
<tr>
<td>Greece</td>
<td>5-11-2014</td>
</tr>
<tr>
<td>Czech Republic</td>
<td>4-5-2013</td>
</tr>
<tr>
<td>Portugal</td>
<td>13-10-2014</td>
</tr>
<tr>
<td>Hungary</td>
<td>31-7-2013</td>
</tr>
<tr>
<td>Sweden</td>
<td>15-9-2015</td>
</tr>
<tr>
<td>Austria</td>
<td>30-10-2013</td>
</tr>
<tr>
<td>Denmark</td>
<td>17-2-2015</td>
</tr>
<tr>
<td>Finland</td>
<td>N/A</td>
</tr>
<tr>
<td>Norway</td>
<td>22-6-2015</td>
</tr>
<tr>
<td>Ireland</td>
<td>N/A</td>
</tr>
<tr>
<td>Luxembourg</td>
<td>N/A</td>
</tr>
</tbody>
</table>

Poland was the first country to implement DCE in early 2013, with Germany following in late 2013. Currently it is 2016, so in a couple of years there should be more data available such that a comparison
with the empirical distribution function is possible beyond the first three years. This evaluation will focus on the first three years (FphV after 1, 2 and 3 years).

### 3.4.2 Evaluating the prediction accuracies of the mathematical model

The performance of the distribution fitting is an important part to evaluate. A better prediction will enable DAF to better calculate the costs of their service contracts. It also helps evaluating the reliability of their components and spot structural problems on their components early. This section evaluates the performance of the predictions for all components. How the evaluations are set up has been explained in the previous section. Table 11 shows the prediction accuracy of the mathematical model of each subset of each failures per hundred vehicles (FphV) time as calculated using Eq. 15. As said in the previous section, the evaluation of the performance of the model is only done for the first three FphV values, as nearly all of the trucks have not been observed longer than three years (see Figure 7). This means that the Kaplan-Meier empirical distribution function used to evaluate the predictions on is not a good representation of the complete dataset after the third year. The only reason it is added to the evaluation is to help explain any interesting phenomena.

**Table 11: Prediction accuracies (Eq. 15) for all i,j (left) and the number of components k in each subset (right)**

<table>
<thead>
<tr>
<th>Subset 1</th>
<th>FphV after 1 year</th>
<th>FphV after 2 years</th>
<th>FphV after 3 years</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.084</td>
<td>0.433</td>
<td>0.227</td>
<td>0.308</td>
</tr>
<tr>
<td>-0.036</td>
<td>0.250</td>
<td>0.121</td>
<td>0.100</td>
</tr>
<tr>
<td>0.086</td>
<td>0.334</td>
<td>0.074</td>
<td>-0.048</td>
</tr>
<tr>
<td>0.192</td>
<td>0.358</td>
<td>0.073</td>
<td>-0.086</td>
</tr>
<tr>
<td>0.403</td>
<td>0.533</td>
<td>0.123</td>
<td>-0.073</td>
</tr>
<tr>
<td>0.517</td>
<td>0.635</td>
<td>0.140</td>
<td>-0.067</td>
</tr>
<tr>
<td>0.535</td>
<td>0.614</td>
<td>0.154</td>
<td>-0.058</td>
</tr>
<tr>
<td>0.524</td>
<td>0.596</td>
<td>0.157</td>
<td>-0.051</td>
</tr>
</tbody>
</table>

In general, it can be seen that the prediction accuracy with less data (lower subset number) is worse than the predictions with more data. It can also be seen that the differences for the FphV after 1-3 years are negative. For subset two, “FphV after 1 year”, the real FphV calculated with the Kaplan-Meier function is 49.6% lower than the predicted FphV value. As more data becomes available for the model, the accuracy increases. Subset 6, “FphV after 1 year” has only a 4.8% difference. Because the differences are negative, it can be concluded that the mathematical model consistently underestimates the true component lifetime, or consistently overestimates the number of FphV.

To get an indication to the spread of the errors for every individual component, a histogram is shown in Figure 16 for all components the third subset for “FphV after 2 years”, using Eq. 16.
Figure 16: Histograms of the prediction errors per component for subset 3, FphV after 2 years. Left histogram shows every component, right histogram is without the one outlier at around -15 to improve readability.

The histogram shows that the differences between the predictions and the realizations are normally distributed with a slight negative skewness. There is one extreme outlier at around -15 which has a big influence on the cumulative FphV comparison that is done in Table 11. Ignoring the outlier, the predicted FphV is higher than the real Kaplan-Meier FphV for most of the components since the values in the histogram are mostly negative. Still, the differences are close to each other and look to be normally distributed, which means that the mathematical method is biased but consistent for the majority of the components.

Figure 17: Histogram of the prediction errors for all components for subset 5, FphV at year 1

Figure 17 shows how the individual differences behave for subset 5, “FphV after 1 year”, as calculated using Eq. 16. The predictions become increasingly more accurate as data gets added to the model, and the overestimation of the FphV becomes smaller. Table 12 shows the standard deviation for the comparison of each individual component, as calculated using Eq. 17.
Table 12: Standard deviation of the prediction errors for all subsets j and all FphV times i (Eq. 17)

<table>
<thead>
<tr>
<th>Subset</th>
<th>FphV after 1 year</th>
<th>FphV after 2 years</th>
<th>FphV after 3 years</th>
<th>FphV after 4 years</th>
<th>FphV after 5 years</th>
<th>FphV after 6 years</th>
<th>FphV after 7 years</th>
<th>FphV after 8 years</th>
</tr>
</thead>
<tbody>
<tr>
<td>Subset 1</td>
<td>0.267</td>
<td>0.797</td>
<td>1.894</td>
<td>6.656</td>
<td>9.570</td>
<td>9.769</td>
<td>14.862</td>
<td>15.071</td>
</tr>
<tr>
<td>Subset 2</td>
<td>0.272</td>
<td>1.348</td>
<td>3.633</td>
<td>6.603</td>
<td>7.830</td>
<td>8.912</td>
<td>10.504</td>
<td>10.701</td>
</tr>
<tr>
<td>Subset 3</td>
<td>0.227</td>
<td>1.033</td>
<td>2.453</td>
<td>5.546</td>
<td>7.821</td>
<td>8.795</td>
<td>10.486</td>
<td>11.501</td>
</tr>
<tr>
<td>Subset 4</td>
<td>0.124</td>
<td>0.606</td>
<td>1.594</td>
<td>4.259</td>
<td>6.770</td>
<td>9.320</td>
<td>12.462</td>
<td>14.099</td>
</tr>
<tr>
<td>Subset 5</td>
<td>0.073</td>
<td>0.340</td>
<td>0.994</td>
<td>3.373</td>
<td>5.635</td>
<td>8.683</td>
<td>11.265</td>
<td>13.174</td>
</tr>
<tr>
<td>Subset 6</td>
<td>0.052</td>
<td>0.197</td>
<td>0.605</td>
<td>2.796</td>
<td>4.957</td>
<td>7.233</td>
<td>10.261</td>
<td>12.026</td>
</tr>
<tr>
<td>Subset 7</td>
<td>0.084</td>
<td>0.158</td>
<td>0.265</td>
<td>2.457</td>
<td>4.737</td>
<td>6.522</td>
<td>9.935</td>
<td>11.905</td>
</tr>
<tr>
<td>Subset 8</td>
<td>0.032</td>
<td>0.052</td>
<td>0.036</td>
<td>2.390</td>
<td>4.788</td>
<td>6.395</td>
<td>9.311</td>
<td>10.912</td>
</tr>
</tbody>
</table>

The standard deviation of the second column is about twice as large as the first column, and the third column is about three times as large as the first column. This is expected since the underlying FphV values are cumulative, as can be seen in Table 8 and Table 9. Overall, the standard deviations are low, which could already be seen in the previous histograms. Also, the standard deviation of the 4th to 8th FphV value shoots up as the number of observed trucks are greatly reduced because most of the trucks are only observed for the duration of the three year warranty. This only says that the Kaplan-Meier empirical distribution is disagreeing with the predictions, but not whether the predictions are bad.

3.4.3 Evaluating the distribution of the chosen lifetime distributions

DAF has always assumed a Weibull distribution for all of the components. To test whether this is true, other suitable distributions have been added to the mathematical model: the gamma, lognormal and the exponential distribution. It has already been explained how the distributions fits are compared and how the best distribution is chosen. Now, it is interesting to see how many times each distribution is seen as the best fit, and how often this changes when more data is added to the mathematical model.

The chosen distribution as shown in the summary screen for each component (Table 7) can be used to get the distribution of the chosen distributions for each subset. Subset 1 has 26 components which have more than the minimum number of observed defects. For this comparison, components that had an improvement at any point in the dataset are not excluded, which explains why there are 7 more components in the first subset compared to Table 8. Looking at the individual component summary screens, 17 had the exponential distribution chosen by the mathematical model, 0 had gamma, 8 had lognormal and 1 had Weibull. This corresponds to 65% (17/26), 0% (0/26), 31% (8/26) and 4% (1/26) percentage of total respectively. The calculation that is made for subset 1 is repeated for subset 2 to 8 and the results are given in a graph, shown in Figure 18.
Figure 18 shows that all four distributions are used by the mathematical model. The gamma distribution is the least popular choice. The choice of the distributions are stable except for the early subsets where the exponential distribution is dominating the Weibull distribution. Subset 8, the subset with all the available data, is more closely examined in Figure 19 in order to get a more detailed overview.

In the above graph each component is categorized by the number of observed defects. The exact categories cannot be displayed due to confidentiality reasons. Over the different categories, the graph is similar to Figure 18 where the exponential distribution is popular with less data, and less popular with more data. The Weibull distribution is taking over the exponential distribution as more data is available, while the lognormal and the gamma distribution remain stable as more data is available. One
explanation for the higher number of exponential distributions with less data could come from the way that AIC compares the fits and that the exponential distribution is a special case of the Weibull distribution when the shape ($\beta$) is fixed to 1. The AIC penalizes distributions that have more parameters to prevent over-fitting. The exponential distribution has one parameter while the others have two. In the case that the Weibull shape is close to one after fitting, it will be the same as the exponential distribution. Then, because the exponential has fewer parameters, it will be preferred over the Weibull distribution. As more becomes known about the true failure distribution, the two-parameter distributions will be more flexible to approximate the true distribution.

The fact that the Weibull distribution has only a share of 34% as being the best fit for all the components means that the DAF’s assumption for always choosing the Weibull might not hold. It is interesting to see what the effects are on the prediction accuracies when the mathematical model is configured by locking the distribution choice to only Weibull. This means that the CDF calculations done in Eq. 9 are always Weibull. The Weibull prediction accuracy table is shown in Table 13. The Weibull standard deviation for each individual FphV error is shown in Table 14.

<table>
<thead>
<tr>
<th>Table 13: Prediction accuracies for all i,j (Eq. 15) when choosing only Weibull distribution</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
</tr>
<tr>
<td>subset 1</td>
</tr>
<tr>
<td>subset 2</td>
</tr>
<tr>
<td>subset 3</td>
</tr>
<tr>
<td>subset 4</td>
</tr>
<tr>
<td>subset 5</td>
</tr>
<tr>
<td>subset 6</td>
</tr>
<tr>
<td>subset 7</td>
</tr>
<tr>
<td>subset 8</td>
</tr>
</tbody>
</table>

The results for each subset in the first three FphV years are nearly identical to the results in Table 11. The prediction accuracies show larger differences in the later FphV values, but because the empirical distribution that is used for the evaluation is not strong enough in these years, no conclusion based on empirical evidence can be drawn about whether the Weibull distribution is always the best choice.
Table 14: Standard deviation of the prediction errors for all subsets j and all FphV times i (Eq. 17) when choosing only Weibull distribution

<table>
<thead>
<tr>
<th>Subset</th>
<th>FphV after 1 year</th>
<th>FphV after 2 years</th>
<th>FphV after 3 years</th>
<th>FphV after 4 years</th>
<th>FphV after 5 years</th>
<th>FphV after 6 years</th>
<th>FphV after 7 years</th>
<th>FphV after 8 years</th>
</tr>
</thead>
<tbody>
<tr>
<td>Subset 1</td>
<td>0.313</td>
<td>1.148</td>
<td>2.706</td>
<td>7.825</td>
<td>11.337</td>
<td>12.617</td>
<td>18.093</td>
<td>19.460</td>
</tr>
<tr>
<td>Subset 2</td>
<td>0.271</td>
<td>1.349</td>
<td>3.642</td>
<td>6.616</td>
<td>7.816</td>
<td>8.952</td>
<td>10.663</td>
<td>11.002</td>
</tr>
<tr>
<td>Subset 3</td>
<td>0.226</td>
<td>1.033</td>
<td>2.452</td>
<td>5.550</td>
<td>7.938</td>
<td>9.312</td>
<td>11.881</td>
<td>14.197</td>
</tr>
<tr>
<td>Subset 4</td>
<td>0.124</td>
<td>0.605</td>
<td>1.593</td>
<td>4.263</td>
<td>6.844</td>
<td>9.688</td>
<td>13.296</td>
<td>15.233</td>
</tr>
<tr>
<td>Subset 5</td>
<td>0.073</td>
<td>0.340</td>
<td>0.994</td>
<td>3.375</td>
<td>5.663</td>
<td>8.799</td>
<td>11.618</td>
<td>13.888</td>
</tr>
<tr>
<td>Subset 6</td>
<td>0.052</td>
<td>0.197</td>
<td>0.606</td>
<td>2.797</td>
<td>4.964</td>
<td>7.255</td>
<td>10.322</td>
<td>12.182</td>
</tr>
<tr>
<td>Subset 7</td>
<td>0.084</td>
<td>0.158</td>
<td>0.266</td>
<td>2.455</td>
<td>4.735</td>
<td>6.534</td>
<td>10.002</td>
<td>12.127</td>
</tr>
<tr>
<td>Subset 8</td>
<td>0.052</td>
<td>0.054</td>
<td>0.035</td>
<td>2.387</td>
<td>4.784</td>
<td>6.401</td>
<td>9.355</td>
<td>11.066</td>
</tr>
</tbody>
</table>

Table 14 also does not show any big differences in the standard deviation of the prediction errors for the first three FphV values compared to Table 12.

During the fitting procedure in the mathematical model each distribution was tried on every component. The distributions were then already compared with the Akaike’s Information Criterion (AIC). It showed that the Weibull distribution was not always the best fit for every component. This in itself is already a clear signal that DAF’s assumption for always using the Weibull distribution might not hold. An empirical conclusion with trucks that are mostly three years observed still leaves things open to discussion. A better conclusion will be possible in a few years because then it will be possible to add the trucks with a longer repair & maintenance contract (up to 8 years) to the evaluation from the countries that use the new claim system DCE. Therefore, it will be recommended that this evaluation is run again in several years when it is possible to empirically evaluate the mathematical model for more than just the first 3 years.
4 Determining component lifetime maturity

DAF currently classifies a component as mature when the engineers and after sales analysts are convinced that the latest predicted component lifetime stays stable in the future. A prediction done currently does not give this indication of stability since it only gives the parameters of the fitted distribution based on the data that is inserted in the mathematical model. This makes it difficult for DAF to know how much they can trust the predicted distribution parameters.

The mathematical model has to give an indication of component lifetime maturity. DAF currently does not have a clear operational definition for when a component is mature, or how to measure this. In this chapter an operational definition for component lifetime maturity will be developed. Then, the performance of the lifetime predictions for only mature components will be compared to the performance of the lifetime predictions as evaluated in chapter 3.

4.1 Defining operational definition for component lifetime maturity

The method for the component lifetime prediction is explained in the previous chapter. Four distributions are fitted on the lifetime data using the maximum likelihood estimation (Eq. 4) and the best distribution is chosen via the Akaike Information Criterion (Eq. 6) to predict the component lifetime. A cumulative distribution function (CDF) is calculated (Eq. 10) using the predicted distribution parameters which show the probability of a defect as a function of time (in days). This CDF is an estimation based on the available data at the time of the prediction. A mature component should be such that the estimated CDF predicted now is the same as the estimated CDF in the future.

\[
\text{Component maturity: } CDF_t - CDF_{t+1} \cong 0 \tag{18}
\]

where \( t \) is the time when the component lifetime prediction is made. It is not required for DAF to have the CDF stable over the whole function, they are mostly interested in the CDF of the fourth year because most of the repair & maintenance contracts are sold for a length of four years. Also, DAF uses the failures per hundred vehicles (FphV) often internally to indicate the component lifetime, so it will be used in this chapter to indicate the component lifetime as well. FphV is calculated the same as in chapter 3:

\[
FphV = CDF \times 100 \tag{19}
\]

except that in this chapter the estimated CDF and FphV value are a point estimate of the fourth year (as explained above). So for DAF, component maturity means that the FphV after four years (1461 days) predicted now is equal to the predicted FphV after four years predicted in the future.

\[
\text{Component maturity: } FphV_t(1461) - FphV_{t+1}(1461) \cong 0 \tag{20}
\]

where \( t \) is the time when the component lifetime prediction is made.

As it is not possible to look into the future, the component lifetime maturity will be determined by looking at the estimated FphV(1461) values in the past, with different amount of data. A series of component lifetime predictions in the past will be compared with each other. If the variation between
each of the estimated FphV values is low, it means that the component lifetime prediction has been stable in the past. When this is the case, it is probable that the component lifetime prediction will remain stable in the future as well.

Component maturity: \( s(F_{phV_{t-N}}; ...; F_{phV_{t-2}}; F_{phV_{t-1}}; F_{phV_{t}}) < \alpha \) \hspace{1cm} \text{Eq. 21}

where \( s \) is the standard deviation of the FphV values of the fourth year (1461 days), \( t \) is the time when the component lifetime prediction is made, \( N \) is the number of component lifetime predictions done in the past and \( \alpha \) is the cut-off value of the standard deviation to indicate maturity.

It is possible that the mathematical model switches the chosen distribution between one subset and the next. This could lead to a sudden change in estimated FphV which wouldn’t have happened if the mathematical model would have used either of the two distributions throughout all subsets. To prevent that the mathematical model switches the chosen distribution between two subsets during the maturity test, it is better to use just one distribution for all subsets instead of the best of four for each subset. The Weibull distribution is chosen as the distribution to compare the stability of the predictions. The Weibull distribution was most often chosen as the best distribution for a component during the evaluation of the distributions in section 3.4.3. Besides that, it is possible to model the exponential distribution with it as well when setting the Weibull shape parameter equal to one. The Weibull distribution can therefore be used as the best distribution for 64% (36%+28%) of the components as shown in Figure 18. This is a lot higher compared to the 32% for the lognormal and 4% for the gamma. For the component lifetime maturity test, the Weibull distribution will be used for all predictions in the past. As an added effect, using only one distribution makes it possible to have more predictions in the past as only 1 distribution needs to be fitted instead of 4 (detailed explanation in section 4.1.1).

Each component lifetime prediction done in the past is done by creating a subset of the complete dataset. This subset simulates how a component lifetime prediction would have looked like when made \( x \) time ago. This is also done during the evaluation of chapter 3 where 8 subsets were created (subset 1 was the smallest, subset 8 the largest). In this chapter, the estimated FphV values in the past are also calculated by using subsets of the complete database. The difference to chapter 3 is that the number of subsets and the time between each subset will be redefined. The subset index will be called \( b \). There will be 12 subsets for determining component maturity, and each next subset will have 2 months more data compared to the previous subset (why 12 subsets and 2 years is explained in section 4.1.1). So in total there is a two year time difference between the first subset and the 12th subset. The data in each subset will be fixed between time 0 and will be fixed relatively to the end-observation time based on the subset index number. This means that subset 1 (smallest subset) is like making a component lifetime prediction as if it were 22 months ago, subset 2 is like making a prediction as if it were 20 months ago, ..., subset 12 is like making a prediction as if it were today.

A graphical example is given in Figure 20 with 4 subsets (instead of the 12 used). The upper timeline shows that there is a minimum time required before the maturity test can start in order to have all subsets filled with data. The lower timeline in the figure shows the same 4 subsets only now after a longer observation period. The data used for the estimated FphV of the first subset is much larger as the
Subsets are relatively fixed to the end-observation time, but always starts at time 0. The second subset contains the data of the first subset, including two months extra.

By using the subset index for the past predictions, the component maturity formula can be simplified:

\[ \text{Component maturity: } s(FphV_b) < \alpha, \quad \text{for all } b \]  \hspace{1cm} \text{Eq. 22}

where \( s \) is the standard deviation of the estimated FphV values over all subsets \( b \), \( FphV_b \) is still the estimated FphV of the fourth year (1461 days) for subset \( b \). The standard deviation of all estimated FphV values has to be smaller than the cut-off value \( \alpha \). The estimated FphV for subset 1 will always have the least amount of data, the estimated FphV for subset 2 will have the data of subset 1 plus some extra data (depending on the time between two predictions), the estimated FphV for subset \( B \) will have all the available data.

An acceptable \( \alpha \) has to be determined that works for each individual component that DAF has on a truck. If one component converges to an estimated FphV value of 5, the accepted \( \alpha \) of the estimation will be smaller than a component which estimated FphV converges to 5000. There will be a certain amount of noise in the prediction which is small enough to be acceptable, but this amount is dependent on the estimated FphV value. The component maturity formula can be expanded by standardizing the variance by dividing the standard deviation of the estimated FphV values over the mean \( \bar{X} \) of all the estimated FphV values. The result that comes from the standard deviation over the mean is called “coefficient of variation”. The coefficient of variation of the component is then compared to the \( \alpha \) in order to determine component maturity. This means that component maturity is still tested by looking at the variance of the component lifetime predictions in the past, but the standardization makes it possible to determine just one \( \alpha \) value that works for all components.
Component maturity: \( \frac{s(FphV_b)}{\overline{X}(FphV_b)} < \alpha, \quad \text{for all } b \)  

A hypothesis can be set for component lifetime maturity:

H0: Component lifetime predictions are not in the mature phase
H1: Component lifetime predictions are in the mature phase

\[
\text{Component maturity (reject H0): } \frac{s(FphV_b)}{\overline{X}(FphV_b)} < \alpha, \quad \text{for all } b
\]

where \( s \) is the standard deviation of the estimated FphV of the fourth year (1461 days) over all subsets \( b \), \( \overline{X} \) is the mean estimated FphV value of the fourth year (1461 days) over all subsets \( b \). If the result of the division, called coefficient of variation, is smaller than the cut-off value \( \alpha \) then H0 will be rejected, meaning that the component lifetime prediction is in the mature phase.

The methodology indicates when the predictions are stable, which gives a good indication of the reliability of the prediction. This methodology focuses on the predicted value over four years. Maturity for this period will not imply stability for an eight years period. If a maturity for eight years is required, then the test parameters should be re-evaluated.

To be able to use the methodology for DAF, the number of subsets have to be determined and the time between each subset. Then, an appropriate cut-off value \( \alpha \) has to be chosen to determine component lifetime maturity. This will be done in the following two subsections. Afterwards, the operational definition will be given again with the chosen parameters.

### 4.1.1 Determine the number of subsets and time between subsets

A choice has to be made on how long the measurement period should be, and how many predictions (subsets \( b \) in Eq. 24) should be done during this measurement period.

The measurement period has to be such that there is enough time to see the changes in stability, but not so much time that it takes forever to determine maturity. The measurement of component maturity in an empty system (for example when a new component is introduced) can only start when the observation time is longer than the measurement period, otherwise there will not be an equal number of subsets which makes determining a cut-off value difficult. By making the measurement period longer there is more time available for to measure the stability of the predictions at the start of the maturity determination, but it also means a longer waiting time before the maturity determination can begin. DAF currently waits for about 1,5 to 2 years before they feel confident enough to start making early changes to the costing of their repair & maintenance contracts for new trucks based on the data they have available at that point. The 1,5 to 2 years is based on the experience that DAF has in bringing new trucks or components to the market. In consultation with DAF, two years between the first (smallest) subset and the last (largest) subset seems to be a good measurement period such that the stability can
be observed properly. Further research is recommended on fine-tuning the measurement period, but from the experience of DAF it is expected that this is a good value.

The number of predictions is a choice between required computational power and the level of detail. The highest level of detail is gained by doing a prediction for each day since the start of the measurement. This also requires immense computational power. The dataset is very large and fitting a distribution is computationally relatively difficult. This is not a problem when making just one or a couple predictions, but with many observed defects it would be very unpractical to make a prediction for every new defect. Running the mathematical model for every component takes about three hours, which is acceptable. The maturity test, which requires more predictions to be done, will be added on top of these three hours. DAF prefers that the model can be run in one night outside the office hours, as stated in the scope of this project. Therefore, the model should run within 10 hours but preferably less.

A rough estimate on the required time can be given on how many predictions can be done without exceeding the 10 hours. It starts with the 3 hours that is required to run the model on the complete dataset. Measuring the time spent on each component prediction shows that 80% is spent on fitting predictions and 20% is spent on data management such as filtering the source data for the right trucks and claims of the requested component number. The data managing is only needed once per component, so the extra time for each added maturity prediction is only 80% of the 3 hours. On top of this, in the previous section it is explained that only Weibull is fitted on the data, the other three can be skipped. This means that fitting the distributions requires 25% of the original time. Finally, because the predictions are done in the past, there is less data. Fitting distributions with less observed trucks is faster. Predictions are done over the last two years, so the maximum number of observed trucks is the average of subset 7 and 8 (subsets from chapter 3, Table 5) for most of the components. The time required for the maturity test needs to be corrected for this smaller subset. The number is trucks observed in subset 7 is only 82% (see Table 5: \(\frac{32782 + 4068 + 5448}{39948 + 4701 + 6688} = 0.82\)) compared to the trucks observed in subset 8, so the time required to do all the predictions in subset 7 is 2.47 hours \((0.82 \times 3 = 2.47)\).

With this information, it is possible to determine the maximum number of predictions \(x\) that meets the 10 hour constraint:

\[
\begin{align*}
\text{maximize } x, \text{ subject to constraint:} \\
3 + \frac{2.47 + 3}{2} \times 0.8 \times 0.25 \times x &< 10 \\
\text{maximized } x & = 12.79
\end{align*}
\]

Eq. 25

The optimization shows that \(x = 12.79 \rightarrow 12\) is the highest number of predictions possible without exceeding 10 hours. This means that one prediction is made per two months, for up to two years in the past.

Concluding, 12 subsets (subsets b in Eq. 24) are created and spread evenly over two years. This means that subset 1 is like making a component lifetime prediction as if it were one year and ten months ago, and subset 2 is like making a prediction as if it were one year and eight months ago. These subsets
simulate how the predicted component lifetimes have been changing over the past two years. This means that maturity can only be determined if there is a minimum of two years data available.

![Diagram](image)

**Figure 21: Spread of the previous predictions for determining maturity**

### 4.1.2 Determining the cut-off value for maturity

There is no universal standard to know when the coefficient of variation is low enough to determine component lifetime maturity. In order to make one, an empirical optimization will be done using the large available database that DAF has as a sample for the entire population. In this optimization, every cut-off value will be tried until the cutoff value is found that results in the optimal performance of all the component lifetime predictions.

To do this, the component lifetime maturity test (Eq. 24) will be done for all components for all eight evaluation subsets as created in section 3.4.1. The coefficient of variation will be stored in the summary screen of the component (called maturity coefficient in the summary screen). An example summary screen is shown in Table 7.

With the coefficients of variation for every component over all eight subsets (subsets of section 3.4.1) available, the evaluation done for the performance of the predictions in section 3.4.2 can be mostly repeated. Though, this time only the estimated FphV values of mature components will be compared to the Kaplan-Meier estimated FphV values. The two formulas for the aggregate FphV values $AFphV_{i,j}$ (Eq. 12) and $AFphVKM_{i,j}$ (Eq. 14) have to be changed such that it is only for components k with a coefficient of variation smaller than $\alpha$ (H0 rejected). To avoid confusion with the equations from chapter 3, these will be called “Maturity aggregated FphV values”, or $MAFphV_{i,j}$ and $MAFphVKM_{i,j}$.

\[
MAFphV_{i,j} = \sum_k FphV_{i,j}, \quad \text{for all } k \text{ with } H0 \text{ rejected}; \text{ for } i = 1, \ldots, 8; \text{ for } j = 1, \ldots, 8 \quad \text{Eq. 26}
\]

\[
MAFphVKM_{i,j} = \sum_k FphVKM_{i,j}, \text{ for all } k \text{ with } H0 \text{ rejected}; \text{ for } i = 1, \ldots, 8; \text{ for } j = 8 \quad \text{Eq. 27}
\]

where i is the “FphV after i years” and j is the subset number (8 in total, as shown in Figure 15). The prediction accuracy is then calculated as explained in Eq. 15 only now with the input of Eq. 26 and Eq. 27.
\[
Prediction\ accuracy_{i,j,\alpha} = \frac{MAFphVKM_{i,j}}{MAFphV_{i,j}} - 1, \quad \text{for all } i,j
\]

Eq. 28

where \(\alpha\) is the maturity cut-off value that is chosen for rejecting H0. By changing the maturity cut-off value \(\alpha\), different prediction accuracies will be calculated. The first cut-off value \(\alpha\) is set at 0.01. This would remove all components \(k\) with a coefficient higher than 0.01 from the evaluation, which would likely be every component. The next cut-off would be 0.02, then 0.03, repeat until cut-off value of 1 is reached, which would likely include all components \(k\) in the evaluation.

Choosing 100 cut-off values means that there will be 100 tables with the prediction accuracy results (Eq. 15), just like the one shown in Table 11. The goal is to choose a cut-off value for the coefficient of variation which results in the prediction accuracy values closest to zero in the prediction performance table. To get a comparable measure between each of the prediction performance tables, a graph can be made. The largest uncertainties are found in the earlier subsets as these have the least amount of data.

The mean of the prediction accuracy of the first three “FphV after i years” and the first four subsets \(j\) values are taken for every prediction accuracy result as the first four years after the launch of a new component are the most important years for DAF to decide maturity on, as DAF would like to be able to determine maturity before 4 or 5 years.

\[
\text{mean prediction accuracy} = \bar{X}(\text{Prediction accuracy}_{i,j,\alpha}), \text{for all } \alpha, \text{for } i,1,\ldots,3, \text{for } j,1,\ldots,4
\]

Eq. 29

where \(\bar{X}\) is the mean of the prediction accuracies for all 100 \(\alpha\) cut-offs, for the first three “FphV after i years” columns and the first four subset rows \(j\). A graph can be shown that shows the mean prediction accuracy as a function of the chosen cut-off value.

Figure 22: Performance of the predictions for different maturity cut-off values, lower is better
As was proposed, the prediction performance increases (goes closer to 0) as the cut-off value becomes more strictly. This is good, as it means that the proposed method to indicate component maturity works as intended. Still, the more strictly the cut-off value, the less components will be classified as mature. If the cut-off value is too strict, components would never become mature which would make definition useless for DAF. Figure 23 shows the number of components that are mature for each cut-off value in subset two, three and four (left, middle and right graph respectively) (subsets from chapter 3), subset one is not shown as it had no mature components (maturity measure needs two years data).

Figure 23: Number of components that have H0 rejected per cut-off value

The number of components in subset three and four go up nicely, while the second subset lags. This makes the prediction accuracy uncertain at the start. Comparing the prediction accuracy graph (Figure 22) and the component quantity graphs, the coefficient of variation cut-off value should be less than 0.5 as otherwise the cut-off is not strict enough to be usable for DAF. In the same reasoning, it should be at least higher than 0.1 otherwise the cut-off value is too strict. Having a coefficient of variation cut-off value of 0.3 seems to show good performance as seen in Figure 22 while also having an acceptable number of components in the comparison to make the measure useful for DAF. Therefore, H0 will be rejected when the coefficient of variation (\(\alpha\) in Eq. 24) is below 0.3.

4.1.3 Operational definition for component lifetime maturity

With all the parameters chosen, the operational definition for component lifetime maturity is as follows.

H0: Component lifetime predictions are not in the mature phase
H1: Component lifetime predictions are in the mature phase

The formula to test the hypothesis:

\[
\text{Component maturity (reject } H_0) = \frac{s(F_{\text{ph}V_b})}{\bar{X}(F_{\text{ph}V_b})} < 0.3, \quad \text{for } b = 1, \ldots, 12
\]

Eq. 30

where \(s\) is the standard deviation of the estimated \(F_{\text{ph}V}\) of the fourth year (1461 days) over all 12 subsets \(b\), \(\bar{X}\) is the mean estimated \(F_{\text{ph}V}\) value of the fourth year (1461 days) over all 12 subsets \(b\). If the result of the division, called coefficient of variation, is smaller than the cut-off value \(\alpha\) then \(H_0\) will be rejected, meaning that the component lifetime prediction is in the mature phase. The 12 subsets \(b\) are subsets of the complete dataset, starting from the start of the observation and ending relatively fixed to the end of the observation. The subsets are spread equally over 2 years with the smallest subset ending.
22 months before the end of the observation, and the largest subsets ends at the end of the observation (= whole dataset).

4.1.4 Procedure flowchart for determining component lifetime maturity
A flowchart is shown in Figure 24 that summarizes every step in the determination of the component lifetime maturity. The component lifetime maturity determination will happen every time a component lifetime prediction is made. The result of the maturity test is given as extra information on the prediction of a component lifetime. This means that the predicted FphV values are still available while the component is not mature, but it will be advised that the estimated FphV values should not be used at face-value.

![Flowchart for determining component lifetime maturity](image)

4.2 Numerical example for determining component lifetime maturity
The previous section discussed the decisions on how component lifetime maturity should be evaluated. This section explains the method how to determine maturity with a concrete numerical example. The steps that are taken are shown in Figure 24. To explain every step of the component lifetime determination with a numerical example, component 63402 – “ignition for the cab heater” is used for the DAF XF FT trucks. The values used in this numerical example are changed due to confidentiality reasons.

**Create 12 subsets (index b)**
Twelve subsets (with index b) are created each with cumulatively more data, as shown in Figure 21. The first subset has the data between the start of the observation until 22 months before the end of the observation. The second subset has the data between the start of the observation until 20 months before the end of the observation. Each new subset has 2 months more data, the 12th subset has all the available observed data. In total, the end-points of the 12 subsets are spread evenly between two years before the end of the observation period and the end of the observation period, and all subsets start at the start of the observation period.

The number of observed trucks and observed defects for each subset b are shown in Table 15. The number of observed defects per subset b is fictitious due to confidentiality reasons. The main difference between these subsets is that the earlier subsets have fewer trucks and defects observed.
Table 15: Subsets for component 63402 used as data for determining component lifetime maturity

<table>
<thead>
<tr>
<th>Subset (b)</th>
<th>Subset 1 (smallest)</th>
<th>Subset 2</th>
<th>Subset 3</th>
<th>Subset 4</th>
<th>Subset 5</th>
<th>Subset 6</th>
<th>Subset 7</th>
<th>Subset 8</th>
<th>Subset 9</th>
<th>Subset 10</th>
<th>Subset 11</th>
<th>Subset 12 (largest)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of observed defects</td>
<td>238</td>
<td>244</td>
<td>246</td>
<td>247</td>
<td>253</td>
<td>260</td>
<td>269</td>
<td>275</td>
<td>276</td>
<td>278</td>
<td>283</td>
<td>293</td>
</tr>
</tbody>
</table>

Estimate Weibull parameter values
A Weibull distribution will be fitted on each of the subsets b, the same way as it has been explained in the previous chapter, using the maximum likelihood estimation (Eq. 4). After fitting the distribution, there will be a shape and scale parameter, called beta $\beta_b$ and eta $\eta_b$ respectively for all subsets b (1,..,12). The Weibull parameter values of the subsets are as follows:

Table 16: Predicted shape and scale parameters of the Weibull distribution after fitting it on the subsets

<table>
<thead>
<tr>
<th>Subset (b)</th>
<th>Subset 1 (smallest)</th>
<th>Subset 2</th>
<th>Subset 3</th>
<th>Subset 4</th>
<th>Subset 5</th>
<th>Subset 6</th>
<th>Subset 7</th>
<th>Subset 8</th>
<th>Subset 9</th>
<th>Subset 10</th>
<th>Subset 11</th>
<th>Subset 12 (largest)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Shape ($\beta$)</td>
<td>1.95</td>
<td>1.93</td>
<td>1.92</td>
<td>1.91</td>
<td>1.92</td>
<td>1.93</td>
<td>1.95</td>
<td>1.95</td>
<td>1.95</td>
<td>1.95</td>
<td>1.95</td>
<td>1.92</td>
</tr>
<tr>
<td>Scale ($\eta$)</td>
<td>27.019</td>
<td>28.566</td>
<td>30.305</td>
<td>32.297</td>
<td>31.072</td>
<td>29.677</td>
<td>27.957</td>
<td>28.050</td>
<td>28.919</td>
<td>28.836</td>
<td>29.191</td>
<td>32.244</td>
</tr>
</tbody>
</table>

Calculate failures per hundred vehicles
Calculating the estimated failures per hundred vehicles (FphV) for the first four years for the Weibull distribution is done by calculating the Weibull cumulative distribution function value at time $t = 1461$ (=4 years in days) and multiplying it by 100 (see Eq. 19). The FphV after four years is taken as this is the most sold length for repair & maintenance contracts. The FphV will be calculated for each subset b. The results are shown in Table 17.

\[
FphV_b(1461) = \left(1 - e^{-\left(\frac{1461}{\eta_b}\right)^{\beta_b}}\right) \times 100, \quad \text{for all } b
\]

Table 17: Predicted failures per hundred Vehicles (FphV) for the first four years for each subset b

<table>
<thead>
<tr>
<th>Subset (b)</th>
<th>Subset 1 (smallest)</th>
<th>Subset 2</th>
<th>Subset 3</th>
<th>Subset 4</th>
<th>Subset 5</th>
<th>Subset 6</th>
<th>Subset 7</th>
<th>Subset 8</th>
<th>Subset 9</th>
<th>Subset 10</th>
<th>Subset 11</th>
<th>Subset 12 (largest)</th>
</tr>
</thead>
<tbody>
<tr>
<td>FphV</td>
<td>0.34</td>
<td>0.32</td>
<td>0.30</td>
<td>0.27</td>
<td>0.28</td>
<td>0.30</td>
<td>0.32</td>
<td>0.32</td>
<td>0.30</td>
<td>0.29</td>
<td>0.29</td>
<td>0.27</td>
</tr>
</tbody>
</table>
Calculate component maturity
With the FphV values, calculate the coefficient of variation of the whole series to get the indicator of stability. For the example values shown in Table 17 the mean $\bar{X}$ over all subsets $b$ is 0,3 and the standard deviation over all subsets $b$ is 0,022. The coefficient of variation is then:

$$\text{Coefficient of variation: } \frac{s(FphV_b)}{\bar{X}(FphV_b)} \text{ for all } b = \frac{0,022}{0,3} = 0,075$$

Eq. 32

This coefficient of variation is compared to the cut-off value $\alpha$ to determine whether the component lifetime prediction is mature. The cut-off value has been calculated to be 0,3.

$$\text{Component maturity (reject H0) } = 0,075 < 0,3$$

Eq. 33

Because 0,075 is smaller than 0,3, H0 will be rejected. The lifetime prediction for this component is in the mature phase. The coefficient of variation value (0,075) will be shown in the summary screen (Table 7).

4.3 Evaluating the performance of the predictions for mature components
To evaluate the performance of the predictions for mature components, the same setup from section 3.4.1 is used. However, instead of comparing every component against the Kaplan-Meier empirical distribution, only the components that are mature will be compared to the Kaplan-Meier. Mature components are these components that have the null hypothesis “Component lifetime predictions are not in the mature phase” rejected. As explained earlier, because of the limitations of the data, it does not guarantee that the predictions remain stable as more data is added. It should however remove any components which predictions have not been stable in the past from the comparison, which should lead to better overall results. The prediction accuracies are shown in Table 18.

Table 18: Prediction accuracies (Eq. 15) for all i,j for only components k that have H0 rejected (Eq. 24) (left) and the number of components compared (right)
The overall model accuracy improves slightly compared to the prediction accuracies of all components shown in Table 11. Looking more closely at the different subsets and FphV values by making a histogram of the prediction errors, using Eq. 16, there are some interesting differences to be seen. The histogram of the prediction errors are shown in Figure 25. The left histogram is for all components k, the right histogram is for components k that have H0 rejected (are mature).

![Histogram of the differences per component for subset 3, FphV at year 2](image1)

![Histogram of the differences per component for subset 3, FphV at year 2](image2)

*Figure 25: Histograms of the prediction errors per component for subset 3, FphV after 2 years. Left histogram shows every component, right histogram shows only components with H0 rejected*

The mathematical model has been successful in recognizing and removing the outlier that was seen earlier in the comparison of all components, which has a very positive effect on the standard deviation of the prediction errors (see Table 19). The method is not always successful in finding the outliers, as can be seen in Figure 26.

![Histogram of the differences per component for subset 5, FphV at year 1](image3)

![Histogram of the differences per component for subset 5, FphV at year 1](image4)

*Figure 26: Histograms of the prediction errors per component for subset 5, FphV after 1 year. Left histogram shows every component, right histogram shows only components with H0 rejected*

It can be seen that the outlier on the left side is not removed. This means that the predictions have been stable up until the fifth subset and something changed to component lifetime afterwards that changed the component lifetime distribution, which led to the prediction error between the estimated FphV in subset 5 and the Kaplan-Meier FphV value in subset 8. It is hard to draw any conclusions why this happens because there are many unknown factors that could influence this. The standard deviations of all the FphV error are shown in Table 19.
The standard deviation of the prediction errors are significantly improved compared to the original standard deviations in Table 12, especially in the earlier subsets. It does not recognize all the outliers, but it is still a positive improvement over the original model. The maturity coefficient that is calculated in this chapter has an added benefit of being interpretable by users. From Figure 22 it was shown that using a lower coefficient values resulted in better prediction accuracy. From Table 19 it can be seen that standard deviation of the prediction errors also becomes smaller for mature components. Right now the model rejects H0 when the coefficient is lower than 0.3, but it is known that the performance is better with an ever lower coefficient and worse with a higher coefficient. When evaluating components individually, looking at the actual coefficient value will give a better indication than just seeing whether H0 is rejected or not, which in turn adds value for DAF.
5 Software implementation of the mathematical model (R code)

The mathematical model has been constructed to be completely automatic. It retrieves the data via a SQL query from the database and then calculates everything using the statistical software program R-Studio (R-Studio, 2016). The choice for using R-Studio was compared with other popular statistical programs such as SPSS, SAS and Minitab. The comparison was based on accessibility and statistical content. Minitab seemed to have less functionality for the statistical methods used in this research compared to the other three. The statistical content for R, SAS and SPSS all seemed to be sufficient. R-Studio as being open-source with a large community does seem to have the most possibilities for doing the required statistical methods and being able to present them nicely. For the accessibility, R-Studio easily outperforms all others as it is the only program in this comparison that is completely free to use. Being open-source, it is also very easy to look at, configure, and validate any methods that are in the program. This accessibility also allows everyone that is interested in this mathematical model to access it freely.

The mathematical model can be run and re-run with different datasets containing different truck types, different components, or newer data. The model provides and stores for each component all the results of the statistical tests, plots, and choices made. The fact that it runs without human intervention makes it a valuable asset for DAF. It could be used to predict the component lifetime monthly, enabling DAF to always have the latest predictions available. It could also be used to see clearly whether a component improvement has a positive effect on the component lifetime. How the program works exactly will be explained in this chapter. The explanation of the code is useful for DAF to offer complete transparency on the processes, and enables them to do modifications easier. Also, other parties within DAF or outside DAF might also be interested in this tool, or parts of this tool (for example the newly developed maturity tests) for use in this own business. Figure 27 shows the flowchart of the global processes that are executed within the mathematical model.

![Flowchart of the global processes within mathematical model](image-url)
It starts by loading the libraries used by the program, and reading the configuration settings set by the user. This is currently done via variables inside the code, but this could easily be extended to a user interface input screen if DAF decides to implement it in their reporting server. Then it loads the two data files that have been created via the SQL query, as explained in section 2.2. After it has the two large data files, the model enters a loop. In this loop, each component is handled one by one. This is where all the statistical tests and predictions will happen that have been discussed in this thesis. First it creates a smaller data frame that has only the data from the components and trucks that are currently being processed to increase the performance of the program. Then, it looks in this data whether there was a recorded improvement. If so, it executes the Kaplan-Meier comparison as explained in section 3.3. If the old and new data is comparable, it does nothing. If it is not comparable, it will remove the trucks and claims that are not linked to the new component. The model notes the improvement date and the comparison results in the component summary data. After this step, some very basic information such as the number of trucks observed, defects observed, mean claim cost, etc. is stored in the component summary. The phenomenon for having many observed defects at the end of the contract follows afterwards. It does the actions as described in “APPENDIX C: Many claims at end of warranty or repair & maintenance contract”. The results are also stored in the component summary. The model continues with calculating the component maturity coefficient, as described in section 4.2. The coefficient of variation value is stored together with the note whether the hypothesis was accepted or rejected. Then, it fits the four chosen distributions and compares them as explained in section 3.2. The best distribution is chosen and its parameters are used to calculate the component lifetime. The results are stored in the component summary data. All the processes within the loop are repeated for every component number that is in the large data file. In the end, there will be a long list of component summaries that can be used for various purposes.

The actual program R code is 350 lines long and contains 23,974 characters which are fully shown in “APPENDIX G: Mathematical model R code”. The different parts of the model are indicated with comments in the R code.
6  Research conclusions and recommendations

In this chapter the overall research conclusion is shown in section 6.1. Recommendations for DAF are given in section 6.2. The academic relevance of this research is described in section 6.3. Finally, any interesting follow-up researched are discussed in section 6.4.

6.1  Research conclusions

This master thesis had one main assignment divided into three deliverables. Each deliverable had several research questions listed in section 1.2.2. The main assignment:

Construct a mathematical model that predicts component lifetime and gives an indication of component lifetime maturity.

Throughout the report, the research questions are answered. The main assignment has been successfully completed. The mathematical model is able to predict the component lifetime and a method has been developed that gives an indication of component lifetime maturity. This section will describe all important findings and conclusions that were made throughout the report.

An empirical evaluation of the mathematical model has been made. The component lifetime predictions are overestimating the number of failures per hundred vehicles compared to the empirical distribution function. It is difficult to draw conclusions why this happens. The overestimation seems to be equal for most of the individual component predictions. The prediction errors are normally distributed with a small spread, so the overestimation is there but it is very constant over all individual component lifetime predictions. When taking in to account this overestimation, the component lifetime predictions are useful for DAF to determine the cost price of the repair & maintenance contracts.

DAF wanted to have an indication of component lifetime maturity for the component lifetime predictions. A method has been developed that gives this maturity indication for every component lifetime prediction. The component lifetime maturity method is based on the coefficient of variation of a series of 12 estimated failures per hundred vehicles (of the fourth year) over the past two years. If the maturity coefficient is below 0.3 then the component lifetime prediction is in the mature phase. Via empirical evaluation it was shown that component lifetime predictions with a lower coefficient have a higher chance of remaining stable, this is very useful for DAF to know how much it can depend on the estimated failure per hundred vehicles.

The mathematical model has been programmed in R-Studio and a SQL query has been written to extract data from the server. Having the mathematical model operational as an independent program allows for easy implementation and complete evaluation of the mathematical model. The programming code is available in this report such that parties within or outside DAF can easily read and use it.

Two fitting techniques have been compared: the Maximum Likelihood Estimation (MLE) and the least squares or rank regression (RR) method. DAF was not sure whether MLE or RR was better in their situation. Via literature research and a case example it was shown that the MLE outperforms RR when censoring is not evenly distributed among observed defects. In this research, the majority of the censored data is at exactly 1096 days (three years) when the truck warranty ends, meaning that it is not
evenly distributed. Therefore, the MLE method is superior to the RR method for fitting the distributions on the data. The MLE method is used in the mathematical model to estimate the distribution parameters.

In the current situation, whenever there is a component improvement, the old data always gets thrown away. Because a component improvement is not necessarily a change in component lifetime, it could lead to throwing away useful information. The mathematical model has been expanded to handle component improvements. The model evaluates whether the data of the old component and the new component are the same by comparing the Kaplan-Meier empirical distribution functions via the log-rank test (Eq. 8). If they show similar failure behavior, the old data is used together with the new data to predict component lifetimes. If the failure behavior is not similar, the component lifetime is changed and the old data is not used for new predictions. This method will prevent information loss for component improvements that did not change the component lifetime.

6.2 Recommendations
Throughout the research, several aspects have been found which can be improved by DAF. This section gives recommendations for the found aspects.

With the rollout of the new claim handling system called Dealer Claim Entry (DCE), there will be more possibilities regarding predicting component lifetimes. Currently it is very hard to get failure data beyond an observation period of three years because the longer period repair & maintenance defect claims do not have an associated component number. With DCE, this component actually is stored. The first DCE rollout was three years ago in Poland, and Germany followed soon after. This means that in a few years there will be a lot of data available for trucks that have an observation period that is longer than three years. DAF should change the current prediction methods to include this data. It is also recommended to re-run the evaluation of the mathematical model to investigate whether choosing different distributions to predict the component lifetimes is helping in making better predictions.

Each truck has a different configuration of installed components, which leads to that not all components are installed on each truck. It is recommended to include this information in the analysis. However, at this moment it is not possible to extract this information.

Predictions are done for component numbers. There are different configurations possible within the same component number, because the underlying part numbers are not always exactly the same. The underlying parts could be of a different brand or a different version number. Different underlying parts could have an influence on the overall lifetime of the component number. It is recommended to include this information in the analysis as well. However, at this moment it is not possible to extract this information.

6.3 Academic relevance
There is very little literature available on determining the maturity of a prediction, specifically with an application to trucks. In this research a new method is presented which tests whether a component lifetime prediction will be stable in the future. The presented method has been empirically tested using data from DAF and the results have been evaluated. The method has been empirically evaluated and
found working as intended in indicating the stability of component lifetime predictions in the future. The newly developed component lifetime maturity method and the evaluation of it contribute to any research projects in the same field.

The empirical evaluation of the mathematical model as a whole, combined with the large and heavily censored dataset is quite unique and will contribute to research done in the same field.

Three prediction methods have been compared in a literature study on the requirements for being able to use them and for which purpose they are most effective. Between lifetime distributions, stochastic processes and artificial neural networks, lifetime distributions have been chosen as the best method to predict component lifetimes as it offered the most flexibility, is not stuck to certain unrealistic assumptions and has less chance of overfitting. The overall results of the literature study are shown in Table 20. The literature study performed for choosing the best prediction method provides colleague researchers with a good source of information for when they are choosing the best method to do predictions with.

Table 20: Advantages and disadvantages of the different prediction methods

<table>
<thead>
<tr>
<th>Prediction technique</th>
<th>Advantages</th>
<th>Disadvantages</th>
</tr>
</thead>
<tbody>
<tr>
<td>Lifetime distributions</td>
<td>Strong theoretical foundation</td>
<td>Assumption that data represents the exact number of failures might not hold</td>
</tr>
<tr>
<td></td>
<td>Many different statistical distributions available</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Method is known to DAF</td>
<td></td>
</tr>
<tr>
<td>Stochastic processes</td>
<td>Strong theoretical foundation</td>
<td>Information loss due to use of failure rates</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Power law assumption might not hold</td>
</tr>
<tr>
<td>Artificial neural networks</td>
<td>Finds complex relationships in large databases</td>
<td>Only exploratory uses</td>
</tr>
<tr>
<td></td>
<td>Require less formal statistical training</td>
<td>No theoretical foundation</td>
</tr>
<tr>
<td></td>
<td>Good for exploratory analysis in large databases</td>
<td>Black box makes training and troubleshooting difficult</td>
</tr>
<tr>
<td></td>
<td></td>
<td>No confidence intervals on results</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Prone to overfitting</td>
</tr>
</tbody>
</table>

6.4 Further research

The mathematical model developed in this research is consistent in how it predicts the component lifetimes. The prediction errors are all very close together and nicely normally distributed, but the predicted failures per hundred vehicles are consistently higher than the real failures per hundred vehicles. In order to improve the prediction accuracy of the mathematical model, it would be interesting to find the source of the underestimation of the component lifetimes in order to correct the bias.

The chosen number of subsets and the time between each subset in the component lifetime maturity test is based on the experience that DAF has with predicting component lifetimes. DAF expects that the chosen values are good, but would like to see whether improvements are possible by doing a sensitivity
analysis for what happens when different parameters are used in the component lifetime maturity method.

During the data analysis, a phenomenon has been found which could be interesting to further investigate for DAF. For 83 component numbers it has been detected that they have an abnormally high number of observed defects just before the end of their repair & maintenance contract. The component numbers are listed in “APPENDIX C: Many claims at end of warranty or repair & maintenance contract” together with the method used to detect them. The effect did not seem to have any negative effect on the performance of the predictions, but it is occurring structurally for these specific components. When warranty items are replaced in a structural pattern, it could be better to switch the category of the component from a warranty component to a maintenance component.
7 Bibliography


APPENDIX A: SQL query for retrieving the data

There are two separate queries, one for the claims and the other for the trucks. They should be executed at the same time when retrieving the data. For overview purposes, the two queries will be showed separately in this appendix. Figure 28 shows the query graphically.

Figure 28: Graphical representation of the query shown in this appendix

SQL query for the claim data

This section shows the actual query that is executed on the server. It returns the claims data that the mathematical model uses as input.

```
DECLARE @endyear as smalldatetime, @ukendyear as nvarchar(15)
SET @endyear = '2015-12-31 00:00:00'
SET @ukendyear = '20151231'

select parc.*, impr, clm, dwh
from

impr (improvements)
clm (claims)
dwh (data warehouse)

parc (truck data)
euchs (EU trucks)
ukchs (UK trucks)

impr.improveddate, impr.improvedchassis, CONCAT(impr.Description1, impr.Description2, impr.Description3, impr.Description4, impr.Description5) as improveddescription, clm.defectcode, clm.defectcodescription, clm.defectdate, DATEDIFF(MONTH, deliverydatefinal, defectdate) as monthstodefect, DATEDIFF(DAY, deliverydatefinal, defectdate) as daysstodefect, dekkingsperiodefinal - DATEDIFF(DAY, deliverydatefinal, defectdate) as daystocompend, clm.parcosts, clm.labourcosts

when clm.misccosts is NULL then 0 else clm.misccosts end as misccosts
, clm.deliverycost, CAST(DATEDIFF(DAY, deliverydatefinal, defectdate) as float) / CAST(dekkingsperiodefinal as float) as perdiffcontractend

from

( select

  case when euchs.ChassisNr is null then ukchs.ChassisNr else euchs.ChassisNr end as chassisnr,
  case when euchs.ChassisNr is null then ukuchs.ChassisModel else euchs.ChassisModel end as chassismodel,
  case when euchs.ChassisNr is null then RTRIM(ukuchs.ProductRange) else RTRIM(euchs.ProductRange) end as productrange,
  case when euchs.ChassisNr is null then ukuchs.serie else euchs.serie end as serie,
  case when euchs.ChassisNr is null then ukuchs.classificationID else euchs.classificationID end as classificationID

  from dbo.Stefan_ukclaims
  left OUTER JOIN dbo.Stefan_euchs ON euchs.ChassisNr = ukuchs.ChassisNr

  --> select case when @CMChassisNo is null then f.@CMChassisNo else @CMChassisNo end as chassisnr,
  --> case when euchs.chassisnr is not null and ukuchs.chassisnr is not null then
  --> case when euchs.dekkingsperiode >= ukchs.contractlength then euchs.dekkingsperiode else ukchs.contractlength end as dekkingsperiodefinal
  --> case when euchs.chassisnr is not null and ukuchs.chassisnr is not null then
  --> case when euchs.dekkingsperiode >= ukchs.contractlength then euchs.dekkingsperiode else ukchs.contractlength end as dekkingsperiodefinal
  --> case when euchs.chassisnr is not null and ukuchs.chassisnr is not null then
  --> case when euchs.dekkingsperiode >= ukchs.contractlength then euchs.dekkingsperiode else ukchs.contractlength end as dekkingsperiodefinal
  --> case when euchs.chassisnr is not null and ukuchs.chassisnr is not null then
  --> case when euchs.dekkingsperiode >= ukchs.contractlength then euchs.dekkingsperiode else ukchs.contractlength end as dekkingsperiodefinal
  --> case when euchs.chassisnr is not null and ukuchs.chassisnr is not null then
  --> case when euchs.dekkingsperiode >= ukchs.contractlength then euchs.dekkingsperiode else ukchs.contractlength end as dekkingsperiodefinal
  --> case when euchs.chassisnr is not null and ukuchs.chassisnr is not null then
  --> case when euchs.dekkingsperiode >= ukchs.contractlength then euchs.dekkingsperiode else ukchs.contractlength end as dekkingsperiodefinal
  --> case when euchs.chassisnr is not null and ukuchs.chassisnr is not null then
  --> case when euchs.dekkingsperiode >= ukchs.contractlength then euchs.dekkingsperiode else ukchs.contractlength end as dekkingsperiodefinal

  from
```
(  
  -- voortuig parc eu (alleen extended warranty)
  SELECT distinct
  chs.ChassisNr, chs.ChassisModel, dtype.classificationID, ProductRange, left(productrange, 2) as serie, DeliveryDate, DeliveryDate as startdate, datediff(month, 36, DeliveryDate) as enddate
  , case when datediff(month, 36, DeliveryDate) < @endyear then 1096 else datediff(day, DeliveryDate, @endyear) end
  as dekkingsperiode
  ,DeliveryCountry
  from fact_Warr_ClaimHeader as clmheader
  INNER JOIN
dbo.dim_Warr_Chassis as chs
  ON clmheader.ChassisID = chs.ChassisID
  INNER JOIN
dbo.dim_Warr_Type as dtype
  ON clmheader.TypeID = dtype.TypeID
  INNER JOIN
dbo.dim_Warr_productrange as prodrange
  ON dtype.productrangeID = prodrange.productrangeID
  INNER JOIN
DM_DAFEHVMSQL1.dbo.dim_Warr_ServiceContract AS servcon
  ON chs.ChassisID = servcon.ChassisID
  INNER JOIN
dim_Warr_ClaimOrigin as clmori
  on servcon.ClaimOriginID = clmori.ClaimOriginID
  WHERE
    (LEFT(chs.ChassisNr, 2) = '0E' OR LEFT(chs.ChassisNr, 2) = '0G' OR LEFT(chs.ChassisNr, 2) = '0L')
  and DeliveryDate >= '2008-01-01 00:00:00'
  and DeliveryDate < @endyear
  and DeliveryCountry
  in('004', '001', '005', '048', '481', '079', '093', '095', '097', '098', '007', '001', '047', '088', '028', '039', '037', '029', '027', '018', '009')
  and left(chs.ChassisModel, 3) in('FA ', 'FT ', 'FE ')
  and (LEFT(dbo.dim_Warr_Chassis, 3) = '5133', '5647', '5896', '5897', '5898', '5899', '5908', '5901', '5988')
) euchs
  full outer join
  ()
  -- voortuig parc uk
  SELECT distinct(concat('0 ', [Chassis])) as chassisnr,min([ChassisModel]) as chassismodel,min([classificationID])
  as classificationID,min(productrange) as productrange, left(min(productrange), 2) as serie,min([Delivery Date])
  as deliverydate,cast(min([Contract Start]) as date) as start,cast(max([Contract End]) as date) as enddate
  , case when max([Contract End]) < @endyear then datediff(day, min([Contract Start]), max([Contract End])) else
  datediff(day, min([Contract Start]), @endyear) end
  as contractlength
  FROM
  [dbo].[test_stefan_ukclaims] as ukclm
  inner join
  ()
  -- uk parc
  SELECT concat('0 ', ukclm.Chassis) as chassisnr = dwh.Chassisnr
  where [Contract package] like '%FULL' and [Delivery Date] >= '20080101'
  and [Delivery Date] < @uke
  and uk parc is already pre-filtered to show only 2x4 trucks and only euro 5 and 6.
  group by [Chassis]
) uckhs on uckhs.chassisnr = euchs.ChassisNr
)
LEFT OUTER JOIN
()  
--uk claims
SELECT concat('0 ', [Chassis]) as chassisnr,min([ChassisModel]) as ChassisModel, min(productrange) as productrange, [Component Code]
  as defectcode, [Component Code] + ' ' + min([ComponentSubGroupDescr]) as defectcodedescription, min(cast([Defect Date] as date)) as
defectdate, max([Approved Claim - Parts Dlr]) as partscoasts, max([Approved Claim - Misc Dlr]) as misccosts, cast(0 as nvarchar) as handlingcost
  from
    fact_Warr_ClaimHeader as clmheader
  INNER JOIN

```
    dbo.dim_Warr_Chassis as chs
ON cleheader.ChassisID = chs.ChassisID
INNER JOIN
    dbo.dim_Warr_Type as dtype
ON cleheader.TypeID = dtype.TypeID
INNER JOIN
    dbo.dim_Warr_productrange as prodrange
ON dtype.productrangeID = prodrange.productrangeID
)
    dw
    inner join
        [dbo].[test_stefan_ukclaims] as ukclm
on concat('0', ukclm.[Chassis]) = dwh.ChassisNr
    inner join
        [dbo].[dim_Warr_ComponentGroup] as cmpgrp
ON [Component Code] = cmpgrp.ComponentSubGroupNr

WHERE [Line Number] = '1'
    and TRY_CONVERT(int([Component Code])) is not null
    -- and [Component Code] like '[0-9]'
    AND ukclm.DefectDate < @endyear
    group by concat('0', [Chassis]), [component code]
    union
    -- eu claims
SELECT claim.ChassisNr, min(chs.ChassisModel) as ChassisModel,
    min(prodrange.ProductRange) as ProductRange,
left(DefectCode,5) as defectcode,
left(DefectCode,5) + '0' + min(ComponentSubGroupDescr)
    as defectcodedescription,
min(DefectDate) as defectdate,
max([ArticlePaidDN_DTNV]) as partscosts,
max([LabourPaidDN_DTNV]) as labourscosts,
max([MiscCostsDN_DTNV]) as misccosts,
max(HandlingCost) as handlingcost
FROM fact_Warr_ClaimHeader as cleheader
    INNER JOIN
        [dbo].[dim_Warr_Claim] as claim
ON cleheader.ClaimID = claim.ClaimID
    INNER JOIN
    dbo.dim_Warr_Chassis as chs
ON cleheader.ChassisID = chs.ChassisID
    INNER JOIN
    dbo.dim_Warr_Type as dtype
ON cleheader.TypeID = dtype.TypeID
    INNER JOIN
    dbo.dim_Warr_productrange as prodrange
ON dtype.productrangeID = prodrange.productrangeID
    INNER JOIN
    dbo.dim_Warr_ComponentGroup as cmpgrp
ON cmpgrp.ComponentGroupID = cleheader.ComponentGroupID

WHERE claim.ClaimSort = 'N'
    AND claim.DefectDate < @endyear
    AND LastClaimStatus = '9'
    AND chs.DeliveryCountry = '8016'
GROUP BY claim.ChassisNr, left(DefectCode,5)

) cc
on parc.chassisnr = cc.ChassisNr
left outer join
(
    select * from
    (SELECT *, ROW_NUMBER() over (partition by ProductRange, componentSubGroup, ImprovedDate order by ImprovedDate) as rn
    FROM
    (Select * from [DM_DAFEHVMSQL1].[dbo].[dim_Warr_Improvements] where ImprovedDate < @endyear
    ) acde
    ) cc where rn < 2
)
impr
ON parc.productrange = impr.ProductRange AND cc defectcode = impr.ComponentSubGroup AND parc.classificationID = impr.ClassificationID

WHERE DATEDIFF(DAY, deliverydatefinal, defectdate) >= 0
AND dokingsperiodfinal - DATEDIFF(DAY, deliverydatefinal, defectdate) >= 0
```
SQL query for the truck data

This section shows the actual query that is executed on the server. It returns the truck data that the mathematical model uses as input.

```
select *
  ,case when RIGHT(parcl.productrange, 1) = 5 then 'EURO 5' else 'EURO 6' end as productrange
from
(
  select
    case when euchs.ChassisNr is null then ukchs.ChassisNr else euchs.ChassisNr end as chassisNr
    ,case when euchs.ChassisNr is null then ukchs.ChassisModel else euchs.ChassisModel end as chassismodel
    ,case when euchs.ChassisNr is null then RTRIM(ukchs.ProductRange) else RTRIM(euchs.ProductRange) end as productrange
    ,case when euchs.ChassisNr is null then ukchs.serie else euchs.serie end as serie
    ,case when euchs.ChassisNr is null then ukchs.classificationID else euchs.ClassificationID end as classificationID

    --select case when e.CMChassisis is null then f.CMChassiselse e.CMChassis is end as chassisNr
    ,case when euchs.chassisnr is not null and ukchs.chassisnr is not null then
case when euchs.dekkperiodeer = ukchs.dekkperiodeer end else
case when euchs.chassisnr is not null and ukchs.chassisnr is not null then
case when euchs.dekkperiodeer = ukchs.dekkperiodeer else ukchs.dekkperiodeer end
when euchs.chassisnr is not null and ukchs.chassisnr is not null then
case when euchs.dekkperiodeer = ukchs.dekkperiodeer then ukchs.dekkperiodeer else ukchs.dekkperiodeer end
when euchs.chassisnr is not null and ukchs.chassisnr is not null then
case when euchs.dekkperiodeer = ukchs.dekkperiodeer then ukchs.dekkperiodeer else ukchs.dekkperiodeer end
when euchs.chassisnr is not null and ukchs.chassisnr is not null then
case when euchs.dekkperiodeer = ukchs.dekkperiodeer then ukchs.dekkperiodeer else ukchs.dekkperiodeer end
when euchs.chassisnr is not null and ukchs.chassisnr is not null then
case when euchs.dekkperiodeer = ukchs.dekkperiodeer then ukchs.dekkperiodeer else ukchs.dekkperiodeer end
when euchs.chassisnr is not null and ukchs.chassisnr is not null then
case when euchs.dekkperiodeer = ukchs.dekkperiodeer then ukchs.dekkperiodeer else ukchs.dekkperiodeer end
when euchs.chassisnr is not null and ukchs.chassisnr is not null then
case when euchs.dekkperiodeer = ukchs.dekkperiodeer then ukchs.dekkperiodeer else ukchs.dekkperiodeer end
when euchs.chassisnr is not null and ukchs.chassisnr is not null then
case when euchs.dekkperiodeer = ukchs.dekkperiodeer then ukchs.dekkperiodeer else ukchs.dekkperiodeer end
when euchs.chassisnr is not null and ukchs.chassisnr is not null then
case when euchs.dekkperiodeer = ukchs.dekkperiodeer then ukchs.dekkperiodeer else ukchs.dekkperiodeer end
when euchs.chassisnr is not null and ukchs.chassisnr is not null then

  ,case when DeliveryCountry is null then '016' else DeliveryCountry end as dekkingsperiode
from
)
```

---voertuig parc eu (alleen extended warranty)
SELECT distinct
  chs.ChassisNr, chs.ChassisModel, dtype.classificationID, ProductRange, left(productrange, 2) as serie, DeliveryDate, DeliveryDate as startdate, dateadd(month, 36, DeliveryDate) as enddate
FROM
  ukchs.Chassis as chs
INNER JOIN
dbo.dim_Warr_Chassis as ch
ON
  clmheader.ChassisID = chs.ChassisID
INNER JOIN
dbo.dim_Warr_Type as dtype
ON
  dtype.TypeID = chs.ChassisID
WHERE
 LEFT(chs.ChassisNr, 2) = '08' OR
  LEFT(chs.ChassisNr, 2) = '09'
and
  DeliveryDate <= (ts '2008-01-01 08:00:00')
and
  DeliveryDate > (ts '2008-01-01 00:00:00')
and
  DeliveryDate < @endyear
and
  DeliveryCountry IN ('047', '001', '005', '048', '110', '115', '089', '003', '002', '057', '081', '047', '088', '028', '039', '037', '029', '027', '018', '009')
and
  chs.ChassisModel IN ('5133', '5647', '5896', '5897', '5898', '5899', '5900', '5901', '5908')
```

)}

---voertuig parc uk
SELECT distinct
  concat('0', (chs.Chassis)) as chassisNr,
  min(chs.ChassisModel) as chassismodel,
  min(dtype.classificationID) as classificationID,
  min(ProductRange) as productrange,
  left(min(productrange), 2) as serie
FROM
  ukclm
inner join
  (select
    FROM
      fact_Warr_ClaimHeader as clmheader
    INNER JOIN
dbo.dim_Warr_Chassis as chs
    ON
      clmheader.ChassisID = chs.ChassisID
    INNER JOIN
dbo.dim_Warr_Type as dtype
    ON
      dtype.TypeID = chs.ChassisID
    WHERE
      LEFT(chs.ChassisNr, 2) = '08' OR
      LEFT(chs.ChassisNr, 2) = '09'
      and
      DeliveryDate <= (ts '2008-01-01 08:00:00')
      and
      DeliveryDate > (ts '2008-01-01 00:00:00')
      and
      DeliveryDate < @endyear
      and
      DeliveryCountry IN ('047', '001', '005', '048', '110', '115', '089', '003', '002', '057', '081', '047', '088', '028', '039', '037', '029', '027', '018', '009')
      and
      chs.ChassisModel IN ('5133', '5647', '5896', '5897', '5898', '5899', '5900', '5901', '5908')
)
```
APPENDIX B: Delayed reporting of defects

A failure can be classified as hard failures or soft failures. Soft failures are failures that result in a degraded performance of the truck, but do not make them nonoperational. Some examples of soft failures are minor oil leaks, slow starting engine, or unusual engine noise. Hard failures are failures that render a truck inoperative until repaired. An engine that does not start, or stops running, belongs to the hard failure category.

Figure 29: Causes of unclean data.

For both hard and soft failures there are different chances of detection by the customer. A rear-brake light failure is harder to detect than an engine that does not start. Undetected failures cause delays in reporting to a dealer or other service point. Therefore, the time to report minus the time to defect is
higher than zero. The delay in reporting will be even stronger if the customer does not see a motivation for immediate reporting after detecting the defect. This effect is generally stronger in soft failures where a failure does not make a vehicle inoperative. Table 21 shows the chance that a reporting delay occurs.

<table>
<thead>
<tr>
<th>Delays in reporting</th>
<th>Low detection chance</th>
<th>High detection chance</th>
</tr>
</thead>
<tbody>
<tr>
<td>Soft failures</td>
<td>High chance for delay</td>
<td>Likely chance for delay</td>
</tr>
<tr>
<td>Hard failures</td>
<td>Low chance for delay</td>
<td>Very low chance for delay</td>
</tr>
</tbody>
</table>

From the position where this research is performed, it is not possible to communicate with customers or dealers to ask which components have reporting delays. Any detection has to come from examining the data. Figure 30 shows the months to defect per claim for the stabilizer rubbers of the front axle for all trucks in the dataset. The graph of the stabilizer rubbers of the rear axle looks very similar.

Figure 30: Component number 10306 has a strong increase in observed defects after every 12 months.

In the above example the exact inspection intervals are unknown because the inspection dates for each truck are not available. From the data it looks like the component is inspected annually, except for the first year. The defects do not strictly break down annually, since there are also observed defects every other month. Therefore, it seems possible that a soft failure can turn in to a hard failure. Hard failures are usually noticed by the user and affect the performance of the vehicle such that a repair is required immediately.

For the observed defects of the component shown in Figure 30 it is likely that the defects observed in the annual peaks actually broke down earlier but were not detected until inspection by the dealer. This effect is also strong for component number 10606. For almost all other components the effect is nonexistent or very weak. The probable reason for this is because this study does not look at wear & tear components or other components that are maintained preventively. Ideally, there should not be any replacements at all for components that have warranty.

Concluding, the effect happens very rarely in the data, it is hard to determine whether they are truly delayed reports and it is not possible to determine the last inspection time via the available data. Therefore, the expected improvements to the mathematical model is low compared to the required
effort for developing a technique that detects and corrects the phenomenon, and as such will not be done in this project.
APPENDIX C: Many claims at end of warranty or repair & maintenance contract

For several components it is observed that many defects are claimed in the very last period of the warranty or repair & maintenance contract. It would be naïve to assume that everyone claims a defect on the day that it occurs, and has the same determination of when a component is defective. In order to determine whether this phenomenon can really be associated with normal failure behavior, several interviews have been done with experts on warranty within DAF.

Remember from section 1.1.2 that all DAF dealers are franchise. Being franchise means that they are an independent organization, responsible for their own profits. With this in mind, the following arguments are formed to explain the phenomenon:

1. A component had a defect just after end of contract. The dealer then predates the claim in order to satisfy his customer and receive compensation from DAF Factory.
2. Dealer buys the truck back from the customer near the end of the contract. The dealer then performs a thorough check of the vehicle and claims many defects.
3. Customers see that their contract is ending and rush to the dealer to get every half-failure repaired for free while they still can.
4. Sometimes it is hard to determine whether something is defective. An example could be excessive engine oil use. A dossier is then opened to track the oil use for a longer period. When the contract or warranty is nearing its end while the measurement is not finished yet, the decision to repair or not is then decided just before the end of the contract.

It can be argued that a large peak in claims near the end of the contract is not a good representation of the normal lifetime of a component. For each of the categories, it could be argued what correction is best to do to repair the data. For the predating of claims they should be removed from the sample since they actually did not break down during the observation period. The second category is already more difficult since more factors play a role in the decision. The dealer will be much stricter in checking the quality of the components, since replacing components is free and would probably result in a higher truck value for reselling. In this process of maximizing profit margins they might also replace a component that could actually work for quite some time still. For example a component that has a squeak or makes more noise than normal is a good example where the word “defect” could be stretched. Also, any defects that the customer did not detect would probably be detected and claimed during this thorough check. The undetected defects should be seen as an interval-censored defect, while the replacements of currently working components could arguably be removed. The third category is about defects that are detected by the customer but were not reported, and thus not observed by DAF. These defects should be modeled as interval-censored defects between the last inspection and the observed defect time. The same for the fourth category: it is hard to determine the actual time of defect since it usually starts as a small problem.

Unfortunately, from the data it is not possible to see enough details of the defect to determine in which of the four categories a defect belongs when observed at the end of the contract. It is also not known which of the four categories described above are most frequent. Without knowing the details of the
defect, the best method to correct the data cannot be chosen. Still, it is interesting to see how often it happens in order to get an indication of the size of the phenomenon. A detection method will be developed in the next section.

**Detecting the peak in claims near the end of a contract**

The peak in claims at the end of the contract only happens in several occasions. A method will developed in this section which is able to recognize when there is such a peak of claims at the end of a contract for a component number. When this is the case, the size of the peak relative to the total number of defects for that component number will be measured.

First, an indicator has to be developed for each defect that shows how much time on a contract has passed until the defect was observed. This is done using a ratio between the time to defect and the contract length. Because there are many different contract lengths, a normalized indicator is used:

\[
Percentage \text{ of contract time used before defect} = \frac{component \text{ time to defect in days}}{contract \text{ length in days}}
\]

This results in a value between 0 and 1 depending on when the component failure was claimed during the contract.

Then, a histogram is created (50 bins) for each component number to visualize the process of the claims during a contract. Figure 31 is an example of such a histogram. In this case, there looks to be a peak in claims near the end of the contract. In the figure it can be seen that there are many observed defects in the last percent of the contract duration. For the most common three year contract this would translate to a defect in the last 11 days of the contract.

**Figure 31: Example histogram showing many defects claimed in the very last period of the contract**

Outlier detection will be used to determine whether the end of the histogram is normal or not. This is very often done using the “3 * standard deviation” rule to determine outliers. In this case, the fact that
it is used so often does not mean that it can work as intended (Leys, Ley, Klein, Bernard, & Licata, 2013). A big pitfall in the “3 * standard deviation” rule is that the data needs to be normally distributed in order to successfully find outliers. Another pitfall is that the calculation of the standard deviation is sensitive to presence of outliers. This is because it takes the squared difference between the sample and the mean.

The Median Absolute Difference (MAD) is a robust measure of the variability of a univariate sample of quantitative data. The formula is given below (Davenport, 2013):

\[
MAD = b \text{Median}_i \left( |x_i - \text{Median}_j(x_j)| \right)
\]

The formula will be explained with a simple example. Imagine the ordered set [2, 6, 6, 12, 17, 25, 32], the median \( M_j \) of this set is 12. Then create a new list by subtracting 12 from each value in the old list, resulting in [-10, -6, -6, 0, 5, 13, 20]. Take the absolute value of this list and sort it: [0, 5, 6, 6, 10, 13, 20]. The median \( M_i \) is then 6. Multiply this number with \( b \) to get the MAD.

To use MAD as a consistent estimator for the estimation of the standard deviation, a constant “\( b \)” is used. For normally distributed data: \( b = 1.4826 \). Because it is not possible to assume that all the data is normally distributed, and because it is also not known which distribution is underlying the data, \( b \) can be ignored, so \( b = 1 \). An advantage of MAD compared to the standard deviation is that MAD uses the median. It is therefore not sensitive to outliers in the data. When looking for outliers on both the left side and the right side of the graph, the data needs to be symmetrical. This is quite logical, because the same cut-off value is used while one tail of the data can be much longer than the other. For this project, only a one-sided outlier test is required so the cut-off value can be configured so that it works well for finding outliers on one tail. An interesting solution for finding outliers in unsymmetrical data is the “Double-MAD” technique by (Rosenmai, 2013).

An algorithm is developed to determine whether there is an outlier in number of claims at the end of the contract for all the component numbers at DAF. It uses the histogram that is shown in Figure 31. First, the median and MAD is calculated using the size of each bin in the histogram. Then, a looping function is developed that starts at the last percent (the last bin) of the histogram to check for outliers. The following is tested (\( t=0 \)):

\[
\text{Count of observed defects in } bin_{100-t} > (\text{median} + 5 \times MAD)
\]

If true, the bin is an outlier. Then, \( t \) is increased by one and the test runs again. If false, no outlier is detected in \( bin_{100-t} \) and the looping function ends. A multiplier of 5 is chosen as it returned the best results for this situation. Because the arguments for the cause of this outlier is only true for the very last part of the data, the function ends when \( t>1 \). When the function ends with \( t>0 \), a notification will be given for the user that indicates the size of the peak. at the tested phenomena is present for this

---

2 This technique divides the data into two halves, separated by the median of the complete data. Then, a separate MAD and cut-off value can be created to find the outliers for each half of the data.
component number, together with a percentage number that tells the user how large the peak is. The peak size is calculated by taking the sum of the defects in each bin that is recognized as an outlier. This number is then shown in the summary screen, shown in Table 7.

\[
Defects \text{ in peak} = \sum_{t=0}^{t} \text{count(bin}_{100-t})
\]

Eq. 37

Sometimes, the number of defects in the last one or two percent of the contract duration account for more than 20% (up to 48%) of the total defects. When a user is making a prediction using the mathematical model, they will receive a notification whether the phenomenon is present in the predicted component lifetime, and how large it is. The following section will give a detailed overview of this phenomenon, such that any further research on this can get a head start.

**Evaluating the predictions for components with many defects at the end of the contract**

To measure the effect of this phenomenon on the performance of the predictions, it is possible to empirically compare the predicted failures per hundred vehicles (FphV) with the data via the Kaplan-Meier empirical distribution function. How the evaluation is setup is explained in section 3.4.1. Because the number of observed trucks is very low after three years, the empirical distribution function is less powerful after the third FphV value. The evaluation therefore focuses on the first three FphV years. For easier readability, the prediction accuracies for all components for all subsets j and for all “FphV after year i”, as well as the standard deviation of the prediction errors for all i,j are repeated from the evaluation and shown in Table 22 and Table 23.

| Table 22: Prediction accuracies (of all components k) for all i,j (left) and the number of compared components (right) |
|---|---|---|---|---|---|---|---|---|---|---|---|
|   | FphV after 1 year | FphV after 2 years | FphV after 3 years | FphV after 4 years | FphV after 5 years | FphV after 6 years | FphV after 7 years | FphV after 8 years |
| subset 1 | -0.641 | -0.593 | -0.429 | 0.084 | 0.435 | 0.227 | 0.308 | 0.157 |
| subset 2 | -0.496 | -0.495 | -0.366 | -0.036 | 0.250 | 0.121 | 0.100 | -0.019 |
| subset 3 | -0.440 | -0.467 | -0.397 | 0.088 | 0.334 | 0.074 | -0.048 | -0.230 |
| subset 4 | -0.266 | -0.362 | -0.286 | 0.192 | 0.358 | 0.073 | -0.086 | -0.275 |
| subset 5 | -0.145 | -0.243 | -0.149 | 0.403 | 0.533 | 0.123 | -0.073 | -0.273 |
| subset 6 | -0.048 | -0.156 | -0.055 | 0.517 | 0.635 | 0.140 | -0.067 | -0.298 |
| subset 7 | 0.093 | -0.058 | 0.007 | 0.535 | 0.614 | 0.154 | -0.058 | -0.300 |
| subset 8 | 0.084 | -0.053 | 0.009 | 0.524 | 0.596 | 0.157 | -0.051 | -0.298 |
| subset components compared | 15 | 83 | 192 | 291 | 398 | 501 | 592 | 651 |
The same prediction accuracy evaluation as done in chapter 3 is also done for only the components that have a detected peak in claims at the end of the contract period. The results are shown in Table 24 and Table 25.

**Table 23: Standard deviation (of all components k) of the prediction errors for all i,j**

<table>
<thead>
<tr>
<th></th>
<th>FphV after 1 year</th>
<th>FphV after 2 years</th>
<th>FphV after 3 years</th>
<th>FphV after 4 years</th>
<th>FphV after 5 years</th>
<th>FphV after 6 years</th>
<th>FphV after 7 years</th>
<th>FphV after 8 years</th>
</tr>
</thead>
<tbody>
<tr>
<td>subset 1</td>
<td>0.267</td>
<td>0.797</td>
<td>1.894</td>
<td>6.656</td>
<td>9.570</td>
<td>9.769</td>
<td>14.862</td>
<td>15.071</td>
</tr>
<tr>
<td>subset 2</td>
<td>0.272</td>
<td>1.348</td>
<td>3.633</td>
<td>6.603</td>
<td>7.350</td>
<td>8.912</td>
<td>10.504</td>
<td>10.701</td>
</tr>
<tr>
<td>subset 3</td>
<td>0.227</td>
<td>1.035</td>
<td>2.453</td>
<td>5.546</td>
<td>7.821</td>
<td>8.795</td>
<td>10.486</td>
<td>11.501</td>
</tr>
<tr>
<td>subset 4</td>
<td>0.124</td>
<td>0.606</td>
<td>1.594</td>
<td>4.259</td>
<td>6.770</td>
<td>9.320</td>
<td>12.462</td>
<td>14.099</td>
</tr>
<tr>
<td>subset 5</td>
<td>0.073</td>
<td>0.340</td>
<td>0.954</td>
<td>3.373</td>
<td>5.335</td>
<td>8.683</td>
<td>11.265</td>
<td>13.174</td>
</tr>
<tr>
<td>subset 6</td>
<td>0.052</td>
<td>0.197</td>
<td>0.606</td>
<td>2.796</td>
<td>4.957</td>
<td>7.233</td>
<td>10.261</td>
<td>12.028</td>
</tr>
<tr>
<td>subset 7</td>
<td>0.084</td>
<td>0.158</td>
<td>0.266</td>
<td>2.467</td>
<td>4.737</td>
<td>6.522</td>
<td>9.935</td>
<td>11.905</td>
</tr>
<tr>
<td>subset 8</td>
<td>0.052</td>
<td>0.052</td>
<td>0.036</td>
<td>2.390</td>
<td>4.788</td>
<td>6.395</td>
<td>9.311</td>
<td>10.912</td>
</tr>
</tbody>
</table>

**Table 24: Prediction accuracies (of only components with many defects at the end of the contract) for the different subsets (left) and the number of compared components (right)**

<table>
<thead>
<tr>
<th></th>
<th>FphV after 1 year</th>
<th>FphV after 2 years</th>
<th>FphV after 3 years</th>
<th>FphV after 4 years</th>
<th>FphV after 5 years</th>
<th>FphV after 6 years</th>
<th>FphV after 7 years</th>
<th>FphV after 8 years</th>
<th>Nr of components compared</th>
</tr>
</thead>
<tbody>
<tr>
<td>subset 1</td>
<td>-0.590</td>
<td>-0.466</td>
<td>-0.139</td>
<td>0.780</td>
<td>1.433</td>
<td>1.039</td>
<td>1.252</td>
<td>0.987</td>
<td>7</td>
</tr>
<tr>
<td>subset 2</td>
<td>-0.525</td>
<td>-0.387</td>
<td>0.092</td>
<td>0.733</td>
<td>1.149</td>
<td>0.711</td>
<td>0.597</td>
<td>0.360</td>
<td>22</td>
</tr>
<tr>
<td>subset 3</td>
<td>-0.409</td>
<td>-0.375</td>
<td>0.013</td>
<td>0.574</td>
<td>0.881</td>
<td>0.494</td>
<td>0.406</td>
<td>0.112</td>
<td>42</td>
</tr>
<tr>
<td>subset 4</td>
<td>-0.214</td>
<td>-0.260</td>
<td>0.118</td>
<td>0.524</td>
<td>0.717</td>
<td>0.395</td>
<td>0.202</td>
<td>-0.102</td>
<td>60</td>
</tr>
<tr>
<td>subset 5</td>
<td>-0.053</td>
<td>-0.140</td>
<td>0.228</td>
<td>0.580</td>
<td>0.729</td>
<td>0.350</td>
<td>0.134</td>
<td>-0.159</td>
<td>72</td>
</tr>
<tr>
<td>subset 6</td>
<td>0.094</td>
<td>-0.049</td>
<td>0.281</td>
<td>0.451</td>
<td>0.463</td>
<td>0.082</td>
<td>-0.114</td>
<td>-0.338</td>
<td>80</td>
</tr>
<tr>
<td>subset 7</td>
<td>0.070</td>
<td>-0.136</td>
<td>0.102</td>
<td>0.183</td>
<td>0.147</td>
<td>-0.166</td>
<td>-0.313</td>
<td>-0.479</td>
<td>83</td>
</tr>
<tr>
<td>subset 8</td>
<td>0.081</td>
<td>-0.181</td>
<td>0.037</td>
<td>0.087</td>
<td>0.042</td>
<td>-0.233</td>
<td>-0.349</td>
<td>-0.495</td>
<td>83</td>
</tr>
</tbody>
</table>
It is interesting to see that the components with many defects at the end of the contract perform better than the rest of the components with better prediction accuracies throughout the comparison. The standard deviations are similar. There will be no investigation why this effect is occurring, since it is not in the scope of this research.

Because this phenomenon is real and sometimes quite heavy, it is recommended for DAF to do further analysis on this phenomenon to see what the effects are on the predictability of these components. Also, when a warranty item is replaced in a structural pattern, it could be better to switch the category of the component from a warranty component to a maintenance component.

**Detailed overview of the components with many defects at the end of the contract**

In order to help DAF analyzing the components where this phenomenon has been detected, an overview of the component numbers is given in the confidential version of this report.

<table>
<thead>
<tr>
<th>Subset</th>
<th>FphV after 1 year</th>
<th>FphV after 2 years</th>
<th>FphV after 3 years</th>
<th>FphV after 4 years</th>
<th>FphV after 5 years</th>
<th>FphV after 6 years</th>
<th>FphV after 7 years</th>
<th>FphV after 8 years</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sub 1</td>
<td>0.218</td>
<td>0.971</td>
<td>2.432</td>
<td>9.153</td>
<td>12.364</td>
<td>12.780</td>
<td>20.305</td>
<td>20.686</td>
</tr>
<tr>
<td>Sub 2</td>
<td>0.338</td>
<td>1.127</td>
<td>3.075</td>
<td>5.745</td>
<td>7.382</td>
<td>8.078</td>
<td>12.195</td>
<td>13.062</td>
</tr>
<tr>
<td>Sub 3</td>
<td>0.225</td>
<td>0.655</td>
<td>1.776</td>
<td>3.958</td>
<td>5.440</td>
<td>6.853</td>
<td>10.680</td>
<td>11.649</td>
</tr>
<tr>
<td>Sub 4</td>
<td>0.148</td>
<td>0.499</td>
<td>1.457</td>
<td>3.444</td>
<td>6.516</td>
<td>10.583</td>
<td>15.977</td>
<td>18.651</td>
</tr>
<tr>
<td>Sub 5</td>
<td>0.102</td>
<td>0.360</td>
<td>1.201</td>
<td>3.301</td>
<td>6.802</td>
<td>10.662</td>
<td>15.635</td>
<td>18.589</td>
</tr>
<tr>
<td>Sub 6</td>
<td>0.083</td>
<td>0.301</td>
<td>1.036</td>
<td>2.962</td>
<td>6.690</td>
<td>10.681</td>
<td>15.806</td>
<td>20.187</td>
</tr>
<tr>
<td>Sub 7</td>
<td>0.047</td>
<td>0.143</td>
<td>0.461</td>
<td>2.842</td>
<td>7.315</td>
<td>11.522</td>
<td>18.300</td>
<td>21.909</td>
</tr>
<tr>
<td>Sub 8</td>
<td>0.028</td>
<td>0.089</td>
<td>0.070</td>
<td>3.194</td>
<td>8.388</td>
<td>12.825</td>
<td>19.254</td>
<td>22.464</td>
</tr>
</tbody>
</table>
APPENDIX D: Basic statistics
There are several terms used throughout this paper. To get a better understanding, explanations on the different terms are explained here.

Skewness and kurtosis
Skewness is a measure of the asymmetry of the probability distribution (Safari). Negative skew, or left skew, indicates that the area under the graph is larger on the left side of the mode. Positive skew, or right skew, indicates that the area under the graph is larger on the right side of the mode.

![Forms of skewness](Safari)

Kurtosis is the degree of peakedness of a distribution (Weisstein). A high kurtosis indicates that the variance results more from infrequent extreme deviations, as opposed to frequent modestly sized deviations. The kurtosis of any univariate normal distribution is 3. Kurtosis less than 3 are platykurtic, more than 3 are leptokurtic. It is a common practice to use excess kurtosis instead, which is simply the original kurtosis minus 3.

![Forms of kurtosis](Safari)

Shape, scale and location parameters
For every different distribution there are different parameters in the function. These parameters can be categorized as the shape, scale and location parameters. Each parameter has a specific influence on how the distribution looks like.

A location parameter controls the position of the distribution on the x-axis (Vose Software, 2007). Therefore, the mean and mode should follow a change in the location parameter exactly. If the location parameter increases by 3, the mean and mode should also increase by 3.
A scale parameter controls the spread of the distribution on the x-axis (Vose Software, 2007). This parameter appears in the equation for a distribution’s variance \( \sigma \) for normal distribution.

A shape parameter controls the shape (e.g. skewness, kurtosis) of the distribution (Vose Software, 2007). It will appear in the pdf in a way that controls the manipulation of x in a non-linear fashion, usually as a coefficient of x. There is no shape parameter in a distribution if the distribution always takes the same shape. This happens for example in the exponential, normal and triangle distribution.

The normal distribution has two parameters location and scale: \( \mu \) (mean), \( \sigma^2 \) (variance).

![Figure 34: The probability density function of a normal distribution. (Damodaran, Statistical Distributions).]

As shown in the graph, changing the parameters changes the behavior of the distribution function.

**Continuous or discrete**

The difference between continuous and discrete can be found in the values that the data can be. Discrete data can only have integer values. Like throwing a die, the values can only take values 1, 2, … 6 but not 1.5. Continuous data on the other hand can take any value, for example the temperature in a room.

**Distribution support**

Support means that the function is not zero-valued on a set interval. The normal distribution has infinite support \((-\infty, \infty)\), meaning each value is possible. The Exponential distribution has only positive values \([0, \infty)\), also called a non-negative distribution. It is also possible to set specific intervals on which the function has support, for example with the triangle distribution.

**Non-parametric statistics**

Nonparametric statistics covers techniques that do not rely on data belonging to any particular distribution. They make no assumptions that the data follow a given distribution. This results in a wider applicability and more robust methods. The downside is that non-parametric tests usually have less power. Several known non-parametrical tests are the Anderson-Darling test, Kaplan-Meier and the Kolmogorov-Smirnov test.
APPENDIX E: List of non-negative distributions

This appendix shows an extensive list of popular non-negative distributions. The probability density function and the cumulative distribution function for every distribution can be found at Mathwave (Mathwave). The description from Wikipedia gives an indication for what the distribution is used for by many. This is one of the few cases where a peer-reviewed website such as Wikipedia is useful to extract information, as these popular statistical distributions have been reviewed by experts many times. The references to the URL are not in the bibliography because it would cause cluttering of the bibliography. The mathematical model does not fit all the distributions in this list. This list is useful for when there are plans for extending the model to fit more distributions.

Table 26: List of non-negative distributions

<table>
<thead>
<tr>
<th>Distribution</th>
<th>Parameters</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Burr</strong></td>
<td>$k, \alpha, \beta$</td>
<td>Used to model household income <a href="https://en.wikipedia.org/wiki/Burr_distribution">link</a></td>
</tr>
<tr>
<td><strong>Chi-Squared</strong></td>
<td>$\nu$</td>
<td>Used in inferential statistics. <a href="https://en.wikipedia.org/wiki/Chi-squared_distribution">link</a></td>
</tr>
<tr>
<td><strong>Dagum</strong></td>
<td>$k, \alpha, \beta, \gamma$</td>
<td>Income studies (personal income) <a href="https://en.wikipedia.org/wiki/Dagum_distribution">link</a></td>
</tr>
<tr>
<td><strong>Erlang</strong></td>
<td>$m, \beta$</td>
<td>Special case of gamma distribution. Used in queueing systems and biomathematics. <a href="https://en.wikipedia.org/wiki/Erlang_distribution">link</a></td>
</tr>
<tr>
<td><strong>Exponential</strong></td>
<td>$\lambda$</td>
<td>Describes time between events in poisson. Special case of gamma distribution. Memoryless. Widely used for many cases. <a href="https://en.wikipedia.org/wiki/Exponential_distribution">link</a></td>
</tr>
<tr>
<td><strong>F Distribution</strong></td>
<td>$\nu_1, \nu_2$</td>
<td>Used in analysis of variance. Frequently used as the null distribution of a test statistic. <a href="https://en.wikipedia.org/wiki/F-distribution">link</a></td>
</tr>
<tr>
<td><strong>Fatigue Life (Birnbaum-Saunders)</strong></td>
<td>$\alpha, \beta$</td>
<td>Inverse Weibull distribution. Used in hydrology for extreme events. <a href="https://en.wikipedia.org/wiki/Fatigue_life_(Birnbaum-Saunders)">link</a></td>
</tr>
<tr>
<td><strong>Frechet</strong></td>
<td>$\alpha, \beta, \gamma$</td>
<td>Three parametrizations in common use. Used in conometrics, accelerated life testing, fading of wifi signal power, neuroscience. <a href="https://en.wikipedia.org/wiki/Fr%C3%A9chet_distribution">link</a></td>
</tr>
<tr>
<td><strong>Gamma</strong></td>
<td>$\alpha, \beta, \gamma$</td>
<td>Three parametrizations in common use. Used in conometrics, accelerated life testing, fading of wifi signal power, neuroscience. <a href="https://en.wikipedia.org/wiki/Gamma_distribution">link</a></td>
</tr>
<tr>
<td><strong>Generalized Gamma</strong></td>
<td>$k, \alpha, \beta, \gamma'$</td>
<td>Generalization of the two-parameter gamma distribution. Since many distributions commonly used for parametric models in survival analysis (such as the Weibull distribution and the log-normal distribution) are special cases of the generalized gamma, it is sometimes used to determine which parametric model is</td>
</tr>
<tr>
<td>Distribution</td>
<td>Parameters</td>
<td>Description</td>
</tr>
<tr>
<td>-----------------------</td>
<td>------------</td>
<td>-----------------------------------------------------------------------------------------------</td>
</tr>
<tr>
<td>Inverse Gaussian</td>
<td>$\lambda, \mu, \gamma$</td>
<td>Also known as Wald distribution. First used as the time to first passage of a brownian motion.</td>
</tr>
<tr>
<td>Levy</td>
<td>$\sigma, \gamma$</td>
<td>Together with the normal and cauchy distribution, it is one of the few distributions that are stable. Used for the frequency of geomagnetic reversals and brownian motion predictions.</td>
</tr>
<tr>
<td>Log-Gamma</td>
<td>$\alpha, \beta$</td>
<td>Used as joint prior distribution in bayesian analysis.</td>
</tr>
<tr>
<td>Log-Logistic</td>
<td>$\alpha, \beta, \gamma$</td>
<td>Also known as Fisk distribution. Used in survival analysis for events whose rate increases initially and decreases later.</td>
</tr>
<tr>
<td>Lognormal</td>
<td>$\sigma, \mu, \gamma$</td>
<td>Continuous probability distribution of a random variable whose logarithm is normally distributed. Also known as Galton distribution. Used widely, also for reliability. Used to model times to repair a maintainable system. Also, coefficients of friction and wear may be treated as having a lognormal distribution.</td>
</tr>
<tr>
<td>Nakagami</td>
<td>$m, \Omega$</td>
<td>Related to the gamma distribution. Used to model attenuation of wireless signals traversing multiple paths.</td>
</tr>
<tr>
<td>Pareto (First Kind)</td>
<td>$\alpha, \beta$</td>
<td></td>
</tr>
<tr>
<td>Pareto (Second Kind)</td>
<td>$\alpha, \beta$</td>
<td></td>
</tr>
<tr>
<td>Pearson Type 5</td>
<td>$\alpha, \beta, \gamma$</td>
<td>Is an inverse gamma distribution. Pearson models are used in financial markets, for example to model the stochastic nature of the volatility of rates and stocks.</td>
</tr>
<tr>
<td>Pearson Type 6</td>
<td>$\alpha_1, \alpha_2, \beta$</td>
<td>Is a beta prime distribution or F-distribution.</td>
</tr>
<tr>
<td>Rayleigh</td>
<td>$\sigma, \gamma$</td>
<td></td>
</tr>
<tr>
<td>Rice</td>
<td>$\nu, \sigma$</td>
<td>Related to Rayleigh</td>
</tr>
<tr>
<td>Weibull</td>
<td>$\alpha, \beta, \gamma$</td>
<td>Used in survival analysis, reliability engineering, failure analysis and various others. A beta &lt; 1 indicated infant mortality. Beta = 1 indicates constant failure rate, beta &gt; 1 indicates an aging process.</td>
</tr>
</tbody>
</table>
APPENDIX F: Why Rank Regression is not suitable for heavily censored data

DAF currently has a prediction model called “Automated Weibulls” that predicts component lifetime using Maximum Likelihood Estimation (MLE) and Rank Regression (RR), also called Least Squares. Both methods give different results. DAF is not sure which of the methods should be used in their situation, so this is investigated. RR mathematically fits the best straight line through a set of points in an effort to estimate the values of the relevant parameters. Drawing a straight line is only possible when the CDF of a distribution can be linearized. The straight line is drawn such that the sum of the squares of the distance of the points to the fitted line is minimized. The minimization can happen in two directions, on the Y axis or the X axis. Figure 35 from Reliawiki (Racaza, 2015) shows the technique very well.

Advantages of the rank regression are that it shows good result and has relatively easy calculations, especially for distributions that can be linearized. It is also possible to use the correlation coefficient ($r^2$), a good method for measuring the goodness of fit of the values of the relevant parameters. Rank regression works well for uncensored data. Unfortunately, it performs poorly when used on censored data where the censored data is not spread evenly throughout the observed failures. This is because the method ranks the failures from shortest to longest in time to defect. When a defect is censored, the failure time is unknown. This makes it possible to have different rankings which are all plausible. It could be that a suspension would have outlived the next observed failure, or it could have died before the next failure. This choice influences the rank that the failure should have, and either choice results in a different fit. A probability is calculated where the suspension and failures ranks have the highest chance of occurring, called Mean Order Numbers. The failures are then placed according to these calculated numbers. A case example from ReliaSoft (ReliaSoft) shows just how disturbing this shortfall is when rank regression is dealing with censored data:
Table 27: Case example from ReliaSoft showing shortfalls when rank regression is dealing with censored data.

<table>
<thead>
<tr>
<th>Case 1</th>
<th></th>
<th>Case 2</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Item Number</td>
<td>State** &quot;F&quot; or &quot;S&quot;</td>
<td>Life of an item, hr</td>
<td>Item number</td>
</tr>
<tr>
<td>1</td>
<td>$F_1$</td>
<td>1,000</td>
<td>1</td>
</tr>
<tr>
<td>2</td>
<td>$S_1$</td>
<td>1,100</td>
<td>2</td>
</tr>
<tr>
<td>3</td>
<td>$S_2$</td>
<td>1,200</td>
<td>3</td>
</tr>
<tr>
<td>4</td>
<td>$S_3$</td>
<td>1,300</td>
<td>4</td>
</tr>
<tr>
<td>5</td>
<td>$F_2$</td>
<td>10,000</td>
<td>5</td>
</tr>
</tbody>
</table>

* F - Failed, S - Suspended

In this example, the suspensions have clearly different suspension times. Calculating the values of the relevant parameters for both cases would lead to a scale parameter lifetime of 11.400 hours for both cases. Using MLE, the scale parameter for case 1 is 6920 hours and 21.300 hours for case 2. The poor performance of the rank regression grows stronger the more censored the data is. This makes rank regression unusable for this research project. It is already known that the used data for this project is heavily censored. This censoring would completely break the rank regression method. MLE does not look at ranks or plotting positions, but instead looks at each unique time to defect or time to suspension. This means that it handles censored data much better. Therefore, MLE will be used to fit the parameters of the distributions to the data.
APPENDIX G: Mathematical model R code

options(scipen = 500)

library("dplyr", lib.loc="~/R/win-library/3.2")
library("stringr", lib.loc="~/R/win-library/3.2")
library("magrittr", lib.loc="~/R/win-library/3.2")
library("stringi", lib.loc="~/R/win-library/3.2")
library("lubridate", lib.loc="~/R/win-library/3.2")
library("ggprop", lib.loc="~/R/win-library/3.2")
library("survival", lib.loc="C:/Program Files/R-R-3.2.3/library")
library("tseries", lib.loc="~/R/win-library/3.2")
library("trend", lib.loc="~/R/win-library/3.2")
library("plotrix", lib.loc="~/R/win-library/3.2")
library("survCl", lib.loc="~/R/win-library/3.2")

#Set working directory
setwd("C:/Documents/Dropbox/Documents/School/TUE/Master/Master thesis/Daf Trucks/R data")

#choose configurations for the model
outlierstep <- 5 # what mad value in last period of contract duration is acceptable (recursive function)
outlierstep <- 5 # stop criteria for the endcontract outlier check
bysize <- 50 #number of bins for the endcontract outlier check histogram
maturitytimesamples <- 10 #minimum observed defects before doing predictions and maturity check
maturitytype <- 20 #interval in years for each subset. make it a multiple or division of 8, <8
testseries <- c("XF", "CF", "LF") #serie to test, choose XF/CF/LF
testchassis <- c("FT") #chassis to test, choose FA/FT

#Load source files (claims and truckpark)
sourceclaims <- list()
sourcetrucks <- list()

sourceclaims[[1]] <- read.csv2("source/data/claims2009.csv", stringsAsFactors=FALSE)
sourceclaims[[2]] <- read.csv2("source/data/claims2010.csv", stringsAsFactors=FALSE)
sourceclaims[[3]] <- read.csv2("source/data/claims2011.csv", stringsAsFactors=FALSE)
sourceclaims[[4]] <- read.csv2("source/data/claims2012.csv", stringsAsFactors=FALSE)
sourceclaims[[5]] <- read.csv2("source/data/claims2013.csv", stringsAsFactors=FALSE)
sourceclaims[[6]] <- read.csv2("source/data/claims2014.csv", stringsAsFactors=FALSE)
sourceclaims[[7]] <- read.csv2("source/data/claims2015.csv", stringsAsFactors=FALSE)
sourceclaims[[8]] <- read.csv2("source/data/claims2016.csv", stringsAsFactors=FALSE)
sourcetrucks[[1]] <- read.csv2("source/data/trucks2009.csv", stringsAsFactors=FALSE)
sourcetrucks[[2]] <- read.csv2("source/data/trucks2010.csv", stringsAsFactors=FALSE)
sourcetrucks[[3]] <- read.csv2("source/data/trucks2011.csv", stringsAsFactors=FALSE)
sourcetrucks[[4]] <- read.csv2("source/data/trucks2012.csv", stringsAsFactors=FALSE)
sourcetrucks[[5]] <- read.csv2("source/data/trucks2013.csv", stringsAsFactors=FALSE)
sourcetrucks[[6]] <- read.csv2("source/data/trucks2014.csv", stringsAsFactors=FALSE)
sourcetrucks[[7]] <- read.csv2("source/data/trucks2015.csv", stringsAsFactors=FALSE)
sourcetrucks[[8]] <- read.csv2("source/data/trucks2016.csv", stringsAsFactors=FALSE)

for (k in 1:length(sourceclaims)){

  #repair data that broke during reading csv
  sourceclaims[[k]]$defectdate <- lubridate::dmy_hm(sourceclaims[[k]]$defectdate)
  sourceclaims[[k]]$deliverydatefinal <- lubridate::dmy_hm(sourceclaims[[k]]$deliverydatefinal)
  sourceclaims[[k]]$startdatefinal <- lubridate::dmy_hm(sourceclaims[[k]]$startdatefinal)
  sourceclaims[[k]]$enddatefinal <- lubridate::dmy_hm(sourceclaims[[k]]$enddatefinal)
  sourceclaims[[k]]$improveddate <- lubridate::dmy_hm(sourceclaims[[k]]$improveddate)
  sourcetrucks[[k]]$deliverydatefinal <- lubridate::dmy_hm(sourcetrucks[[k]]$deliverydatefinal)
  sourcetrucks[[k]]$startdatefinal <- lubridate::dmy_hm(sourcetrucks[[k]]$startdatefinal)
  sourcetrucks[[k]]$enddatefinal <- lubridate::dmy_hm(sourcetrucks[[k]]$enddatefinal)
  sourceclaims[[k]]$chassismodel <- stringr::str_sub(sourceclaims[[k]]$chassismodel,1,2)
  sourcetrucks[[k]]$chassismodel <- stringr::str_sub(sourcetrucks[[k]]$chassismodel,1,2)
  sourceclaims[[k]]$partscosts <- sourceclaims[[k]]$partscosts+sourceclaims[[k]]$laborcosts+sourceclaims[[k]]$misccosts+sourceclaims[[k]]$handlingcost
  #filter data to have only the truck serie that is tested
  sourcetrucks[[k]] <- dplyr::filter(sourcetrucks[[k]], serie == testseries & chassismodel == testchassis)
  sourceclaims[[k]] <- dplyr::filter(sourceclaims[[k]], serie == testseries & chassismodel == testchassis)
}

#create list to put all test combined testresults (defectcodesummarys) in
allsummarys <- list()

for (datayears in 1:8){

  #create table that stores all results for every unique component number
  defectcodesummary <- dplyr::data_frame(defectcode = unique(sourceclaims[[k]]$defectcode), series = testseries, improveddateNA, useolddataNA, useolddataT, outfilesummaryNA, pettittpvalueNA, pettitchangepointNA, corexresultNA, corexvalueNA, corexvalue=8.0,
}

87
ifelse
ifelse
ifelse
na.rm
na.rm
na.rm
percdifftocontractend, improvedate, claimcost

#filter sourceclaims[[datayears]] to have just the claims for one defect code
c = trucksfiltered
#Component lifetime is comparable, so use old data
obscombinedbeforeimprovement <- dplyr::mutate(obscombinedbeforeimprovement, time = ifelse(is.na(dekkingsperiodefinal.y)$dekkingsperiodefinal.x, daystodefect), dead =
obscombinedafterimprovement <- dplyr::select(obscombinedafterimprovement, time, group, dead)
obscombinedafterimprovement$&time <- as.numeric(as.character(obscombinedafterimprovement$&time))
obscombinedafterimprovement$&time <- as.numeric(as.character(obscombinedafterimprovement$&time))
impcombined <- rbind(obscombinedbeforeimprovement, obscombinedafterimprovement)

impurvdiff <- survival::survdiff(Surv(time, dead) ~ group, data=impcombined)
impchisqvalue <- pchisq(impurvdiff$chisq, 1, lower.tail = FALSE)

#Component lifetime is not comparable, do not use old data
obscombined <- dplyr::left_join(trucksfiltered, claimsfiltered, by = "chassinsnr")

#filter sourceclaims[[datayears]] to have just the claims for one defect code
c = trucksfiltered
#Component lifetime is not comparable, do not use old data
obscombined <- dplyr::left_join(trucksfiltered, claimsfiltered, by = "chassinsnr")

#Check for component improvements and check whether the old data is usable for the new data
if(sum(!is.na(claimsfiltered$improvedate))>0) {
  if(is.na(max(claimsfiltered$improvedate, na.rm = TRUE))){
    trucksbeforeimprovement <- filter(trucksfiltered, deliverydatefinal <= max(claimsfiltered$improvedate, na.rm = TRUE))
    trucksafterimprovement <- filter(trucksfiltered, deliverydatefinal > max(claimsfiltered$improvedate, na.rm = TRUE))
    defectsums$improvedate[] = max(claimsfiltered$improvedate, na.rm = TRUE)
    trucksfiltered <- dplyr::left_join(trucksbeforeimprovement, by = "chassinsnr", group="old"")
    trucksfiltered <- dplyr::left_join(trucksafterimprovement, by = "chassinsnr", group="new"")
    trucksfiltered <- dplyr::mutate(trucksfiltered, claimsfiltered, deliverydatefinal)
    depcoheq <- dplyr::select(trucksfiltered, deliverydatefinal, dead)
    depcoheq$&time <- as.numeric(as.character(depcoheq$&time))
    depcoheq$&time <- as.numeric(as.character(depcoheq$&time))
    impcombined <- rbind(trucksbeforeimprovement, trucksafterimprovement)
    impurvdiff <- survival::survdiff(Surv(time, dead) ~ group, data=impcombined)
    impchisqvalue <- pchisq(impurvdiff$chisq, 1, lower.tail = FALSE)
    if(impchisqvalue>0.05) {
      #Component lifetime is comparable, so use old data
      obscombined <- dplyr::left_join(trucksfiltered, claimsfiltered, by = "chassinsnr")
      defectsums$useolddata[1] = TRUE
      defectsums$useolddatavalue[1] = impchisqvalue
    } else {
      #Component lifetime is not comparable, do not use old data
      obscombined <- dplyr::left_join(trucksafterimprovement, claimsafterimprovement, by = "chassinsnr")
      defectsums$useolddata[1] = FALSE
      defectsums$useolddatavalue[1] = impchisqvalue
    }
    par(mfrow = c(1, 1))
    plot(impurvdiff, main = paste("defect ", test, ", use old data: ", defectsums$useolddata[1], ", N/D(old): ", impurvdiff$N[1], ", N/D(new): ", impurvdiff$N[1], ", ylim = c(min(impurvdiff$Surv), 1), xlab = "Time (days)", ylab = "Kaplan-Meier survival probability"
    )
  } else {
    obscombined <- dplyr::left_join(trucksfiltered, claimsfiltered, by = "chassinsnr")
  }
}

```
defectcodesummary$claimcostdev[i] = sd(obscombined$claimcost, na.rm = TRUE)
par(mfrow=c(2,1))

  t <- 1
  defectcodesummary$defectsinpeak[i] = 0
  if(defectcodesummary$observeddefects[i] > maturityminsamples){
    h <- hist(obscombined$percdiff CONTRACTend, breaks=seq(0,1,i+1=i+1), main = paste("endcontract outlier check for ",test," in",trimmean(obscombined$defectcodesummary[i],datayears,i), ylab = "number of observed defects", xlab = "number of contract time used before defect")
  madd <- stats::mad(h$counts, constant = 1)
  medianh <- median(h$counts)
  
  if(madd != 0) {
    t <- 0
    repeat({
      if(!t){
        #hist(claimsfiltered$percdifftocontractend,breaks=binsize, main = paste(test," minus last ","outlierstep"*"bins (n="defectcodesummary$observeddefects[i],"="defectcodesummary$defectsinpeak[i,"="))
      break
    } else if((medianh*madd*outliermultiplier)<h$counts[binsize-1]){
      t <- t+1
    }
  }
  }
  
  defectcodesummary$defectsinpeak[i] = defectcodesummary$defectsinpeak[i] + length(dplyr::filter(obscombined, obscombined$percdifftocontractend >= outlierstep/binsize)$chassisnr)
  
  defectcodesummary$percentageinpeak[i] = defectcodesummary$defectsinpeak[i]/defectcodesummary$observeddefects[i]
  else {
    if(t>0){
      #hist(claimsfiltered$percdifftocontractend,breaks=binsize, main = paste(test," minus last ","outlierstep"*"bins (n="defectcodesummary$observeddefects[i],"="defectcodesummary$defectsinpeak[i,"="))
      break
    }
  }
  }
  
  # else{
  #  h <- hist(0, main = paste("endcontract outlier check for ",test))
  #}
  
  par(mfrow=c(2,1))
  
  defectcodesummary$observedtrucks[i] = length(obscombined$chassisnr)

  if (defectcodesummary$observeddefects[i] >= maturityminsamples) {
    #obscombinedmaturity <- dplyr::select(obscombined[complete.cases(dplyr::select(obscombined, defectdate, daystocheck))], defectdate, daystocheck)
    #obscombinedmaturity <- arrange(obscombinedmaturity, defectdate)
    obscombinedmatweibull <- obscombined
    obscombinedmatweibull <- obscombined[match(matdate, sample(1:nrow(obscombinedmatweibull)))]
  } else {
    matresults <- list()
    for (imat in 1:10) {
      matdate <- as.Date("2015-12-31") + i months((imat-1)::*((i+1)-datayears)*12)
      obscombinedmatweibull$dekingsperiodefinal.x =
        diff(time(as.Date(ifelse(is.na(dekingsperiodefinal.y), dekingsperiodefinal.x, ifelse(is.na(as.Date(dekingsperiodefinal.x), as.Date(as.Date(dekingsperiodefinal.x), x, units = "days"))), x, units = "days")), origini=1970-01-01)), obscombinedmatweibull$deliverydatefinal.x, units = "days")
      obscombinedinterval <- dplyr::mutate(obscombinedmatweibull, time =
        ifelse(is.na(dekingsperiodefinal.y), dekingsperiodefinal.x, ifelse(is.na(as.Date(dekingsperiodefinal.x), as.Date(dekingsperiodefinal.x), x, units = "days")), x, units = "days")), head = ifelse(is.na(dekingsperiodefinal.y), x, dekingsperiodefinal.x, 0,
        ifelse(is.na(as.Date(dekingsperiodefinal.x), as.Date(dekingsperiodefinal.x), x, units = "days"), x, 0, 1)))
      obscombinedinterval <- dplyr::select(dplyr::filter(obscombinedinterval, time > 0), time, head)
      matweibull[i] =
        if (sum(obscombinedinterval$head) >= 10) {
          matweibull <- list(res = NA)
          matweibull <- try(flexsurv::flexsurvreg(formula = Surv(time, head) ~ 1, data = obscombinedinterval, dist = "weibull"))
          if(is.na(matweibull[i])){
            matresults$shape[[i-imat]] = NA
            matresults$scale[[i-imat]] = NA
            matresults$fphv[[i-imat]] = NA
          } else {
          }  
```
if (length(na.remove(matresults$fphv)) > 0) {
  # plot(matresults$fphv, type = "n")
  # lines (matresults$fphv)
  fit = tryCatch(do.call(flexsurvreg, formula = Surv(time, cens) ~ 1, data = censdatasurvfinal, dist = "weibull", method = "Weibler-Mead", na.rm = TRUE, na.rm = TRUE)
  aic = censdatasurvfinal$fphv, na.rm = TRUE)
  if (fitcode$summary$matresult$[1]$< 0.05) {
    fitcode$summary$matresult$[1]$< TRUE
  } else {
    fitcode$summary$matresult$[1]$< FALSE
  }
  # else {
}
  if (fitcode$summary$matresult$[1]$< FALSE)
    distlist$[1]$ <- tryCatch(flexsurvreg(formula = Surv(time, cens) ~ 1, data = censdatasurvfinal$distlist$[1]$< 0.05)
  else {
    fitcode$summary$matresult$[1]$< FALSE
  }
  # else {
}

  # Fit distributions

  # if (length(na.remove(matresults$fphv)) > 0) {

  fit, AIC:

  ylab = AIC:

  par(mfrow = c(4, 4))

  flexsurv::plot.flexsurvreg(distlist$[1]$ <- tryCatch(flexsurvreg(formula = Surv(time, cens) ~ 1, data = censdatasurvfinal, dist = "weibull", method = "Weibler-Mead", na.rm = TRUE, na.rm = TRUE)
    aic$[1]$< Inf)
  distlist$[1]$ <- list()

  distlist$[1]$[1]$ <- tryCatch(flexsurvreg(formula = Surv(time, cens) ~ 1, data = censdatasurvfinal, dist = "weibull", method = "Weibler-Mead", na.rm = TRUE, na.rm = TRUE)
    aic$[1]$< Inf)
  distlist$[1]$[2]$ <- tryCatch(flexsurvreg(formula = Surv(time, cens) ~ 1, data = censdatasurvfinal, dist = "lnorm", method = "Weibler-Mead", na.rm = TRUE, na.rm = TRUE)
    aic$[1]$< Inf)
  distlist$[1]$[3]$ <- tryCatch(flexsurvreg(formula = Surv(time, cens) ~ 1, data = censdatasurvfinal, dist = "exp", method = "Weibler-Mead", na.rm = TRUE, na.rm = TRUE)
    aic$[1]$< Inf)

  plot(flexsurvreg(formula = Surv(time, cens) ~ 1, data = censdatasurvfinal, dist = "weibull", method = "Weibler-Mead", na.rm = TRUE, na.rm = TRUE)
    aic$[1]$< Inf)
  plot(flexsurvreg(formula = Surv(time, cens) ~ 1, data = censdatasurvfinal, dist = "lnorm", method = "Weibler-Mead", na.rm = TRUE, na.rm = TRUE)
    aic$[1]$< Inf)
  plot(flexsurvreg(formula = Surv(time, cens) ~ 1, data = censdatasurvfinal, dist = "exp", method = "Weibler-Mead", na.rm = TRUE, na.rm = TRUE)
    aic$[1]$< Inf)

defectcodesummary$shape[i]), add.line = TRUE)
defectcodesummary$gamma = distlist[[bestfitindex]]$res[i], scale =
defectcodesummary$shape[i]), add.line = TRUE)
defectcodesummary$shape[i]), add.line = TRUE)
defectcodesummary$shape[i]), add.line = TRUE)
defectcodesummary$gamma = distlist[[bestfitindex]]$res[i], scale =
defectcodesummary$shape[i]), add.line = TRUE)
defectcodesummary$gamma = distlist[[bestfitindex]]$res[i], scale =
defectcodesummary$gamma = distlist[[bestfitindex]]$res[i], scale =
defectcodesummary$shape[i]), add.line = TRUE)
defectcodesummary$shape[i]), add.line = TRUE)
defectcodesummary$shape[i]), add.line = TRUE)
defectcodesummary$shape[i]), add.line = TRUE)
defectcodesummary$shape[i]), add.line = TRUE)
defectcodesummary$shape[i]), add.line = TRUE)
defectcodesummary$shape[i]), add.line = TRUE)
defectcodesummary$shape[i]), add.line = TRUE)
defectcodesummary$shape[i]), add.line = TRUE)
defectcodesummary$shape[i]), add.line = TRUE)
defectcodesummary$shape[i]), add.line = TRUE)
defectcodesummary$shape[i]), add.line = TRUE)
defectcodesummary$shape[i]), add.line = TRUE)
defectcodesummary$shape[i]), add.line = TRUE)
defectcodesummary$shape[i]), add.line = TRUE)
defectcodesummary$shape[i]), add.line = TRUE)
defectcodesummary$shape[i]), add.line = TRUE)
defectcodesummary$shape[i]), add.line = TRUE)
defectcodesummary$shape[i]), add.line = TRUE)
defectcodesummary$shape[i]), add.line = TRUE)
defectcodesummary$shape[i]), add.line = TRUE)
defectcodesummary$shape[i]), add.line = TRUE)
defectcodesummary$shape[i]), add.line = TRUE)
defectcodesummary$shape[i]), add.line = TRUE)
defectcodesummary$shape[i]), add.line = TRUE)
defectcodesummary$shape[i]), add.line = TRUE)
defectcodesummary$shape[i]), add.line = TRUE)
defectcodesummary$shape[i]), add.line = TRUE)
defectcodesummary$shape[i]), add.line = TRUE)
defectcodesummary$shape[i]), add.line = TRUE)
defectcodesummary$shape[i]), add.line = TRUE)
defectcodesummary$shape[i]), add.line = TRUE)
defectcodesummary$shape[i]), add.line = TRUE)
defectcodesummary$shape[i]), add.line = TRUE)
defectcodesummary$shape[i]), add.line = TRUE)
defectcodesummary$shape[i]), add.line = TRUE)
defectcodesummary$shape[i]), add.line = TRUE)
defectcodesummary$shape[i]), add.line = TRUE)
defectcodesummary$shape[i]), add.line = TRUE)
defectcodesummary$shape[i]), add.line = TRUE)
defectcodesummary$shape[i]), add.line = TRUE)
defectcodesummary$shape[i]), add.line = TRUE)
defectcodesummary$shape[i]), add.line = TRUE)
defectcodesummary$shape[i]), add.line = TRUE)
defectcodesummary$shape[i]), add.line = TRUE)
defectcodesummary$shape[i]), add.line = TRUE)
defectcodesummary$shape[i]), add.line = TRUE)
defectcodesummary$shape[i]), add.line = TRUE)
defectcodesummary$shape[i]), add.line = TRUE)
defectcodesummary$shape[i]), add.line = TRUE)
defectcodesummary$shape[i]), add.line = TRUE)
defectcodesummary$shape[i]), add.line = TRUE)
defectcodesummary$shape[i]), add.line = TRUE)
defectcodesummary$shape[i]), add.line = TRUE)
defectcodesummary$shape[i]), add.line = TRUE)
defectcodesummary$shape[i]), add.line = TRUE)
defectcodesummary$shape[i]), add.line = TRUE)
defectcodesummary$shape[i]), add.line = TRUE)
defectcodesummary$shape[i]), add.line = TRUE)
defectcodesummary$shape[i]), add.line = TRUE)
defectcodesummary$shape[i]), add.line = TRUE)
defectcodesummary$shape[i]), add.line = TRUE)
defectcodesummary$shape[i]), add.line = TRUE)
defectcodesummary$shape[i]), add.line = TRUE)
defectcodesummary$shape[i]), add.line = TRUE)
defectcodesummary$shape[i]), add.line = TRUE)
defectcodesummary$shape[i]), add.line = TRUE)
defectcodesummary$shape[i]), add.line = TRUE)
defectcodesummary$shape[i]), add.line = TRUE)
defectcodesummary$shape[i]), add.line = TRUE)
defectcodesummary$shape[i]), add.line = TRUE)
defectcodesummary$shape[i]), add.line = TRUE)
defectcodesummary$shape[i]), add.line = TRUE)
defectcodesummary$shape[i]), add.line = TRUE)
defectcodesummary$shape[i]), add.line = TRUE)
defectcodesummary$shape[i]), add.line = TRUE)
defectcodesummary$shape[i]), add.line = TRUE)
defectcodesummary$shape[i]), add.line = TRUE)
defectcodesummary$shape[i]), add.line = TRUE)
defectcodesummary$shape[i]), add.line = TRUE)
defectcodesummary$shape[i]), add.line = TRUE)
defectcodesummary$shape[i]), add.line = TRUE)
defectcodesummary$shape[i]), add.line = TRUE)
defectcodesummary$shape[i]), add.line = TRUE)
defectcodesummary$shape[i]), add.line = TRUE)
defectcodesummary$shape[i]), add.line = TRUE)
defectcodesummary$shape[i]), add.line = TRUE)
defectcodesummary$shape[i]), add.line = TRUE)
defectcodesummary$shape[i]), add.line = TRUE)
defectcodesummary$shape[i]), add.line = TRUE)
defectcodesummary$shape[i]), add.line = TRUE)
defectcodesummary$shape[i]), add.line = TRUE)
defectcodesummary$shape[i]), add.line = TRUE)
defectcodesummary$shape[i]), add.line = TRUE)
defectcodesummary$shape[i]), add.line = TRUE)
defectcodesummary$shape[i]), add.line = TRUE)
defectcodesummary$shape[i]), add.line = TRUE)
defectcodesummary$shape[i]), add.line = TRUE)
defectcodesummary$shape[i]), add.line = TRUE)
defectcodesummary$shape[i]), add.line = TRUE)
defectcodesummary$shape[i]), add.line = TRUE)
defectcodesummary$shape[i]), add.line = TRUE)
defectcodesummary$shape[i]), add.line = TRUE)
defectcodesummary$shape[i]), add.line = TRUE)
defectcodesummary$shape[i]), add.line = TRUE)
defectcodesummary$scale[i] = distlist[[i]]$res[1]

defectcodesummary$FphVafter1year[i] = pexp(365, rate = distlist[[bestfitindex]]$res[1]) * 100

defectcodesummary$FphVafter2years[i] = pexp(730, rate = distlist[[bestfitindex]]$res[1]) * 100

defectcodesummary$FphVafter3years[i] = pexp(1096, rate = distlist[[bestfitindex]]$res[1]) * 100

defectcodesummary$FphVafter4years[i] = pexp(1461, rate = distlist[[bestfitindex]]$res[1]) * 100

defectcodesummary$FphVafter5years[i] = pexp(1826, rate = distlist[[bestfitindex]]$res[1]) * 100

defectcodesummary$FphVafter6years[i] = pexp(2191, rate = distlist[[bestfitindex]]$res[1]) * 100

defectcodesummary$FphVafter7years[i] = pexp(2557, rate = distlist[[bestfitindex]]$res[1]) * 100

defectcodesummary$FphVafter8years[i] = pexp(2922, rate = distlist[[bestfitindex]]$res[1]) * 100

#censdatasurvfinal <- dplyr::mutate(censdatasurvfinal, dead = ifelse(censdatasurvfinal$dead==0, TRUE, FALSE))

#EnvStats::qqPlotCensored(censdatasurvfinal$time, censdatasurvfinal$dead, censoring.side = "right", distribution = "exp", param.list = list(rate = defectcodesummary$scale[i]), add.line = TRUE)
defectcodesummary$expectedlifetime[i] = defectcodesummary$scale[i] ^ (-1)

allsummaries[[datayears]] <- defectcodesummary

year1 <- allsummaries[[1]]
year2 <- allsummaries[[1]]
year3 <- allsummaries[[1]]
year4 <- allsummaries[[1]]
year5 <- allsummaries[[1]]
year6 <- allsummaries[[1]]
year7 <- allsummaries[[1]]
year8 <- allsummaries[[1]]