Reliability Modelling for Aircraft Component Availability Management

O.W.M. Thijssens

TUDelft

Reliability modelling for aircraft component availability management

by

O.W.M. Thijssens

in partial fulfilment of the requirements for the degree of

Master of Science
in Aerospace Engineering,
specialisation Air Transport & Operations,

at the Delft University of Technology,
faculty of Aerospace Engineering,

to be defended publicly on Monday July 29, 2019 at 13:00.

Thesis committee:

Chair & supervisor	Dr. ir. W. J. C. Verhagen,	TU Delft – Air Transport & Operations
Supervisor	Drs. M. M. Bontenbal,	KLM N.V. – Supply Chain Management
Supervisor	Drs. T. D. Knappers,	KLM N.V. – Supply Chain Management
Assistant Professor	Dr. ir. B. F. Lopes dos Santos,	TU Delft – Air Transport & Operations
Assistant Professor	Ir. J. Sinke,	TU Delft – Aerospace Structures and Materials

This thesis is confidential and cannot be made public until July 29, 2024,
after which an electronic version will become available at `www.repository.tudelft.nl`*.*

Preface

The past eight years have truly been an adventure. From being a fanatic race rower, dreaming that I would one day – perhaps next year – race for an Olympic medal, via one semester of airline operation courses in the home city of Airbus headquarters, a one year full-time responsibility of running a student employment agency, starting a masters in Air Transport and Operations in Delft and a masters in Econometrics & Operations Research in Tilburg, improving revenue management models for the flag carrier of the beautiful country of Australia, to where I am right now: a few days left until I hand in my master thesis – the product of a nine-month research project performed at KLM Royal Dutch Airlines – concluding the end of my Aerospace Engineering study at Delft University of Technology. I would like to take some time to express my gratitude to everyone that contributed to this exciting, life-changing journey.

First and foremost, I would like to thank the daily supervisors of this thesis – Wim, Thomas and Martijn – for their guidance throughout this journey. Thomas and Martijn, I will never forget the interview I had with you for this position: via a Skype call from a room at Qantas HQ with a lagging internet connection, failing camera and eight hours of time-zone difference. Thank you for making possible this collaboration between university and industry and handing me this opportunity. Wim, your kindness, patience and critical glance have greatly contributed to the work that is presented. I would also like to thank my two other assessment committee members – Bruno and Jos – for their time to read through this extensive document, and attending my presentation.

Furthermore, I would like to thank Isa, Iris, Jochem, Abe, Jesper, Mark, Jolanda, Gerda, Sterre, Michel, and all other people at KLM that I interacted with – either discussing content or talking about something totally irrelevant to my research. You made the numerous hours at Schiphol-East a very educational and fun time. Special thanks to Bas, Arnold and René, for all the jokes, coffee, and the back and forth football club related teasing.

Then I would like to thank all of my friends for being the awesome people that your are. The adventures I experienced would not have been worthwhile without any of you. Especially Jonghe Acht 136, Huize Duizend, Huize Arnold, StuD18, Laga Onzin, AitAitAit, Qantas co-interns, Chatham and Siri – you made my student life as brilliant as possible.
Finally, I want to thank my family: my parents Wim and Reggy, and my brothers Boyd, Pepijn and Benjamin. Thank you for supporting me in everything I do.

This adventure comes to an end, and new ones lie ahead. Thank you all for being part of this.

Olivier Thijssens,
Delft, 12 July 2019

Contents

List of Figures

List of Tables

Introduction

The global and local aviation traffic is growing while economic and performance pressures on the industry are increasing. This results in airlines maximising their fleet utilization. Therefore, airline operators and Maintenance, Repair and Overhaul (MRO) providers require as much insight as possible in factors affecting component reliability and availability. Reliability analysis in literature is generally limited to statistical models as functions of time only, including strict assumptions on independence of events and underlying distributions. This foregoes the complex nature of aircraft operations, where various operational and maintenance factors may influence the probability of occurrence of a failure. In this research new insights from operational and maintenance data about the impact of operating environment and ageing of components and fleet on reliability of the components will be developed. The main research question is formulated as:

How can operational and maintenance data be leveraged for reliability modelling of aircraft repairables?

The structure of this thesis will be as follows, The main theory, methodology, results and conclusions of the research work are provided in a scientific paper beginning on page 1. A proper scientific paper demands clear and concise information provision. Appendices are supplemented for a more extensive elaboration on specific sections of the performed research. Appendices A and B describe the initial project plan and literature study performed prior to the actual research work. Appendix C covers the data preprocessing steps undertaken, which were essential for the modelling phase. In Appendix D the selection procedure regarding components included in the research is explained. Due to the vast number of different components, the scope of the evaluated components had to be narrowed. Appendix E will elaborate on the hypothesised interrelations between variables from operational and maintenance data and the reliability of the component. Finally, Appendix F provides a thorough explanation of all the reliability modelling steps which were needed in order to obtain the most optimal model from the available data.

Paper

Application of Extended Cox Regression Model to Time-On-Wing Data of Aircraft Repairables.

O.W.M. Thijssens (MSc. Student)
Supervisors: dr. ir. W.J.C. Verhagen, drs. T.D. Knappers, drs. M.M. Bontenbal
Section Air Transport & Operations, Department Control and Operations, Faculty of Aerospace Engineering, Delft University of Technology, Delft, The Netherlands

Abstract— **Global and local aviation traffic is growing while economic and performance pressures on the industry are increasing. As a consequence, airlines try to maximise their fleet utilization. Airline operators and Maintenance, Repair and Overhaul (MRO) providers therefore require as much insight as possible in factors affecting component reliability and availability. Reliability analysis in literature rarely considers the existence of a relation between explanatory variables and component reliability, and includes strict assumptions on independence of events and underlying distributions. This disregards the complex nature of aircraft operations, where the probability of an event may be influenced by various operational and maintenance factors. This paper develops new insights from operational and maintenance data about the impact of operating environment and ageing of components and fleet on reliability of the components by incorporating these factors in an extension of the Cox regression model. Examination of results obtained from analysing historical data of a set of three components with respect to installations and removals indicate that the natural environment at the hub airport, maintenance history of components, the age of the aircraft on which the component is installed and different modification designs are useful significant predictors of the time-on-wing duration of the component.**

I. Introduction

This global and local aviation traffic is growing while economic and performance pressures on the industry are increasing [1]. This results in airlines maximising their fleet utilization. Maintenance costs can contribute considerably to the spending of an airline; historical maintenance cost estimates range from 10% to 15% of total airline spending. of the overall expenditure incurred by airlines [2]. 22% of the maintenance cost comprises component maintenance cost.

An important factor in decreasing component maintenance cost is economies of scale for component availability. The aim of component availability service is to maximise the use of aircraft by maintaining spare units ready for installation whenever necessary [3]. Naturally, spare needs are random because a random occurrence, aircraft component failure, initiates them. By the law of large numbers, variation in demand reduces for an increase in the amount of random events. This phenomenon is depicted in Fig 1, which shows that for a specific aircraft model, the need for spare components per aircraft decreases with an increase in number of aircraft in the fleet. The direct consequence is lower marginal expenses for storage, capital intensity of stock, and costs of obsolescence.

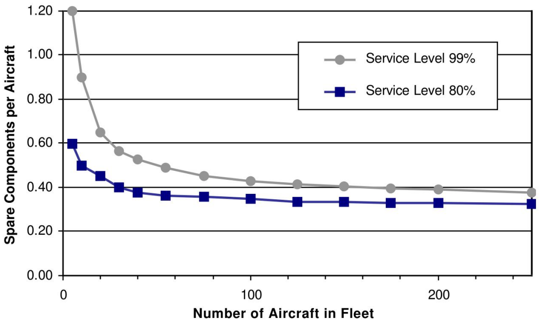


Fig. 1: Economies of scale in component availability service [3]

Apart from standardising fleet composition, a popular way to exploit the scale of economies regarding availability services is a combination of subcontracting the availability service and inventory pooling. Subcontracting component availability replaces capital expenses with a steady cash flow, improving business flexibility. The airline customer pays an MRO service provider in proportion to the number of aircraft flying hours (so called 'Power-by-the-Hour' [4]). A service level agreement dictates the numerous service-performance metrics with corresponding service-level objectives. Inventory pooling between airlines on the other hand means that from a single stock of inventory different airline markets are served, each with their own uncertain demand. Consequentially, components from the pool are used in different (natural) environments and in aircraft varying in age. Such factors could affect the reliability characteristics of the equipment. It is therefore desirable for MRO service providers to quantify the effects of these factors in order to determine the right maintenance price per flight-hour, as well as to help to identify and plan for maintenance events.

Commonly used models for reliability modelling of a component, like the homogeneous Poisson process (HPP) and the renewal process (RP), consider the time to failure variable as the only variable of interest [5]. These models neglect the operational factors that can influence the time to failure of a system. For instance, aircraft operating from hot, sandy regions are affected differently than aircraft operating from humid airports, which could result in distinct failure modes and times for installed components [6], [7]. Thus, for a better estimate of reliability characteristics, the use of regression models is suggested because of the possibility of including covariates [8]. Very popular and well-known regression models often used for reliability analysis are from the proportional hazards family. The proportional hazards model (PHM) was introduced by Cox [9] and created a great deal of interest, though primarily in the field of bio-statistics. Recently, the interest in reliability engineering applications of PHM has risen due to the ability to process reliability data without making prior assumptions about the hazard rate's functional form. Nonetheless, applications in industrial cases appear scarce in reliability literature, especially when time-dependent covariates with recurrent events are present [10]. Furthermore, the assumptions underlying its use are rarely verified, while inappropriate use of this model may lead to biased results, inaccurate risk prediction, and reduced statistical power [11]. Additionally, external validation is not straightforward and is seldom considered for a Cox model, while for it to be useful in practice the model should perform satisfactorily in an external sample [12]. Even more, the hazard ratio estimate is almost routinely used to quantify an covariate effect. However, the interpretation of this ratio is not so straightforward, especially when the proportional hazards assumption is violated [13]. A much more widely used measure of reliability in the aviation industry is the mean time between (unscheduled) removals (MTBR), and therefore expressing the effect of a covariate on the MTBR might be more valuable [3].

This research aims to improve statistical reliability assessment of aircraft components by incorporating operational and maintenance factors. An observational study is performed where historical operational and maintenance data is analysed to identify factors with a measurable influence on the time-on-wing of aircraft repairables. The use of restricted mean survival time ratios is demonstrated as an alternative measure of the covariate effects.

The structure of this paper reflects this focus. In Section II, a theoretical background in survival analysis and the important Cox regression model is given. In Section III, the study design and modelling approach is given, including a discussion of the component scope and data sources used. Section IV provides results for a set of selected components. Finally, Section V discusses the findings and describes the significance of these findings in light of what was already known about the research problem. It also describes the limitations of the performed research.

II. Theoretical Background

Here follows a discussion of the statistical approaches and underlying theory of survival analysis. For a more detailed explanation, the reader is referred to text books of Kalbfleisch and Prentice [14], Klein and Moeschberger [15], Cook and Lawless [16], Kleinbaum and Klein [17] and Therneau and Grambsch [18].

A. Survival Analysis

Survival analysis, or time-to-event analysis, is the study of survival times and of the factors that influence them. The problem of analysing time to event data arises in a number of applied fields, such as medicine, biology, public health, epidemiology, engineering, economics, and demography [15]. A key characteristic of survival data is that the target variable is a non-negative, often continuous, random variable, and represents the time from a well-defined origin to a well-defined event. A second characteristic of survival analysis is censoring, and occurs when the starting or ending events of some subjects are not precisely observed. A key benefit of survival analysis models is that the censored observations are still used for modelling the target variables, as opposed to a logistic or Poisson regression model, and therefore more information from the data is extracted. Furthermore, survival analysis models permit changing covariate values over time, resulting in greater efficiency and accuracy [19]. The most occurring type of censoring in survival data is right-censoring, meaning that the true survival time interval has been cut off at the right side of the observed time interval, resulting in an observed survival time that is shorter than the true survival time. An example is when the experiment, study or operation is stopped at a predetermined time. The theory and application of other censoring types (e.g. random, left, interval) is readily available, however in context to this paper only right censoring will be considered. Figure 2 shows a simplified overview of the survival analysis study performed.

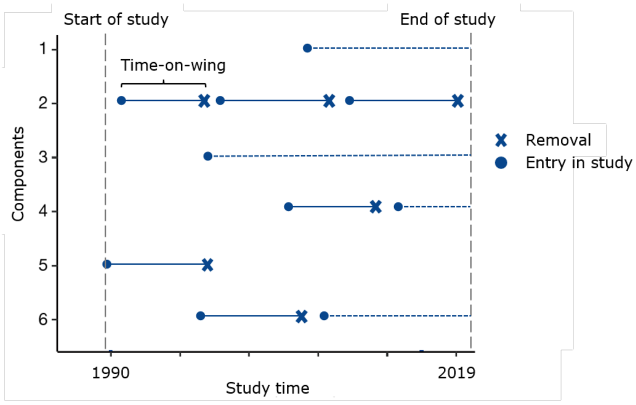


Fig. 2: Simplified overview of survival analysis study specific to this research. The horizontal dashed lines indicate the right-censored observations.

Survival analysis methods depend on the survival distribution, and a key way of specifying it is the survival function. The survival function defines the probability of a subject surviving beyond some specified time t,

$$S(t) = P(T > t) \tag{1}$$

The random variable T in (1) indicates the survival time and t any specific value of interest for the variable T. The survival time scale is often chosen to be calendar time, especially with processes that apply to humans or animals. In technological areas, measures of usage or exposure are often used, e.g. distance, number of cycles etc.. The survivor function is graphed as a decreasing smooth curve, which begins at $S(t) = 1$ at $t = 0$ and heads downward toward zero as t increases toward infinity. Another way to describe the distribution of T is given by the hazard function, which is the conditional probability of failure in the next instant give survival up to a point in time.

$$h(t) = \lim_{\Delta t \to 0} \frac{P(t \leq T < t + \Delta t | T \geq t)}{\Delta t} \tag{2}$$

In literature the hazard function is also referred to as the instantaneous failure rate. There is a clearly defined relationship between *S(t)* and *h(t)*, as can be mathematically expressed by:

$$S(t) = exp\left[-\int_0^t h(u)du\right] \tag{3}$$

$$h(t) = -\left[\frac{dS(t)/dt}{S(t)}\right] \tag{4}$$

The mean time to event or mean survival time, μ, can now be introduced, which is the expected value of T:

$$\begin{aligned}\mu = E(T) &= \int_0^\infty tf(t)dt \\ &= \int_0^\infty \left(\int_0^t ds\right) f(t)dt = \int_0^\infty \left(\int_s^\infty f(t)dt\right) ds \\ &= \int_0^\infty S(t)dt \end{aligned} \tag{5}$$

Thus from (5) it is clear that the mean survival time equals the area under the survival curve, which can be calculated via the trapezoidal or Simpson's rule. If however the longest survival time corresponds with a censored observation, the survival curve will not drop to zero and $S(\infty)$ is undefined. A work-around that is becoming popular in literature is to specify a maximum possible survival time in order to make the integral finite. This restricted mean survival time (RMST) is thus a measure of average survival from time zero to a specified time point. In case of sufficient observations and the survival function near zero toward the end of the survival period, the restricted mean will be close in magnitude to the overall mean.

To assess the relationship of explanatory variables to survival time usually requires some form of regression model. The most widely applied regression model is the Cox proportional hazards model.

B. Cox Proportional Hazard Model

The Cox proportional hazards model is usually written in terms of the hazard model formula shown in (6). This model gives an expression for the hazard at time t for a subject with a given specification of a set of covariates denoted by the bold **x**. β is the unknown parameter of the model, defining the effects of the covariates.

$$h(t|\mathbf{x}) = h_0(t)exp(\beta^T\mathbf{x}) \tag{6}$$

The Cox model formula states that the hazard at time t is the product of two quantities. The first of these, $h_0(t) = h(t|\mathbf{0})$, refers to a common baseline hazard function, the second quantity is the exponential expression which include the explanatory variables. The baseline hazard rate is assumed to be identical and equal to the total hazard rate when the covariates have no influence on the survival time. The covariates may influence the hazard rate so that the observed hazard rate is either greater (e.g. in the case of poor maintenance) or smaller (e.g. a new or improved component of a system) compared to the baseline hazard rate. An important feature of this formula, which concerns the proportional hazards (PH) assumption, is that the baseline hazard is a function of t, but does not involve the explanatory variables. The unique effect of a unit increase in a covariate, also known as the hazard ratio (HR), is multiplicative with respect to the hazard rate and constant over time, ceteris paribus. As mathematically shown in (7) for a specific covariate x_i and its new value x_i^*, this is because the baseline hazard cancels out resulting in an expression which does not involve time t.

$$HR = \frac{h(t|x_i^*)}{h(t|x_i)} = \frac{h_0(t)exp(\beta_i x_i^*)}{h_0(t)exp(\beta_i x_i)} = exp\left(\beta_i(x_i^* - x_i)\right) \tag{7}$$

Another important property of the Cox model is that the baseline hazard, $h_0(t)$, is an unspecified function. It is this semi-parametric property that makes the Cox model so popular, as reasonably good estimates of regression coefficients and adjusted survival curves can be obtained for a wide variety of data situations, even though the baseline hazard is not specified or known. It is a robust model, so that the results from using the Cox model will closely approximate the results for the correct parametric model [17].

Even though the Cox PHM is less restrictive than a full parametric model, it has still various underlying assumptions which have to be valid. First of all, it relies on the assumption of independent censoring for valid inference in the presence of right-censored data, i.e. censored subjects are not at increased risk for failure. Furthermore, an obvious assumption for most regression models, including the Cox PHM, is that the observations are independent. In case of multiple observations per subject, i.e. recurrent event analysis, this is however an invalid assumption. Finally, the most important assumption it has to satisfy is the one of proportional hazards.

C. Maximum Likelihood Estimation

The coefficient estimates of the Cox model parameters are derived by maximising a likelihood function, which is a mathematical expression describing the joint probability of obtaining the data actually observed on the subjects in the study as a function of the unknown parameters, β. The term partial likelihood is used because the likelihood formula considers probabilities only for those subjects experiencing the event, and does not explicitly consider probabilities for censored observations. This partial likelihood (L) can be written as the product of several likelihoods, one for each event time. As in the HR formula in (7), the baseline hazard cancels out of the numerator and denominator:

$$\begin{aligned} L(\beta) &= \prod_{j=1}^{d} \frac{h(t_{(j)})}{\sum_{k \in R(t_{(j)})} h(t_k)} = \prod_{j=1}^{d} \frac{h_0(t_{(j)})exp(\beta^T\mathbf{x}_{(j)})}{\sum_{k \in R(t_{(j)})} h_0(t_{(j)})exp(\beta^T\mathbf{x}_k)} \\ &= \prod_{j=1}^{d} \frac{exp(\beta^T\mathbf{x}_{(j)})}{\sum_{k \in R(t_{(j)})} exp(\beta^T\mathbf{x}_k)} \end{aligned} \tag{8}$$

where d is the observed number of events, $t_{(j)}$ is the jth ordered event time, $h(t_{(j)})$ is the hazard function for the jth ordered event, $\mathbf{x}_{(j)}$ is the covariate vector of the subject with an event at time t_j, and $R(t_{(j)})$ is the risk set [9]. The risk set is the collection of subjects which are still at risk of experiencing the event; their survival time is equal or longer than $t_{(j)}$. A remarkable characteristic of the partial likelihood function in (8) is that only the order of the event times matters; the particular values of the event times do no contribute to the partial likelihood.

$$l(\beta) = \sum_{j=1}^{d} \left[\beta^T \mathbf{x}_{(j)} - \log\left(\sum_{k\in R(t_{(j)})} exp(\beta^T \mathbf{x}_k) \right) \right] \quad (9)$$

The maximisation process is carried out by taking partial derivatives of l with respect to each parameter in the model, and then solving a system of equations. The score function, which is the first derivative of $l(\beta)$, has p components, one for each of the p covariates [20]. The l'th component is given by

$$S_l(\beta) = \frac{\partial l(\beta)}{\partial \beta_l} = \sum_{j=1}^{d} \left[x_{(j)l} - \frac{\sum_{k\in R(t_{(j)})} x_{(j)k} exp(\beta^T \mathbf{x}_k)}{\sum_{k\in R(t_{(j)})} exp(\beta^T \mathbf{x}_k)} \right] \quad (10)$$

where $x_{(j)l} = \partial \beta^T \mathbf{x}_{(j)} / \partial \beta_l$. This function can be interpreted as the sum of residuals, each of which consists of the observed value $x_{(j)l}$ of the covariate minus an expected value.

D. Recurrent Event Survival Analysis

In many research scenarios a subject may experience an event several times over the study time. Examples of recurrent event data include recurrence of bladder cancer tumours, recurrent failures of a system or the advent of economic recession [21], [22]. By considering all events instead of solely the first event, less number of subjects need to be included in the study for reaching an acceptable power of demonstrating a relevant effect. Modelling recurrent event data can be carried out using a Cox PH model with the data layout constructed so that each subject has a line of data corresponding to each recurrent event. A key decision for recurrent event analysis is defining when a subject is at risk of having an event along a given time scale as this directly impacts the likelihood estimation in (8) via the risk set $R(t_{(j)})$. Three main options for this risk interval formulation exist: gap time, total time and counting process formulation [23]. Gap time approach would be preferred if the time interval of interest is the time from the previous event to the next recurrent event rather than time from study entry until each recurrent event in case of total time. The counting process formulation has the same time scale as total time, but recurrent events of a subject are not considered to be at risk before the previous event has happened. This means it models one survival distribution per subject for the whole study time.

In most cases, for a specific subject an event will influence the survival time to the next recurrent event. One way to include this relationship is by introducing covariates that track the number of previous events in the Cox model. These covariates do however need to satisfy the PH assumption. It also makes sense to view the different intervals contributed by a given subject as representing correlated observations on the same subject that must be accounted for in the analysis. Genetic or inherent product factors mean that survival times of a subject are more similar to each other than those from other clusters. This within-subject correlations is generally positive, meaning that the true degree of variability will be underestimated, and may result in inadequate rejection of the null hypothesis [19]. A widely used technique for adjusting for the correlation among outcomes on the same subject is called robust estimation. This technique involves adjusting the estimated variances of regression coefficients obtained for a fitted model to account for misspecification of the correlation structure assumed [24], [25]. Note that the estimated regression coefficients themselves are not adjusted, but only the estimated variances of these coefficients. The general form of this estimator can be most conveniently written in matrix notation as:

$$\mathbf{Var}^*(\hat{\beta}) = \widehat{\mathbf{Var}}(\hat{\beta}) \left[\hat{\mathbf{R}}_S^T \hat{\mathbf{R}}_S \right] \widehat{\mathbf{Var}}(\hat{\beta}) \quad (11)$$

where $\widehat{\mathbf{Var}}(\hat{\beta})$ denotes the variance matrix for $\hat{\beta}$, $\mathbf{Var}^*(\hat{\beta})$ denotes the cluster-adjusted variance matrix, and $\hat{\mathbf{R}}_S$ denotes the matrix of score residuals obtained from partial maximum likelihood estimation of the Cox model being fit [17]. The adjusted standard errors of the parameter estimate $\hat{\beta}$ are the square root of the diagonal elements of $\mathbf{Var}^*(\hat{\beta})$.

E. Extended Cox Model

The Cox proportional hazards model can be extended to allow time-dependent variables as predictors. To accommodate covariates that may change their value over time, special measures are necessary to obtain valid parameter estimates; each subjects survival time has to be subdivided into smaller time intervals to allow for changes in values of time-varying covariates. The extended Cox model is mathematically expressed as

$$h(t|\mathbf{x}(t)) = h_0(t) exp(\beta^T \mathbf{x} + \delta^T \mathbf{x}(t)) \quad (12)$$

where $\mathbf{x}(t)$ are the time-dependent predictors. Due to the time-dependent predictors, the proportional hazards assumption is no longer satisfied as the hazard ratio is now a function of time:

$$\begin{aligned} HR(t) &= \frac{h(t|\mathbf{x}^*(t))}{h(t|\mathbf{x}(t))} = \frac{h_0(t) exp(\beta^T x^*(t))}{h_0(t) exp(\beta^T x(t))} && (13) \\ &= exp\left[\left(\beta^T (\mathbf{x}^* - \mathbf{x}) \right) + \left(\delta^T (\mathbf{x}^*(t) - \mathbf{x}(t)) \right) \right] && (14) \end{aligned}$$

The partial likelihood function will be still similar to (8), except that the contributions of each subject in the risk set can change from one event time to the next.

F. Assessing the Model Adequacy

Before the Cox model results can be used, it is important to optimise the model and verify the assumptions. This encompasses three main points: covariate selection, checking the functional form of continuous covariates, and validating the proportional hazard assumptions.

1) Covariate selection: First, it is important to test which covariates should be included in the regression model. One popular hypothesis test is the likelihood ratio test, which is a test of $H_0 : \beta = 0$ for a certain covariate. Even though the Cox model only gives a partial likelihood, it is possible to compare nested Cox models using a likelihood ratio test [26]. In nested models the covariates of one model are a subset of the covariates in the other. The likelihood ratio test uses the result from statistical theory that $2\left[l(\beta = \hat{\beta}) - l(\beta = 0)\right]$ follows approximately a chi-square distribution with one degree of freedom.

Another manner of optimising the model makes use of the Akaike's Information Criterion (AIC). The AIC is depended on likelihood as well, but also on the amount of degrees of freedom used by the model. A benefit of using AIC is the ability of comparing unnested models made for the same outcome on the same data due to its correction for the amount of parameters used. It is defined as

$$AIC = 2k - 2l(\hat{\beta}) \tag{15}$$

where k is the amount of degrees of freedom used and $l(\hat{\beta})$ the partial likelihood of the model at the maximum partial likelihood estimation. The value of the AIC balances two quantities: the goodness of fit term, which quantity is smaller for models that fit the data well, and a penalty term for the number of parameters as a measure of complexity. Smaller values of AIC should in theory indicate better models. Generally, it is computationally impractical to compute the AIC for all possible combinations of covariates. An alternative is to use a stepwise procedure which at each step tests if a covariates should be added or deleted. However, blindly selecting the model with the smallest AIC is not advised, as sometimes there will be good scientific or practical reasons for preferring one model to another. The AIC solely provides an objective evaluation of the model given the current data [27]. An alternative to the AIC is the Bayesian Information Criterion (BIC). The key difference is that the BIC penalises the number of parameters by a factor of $\log(n)$ rather than by a factor of 2 as in the AIC. As a result, using the BIC in model selection will tend to result in models with fewer parameters as compared to AIC.

2) Functional form of continuous variables: One assumption of the Cox proportional hazards model is that additive changes in values of covariates are assumed to have constant multiplicative effects on the hazard rate. However, it is quite possible that the hazard rate and the covariates do not have such a log-linear relationship, but e.g. log-quadratic. Furthermore, it is not always possible to know a priori the correct functional form that describes the relationship between a covariate and the hazard rate. One method to test the functional form of a continuous covariate is by categorising the covariate in non-overlapping intervals, e.g. by quantiles. The Cox model is fit with dummy variables for each category instead of the continuous covariate itself. A plot of the β-estimates by mean covariate value of the intervals, with β=0 for the reference category, should lie on a straight line for the log-linear assumption to be satisfied.

3) Validation of PH assumption: A number of different tests for assessing the PH assumption have been proposed in the literature. A popular test makes use of scaled Schoenfeld residuals and is proposed by Gramsch and Therneau [28], a variation of a test originally proposed by Schoenfeld [29]. For each covariate in the model, Schoenfeld residuals are defined for every subject who has an event. The Schoenfeld residuals are the individual terms of the score function as given in (10). In case of one covariate, each term is the observed value of that covariate for event j minus the expected value, which is a weighted sum, with weights given by $p(\beta, x_k)$.

$$\hat{r}_{(j)} = x_{(j)} - \sum_{k \in R(t_{(j)})} x_k \cdot p(\beta, x_k) \tag{16}$$

where

$$p(\beta, x_k) = \frac{exp(\beta x_k)}{\sum_{m \in R(t_{(j)})} exp(\beta x_m)} \tag{17}$$

Note that these residuals are defined only for the event time, not for censoring times. If there are multiple covariates, then one obtains a series of residuals for each of the p covariates. Each residual is scaled by an estimate of its variance, which is approximated via

$$\hat{r}^*_{(j)} = \hat{r}_{(j)} \cdot d \cdot var(\hat{\beta}) \tag{18}$$

A plot of theses residuals versus the covariate $x_{(j)}$ will yield a pattern of points that are centered at zero, if the proportional hazards assumption is correct. An approximate estimate of the coefficient of a covariate over time, $\beta(t)$, can be calculated by adding the estimate $\hat{\beta}$ to the standardised residuals [28]:

$$\hat{\beta}(t) \approx \hat{\beta} + E(r^*_{(j)}) \tag{19}$$

Plotting $\hat{\beta}(t)$ against time enables detection of departures from the PH assumption. A statistical test can be performed by fitting a straight line trough the residuals and testing for a significant slope coefficient, leading to a more objective approach compared to the subjectivity interpreting graphs.

Another test regarding the PH assumption makes use of time-dependent variables. For this test, the Cox model is extended to contain product terms involving the time-independent covariate being assessed and some function of time $g(t)$.

$$h(t|\mathbf{x}) = h_0(t) exp\left(\beta^T \mathbf{x} + \delta(x \times g(t))\right) \tag{20}$$

The PH assumption is tested by testing for the significance of the product term δ, meaning the null hypothesis equals $H_0 : \delta = 0$. The test can be carried out using a likelihood ratio statistic, and the test statistic has a chi-square distribution with one degree of freedom under the null hypothesis. A drawback of the use of time-dependent variables for assessing the PH assumption is that different choices of the functions $g(t)$ may result in different conclusions about whether the PH assumption is satisfied.

G. Stratification

The stratified Cox model is a modification of the Cox proportional hazards model that allows for control by stratification of a covariate that does not satisfy the PH assumption. Stratification can only be performed for discrete-valued covariates. Covariates that are assumed to satisfy the PH assumption are included in the model, whereas the covariate being stratified is not included. The hazard function formula will now contain a subscript g that indicates the g'th stratum.

$$h_g(t|\mathbf{x}) = h_{0_g}(t) exp\left(\beta^T \mathbf{x}\right) \tag{21}$$

Because stratified variables are not included in the model, it is not possible to obtain a hazard ratio value for the effect of those variables. The assumption is that the covariates affect all strata equally, and therefore the hazard ratios are also identical for different strata. This assumption can be tested via interaction terms between covariates and strata; insignificance of estimates means the assumption is valid.

H. Measure of Effect

In survival analysis, the measure of effect of a covariate on the survival time is typically expressed by the hazard ratio. An increasingly popular alternative to the hazard ratio is the ratio of restricted mean survival times between different values of a covariate [30], [31]. In order to compute these mean survival times when a Cox model is used, first survival curves need to be obtained that adjust for the explanatory variables used as covariates. These adjusted survival curves can be computed via

$$S(t|\mathbf{x}) = [\hat{S}_0(t)]^{exp(\hat{\beta}^T \mathbf{x})} \tag{22}$$

where $\hat{S}_0(t)$ indicates the predicted baseline survival function. Typically, when computing adjusted survival curves, the value chosen for a covariate being adjusted is an average value like an arithmetic mean or a median. Most computer programs for the Cox model automatically use the mean value over all subjects for each covariate being adjusted. An issue with this is that for categorical covariates it is not clear what a mean value, e.g. ”0.6 male”, represents. A solution is the use of a hypothetical group of subjects for which predicted survival curves are produced. By taking the average over these curves, instead of over individual covariate values, a proper average survival curve can be established [18]. The hypothetical group of subject can be based on empirical distribution of used data, a distribution from some external study or standard, or a factorial distribution in case of a model solely containing categorical variables. Computing standard errors or confidence intervals for the RMST-ratio is challenging, as the predicted survival curves for subjects in the hypothetical group are correlated due to their common dependence on the model's coefficient vector β. Even though standard errors for the RMST for a specific subject can be derived from the Cox model fit, deriving the distribution of the ratio of two correlated RMST values is complicated. One feasible solution is using the bootstrap method, a resampling method which independently samples with replacement from an existing sample data and performs inference among these resampled data [32]. In this context, it means fitting the Cox model on hundreds of samples from the original survival data, and using those different fits to compute the RMST-ratio. A histogram of all RMST-ratios will approximate the real distribution of this ratio.

I. External validation

An important extension to assessing model fit on a given dataset is external validation. This entails evaluating the performance of a model in a sample independent of that used to develop the model, and consists of discrimination and calibration. Discrimination is the extent to which risk estimates from a model characterise different subject prognoses. Subjects predicted to be at higher risk should exhibit higher event rates than those deemed at lower risk. Calibration refers to the predictive accuracy of survival probabilities. A well-calibrated risk score or prediction rule assigns the correct event probability at all levels of predicted risk. Assessing calibration of Cox models is tricky, because, apart from difficulties induces by censored observations, the Cox model estimates event probabilities indirectly and only relative to an unspecified baseline survival function. One approach to assess discrimination and calibration for a Cox model well described by Royston [12], [33] is to compare observed and predicted survival probabilities from a Cox model in several prognostic groups derived by placing cut points on the prognostic index, $\hat{\beta}^T\mathbf{x}$. First, it necessary to assess whether the validation data can be described by a similar probabilistic data generating mechanism as the training data. This can be done graphically, by displaying histograms of the prognostic index for both sets and comparing their distributions. Second, individual predicted survival functions are computed from the Cox model which was fit on the derivation set via (22). Third, for a given risk group with subject indices belonging to a set G, the individual survival functions, $\{\hat{S}(t|\mathbf{x}_i)\}_{i\in G}$, are averaged over the risk group at the observed event or censoring times. Fourth, the observed survival probabilities of those same risk groups are estimated via the nonparametric Kaplan-Meier method, which takes the product over the failure times of the conditional probabilities of surviving to the next failure time:

$$\hat{S}(t) = \prod_{t_i \leq t} \left(1 - \frac{d_i}{n_i}\right) \tag{23}$$

where n_i is the number of subjects at risk at time t_i and d_i is the number of subjects who experience the event at time t_i. Finally, a graphical comparison is performed based on plots of the predicted and observed survival curves against survival time within each risk group. The plot of Kaplan-Meier curves by group indicates the discrimination available with the model and the appearance of the survival curves. The risk-group predicted survival curves should show the same separation and should closely follow the observed estimated curves in order for the Cox model to be successfully validated. Generally, three to five risk groups are created. With a larger number of groups, the survival curves may be

TABLE I: General information on data of selected B737 components

Name	ATA chapter	Fleet size	QPA	Censored	Recurrent	MTBR (FH)	Inventory Value
ACM	Air Conditioning (21)	320	2	22%	39%	37,800	$425,000
ADIRU	Navigation (34)	465	2	16%	47%	18,000	$450,000
Display Unit	Indicating/Recording System (31)	404	6	21%	30%	80,000	$370,000

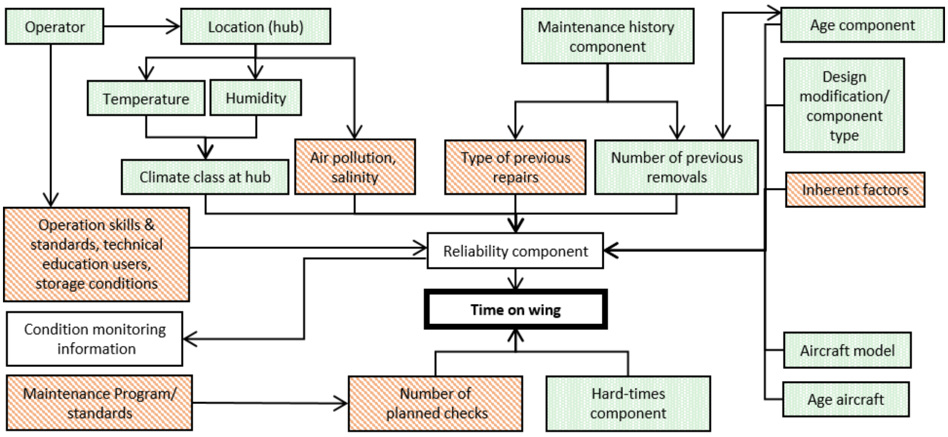


Fig. 3: Predictors for time-on-wing target variable, and their interrelation. The green (full) fields indicate included predictors, the red (downward diagonal) fields indicate covariates which are not controlled for.

unstable and the discrimination between neighbouring groups is likely to be poor [12].

III. Materials and Methods

A. Study design

This study is performed in collaboration with KLM Engineering & Maintenance (KLM E&M), a Dutch MRO company. In this observational study the author retrospectively reviewed data of all B737 aircraft which were in contract with KLM E&M between January 1995 and March 2019 and had a removal of the considered component. Over this time frame the number of aircraft in contract varied, as new customers were introduced and old aircraft were phased out. Depending on the component and moment in time, around 5%-10% of all B737 worldwide were in contract with KLM E&M [34]. Furthermore, these aircraft are based throughout the world, indicating a representative cross-section of the entire B737 fleet worldwide.

B. Data

Three main data sets were used for this study. The maintenance information is obtained from a large dataset of installation and removal data of components of the B737, logged for maintenance administration by KLM E&M. It contains information on airline, registration number of aircraft in which the component has been installed, the aircraft model, identification codes of parts, installation and removal dates, installation and removal age aircraft as indicated by Time-Since-New (TSN), manufacturing date of the part and specification on whether a removal was scheduled or unscheduled. A Köppen climate classification data set was used for information on the Köppen climate type per geographical location grid box of size 0.5° longitude times 0.5° latitude [35]. Lastly, an airport dataset from the OpenFlights Airports Database was utilised, containing information on location of airports (city, country and geographic coordinates) as well as IATA and ICAO airport codes [36]. For a complete description of the data preprocessing steps taken, the reader is referred to Appendix C.

C. Component scope

Based on data availability, maintenance cost and component type, three components are selected for this reliability study: the Air Cycle Machine (ACM), the Air Data Inertial Reference Unit (ADIRU) and the Display Unit. These components are regarded as self-contained units and are not analysed in terms of the functioning of their constituents. General characteristics of these components, including descriptive numbers on their maintenance data records, are presented in Table I. For a comprehensive description of the scoping process, the reader is referred to Appendix D.

D. Statistical Method

The data was randomly split up into a derivation (train) set and validation set containing 80% and 20% of the total data respectively. A Cox regression model was fit to the derivation data to assess effects of covariates on component reliability. Potential prognostic factors, including different modification designs, part-numbers and age of the components, number of previous removals, climate at hub airport and age of aircraft were selected according to data availability and current knowledge about the risks of failure. Climate at hub was categorised to desert, humid and temperate climate, were the latter was selected as reference level to which the others were compared. For component modification design or part numbers, the reference level was taken as the first

TABLE II: Information on covariates values and distributions for the three selected components

Characteristics of components at	**ACM**		**ADIRU**		**Display unit**	
moment of installation	*Derivation*	*Validation*	*Derivation*	*Validation*	*Derivation*	*Validation*
Mean (sd) age component in flight hours	7947 (13255)	7071 (12648)	9239 (14639)	8754 (14301)	5241 (11105)	3891 (9775)
Mean (sd) age aircraft in flight hours	12165 (14963)	11874 (16724)	17887 (18183)	17145 (18257)	15283 (16958)	14674 (16736)
Repaired before, n (%)						
No	403 (61)	102 (62)	713 (52)	185 (54)	1253 (70)	327 (73)
Yes	254 (39)	63 (38)	665 (48)	160 (46)	539 (30)	121 (27)
Number of previous repairs, n (%)						
0	403 (62)	102 (62)	713 (52)	185 (54)	1253 (70)	327 (73)
1	196 (30)	55 (33)	363 (26)	92 (27)	403 (22)	89 (20)
2	49 (7)	7 (4.4)	170 (12)	34 (10)	99 (6)	23 (5)
3	7 (0.7)	1 (0.6)	77 (6)	18 (5)	23 (1.3)	7 (1.5)
4+	2 (0.3)	0 (0)	55 (4)	16 (4)	14 (0.7)	2 (0.5)
Modification design, n (%)						
0	-	-	61 (4)	14 (4)	276 (15)	71 (15.7)
1	-	-	1193 (87)	299 (87)	796 (44)	172 (38)
2	-	-	124 (9)	32 (9)	457 (26)	133 (30)
3	-	-	-	-	243 (14)	66 (15)
4	-	-	-	-	20 (1)	6 (1.3)
Natural climate, n (%)						
Temperate	337 (51)	73 (44)	838 (61)	210 (61)	1118 (62)	258 (58)
Humid	223 (34)	61 (37)	408 (30)	107 (31)	503 (28)	145 (32)
Desert	97 (5)	31 (19)	132 (9)	28 (8)	171 (10)	45 (10)

and original design. Age of component and aircraft were taken at moment of installation, since manufacture date, and expressed in flight hours. Next to the number of previous removals, a binary indicator was introduced with value of one if the component has been removed before. Table II gives an overview of distribution of these covariate values for the selected components in both the train and validation set. Time-on-wing since installation in flight hours was the response variable of interest, meaning that every risk interval starts at survival time of zero (gap time risk interval). A removal was considered an event, and components still installed at moment of data extraction were considered as censored. The risk set was chosen to be unrestricted, meaning that all the component's risk intervals may contribute to the risk set for any given removal [23]. All dependence between recurrent removals was mediated through the covariates indicating the number of previous removals. Due to unavailability or unreliability of data, for some expected predictors the model could not be controlled. An overview of the predictors and their interrelation is given in Figure 3. For the reasoning behind these interrelations, the reader is referred to Appendix E. The performance of the model on the derivation data was evaluated and optimised with the AIC as an objective measure, and using scientific reasoning. Residual analysis was performed to evaluate the internal validity of the model. Values of $p < 0.05$ were considered statistically significant. The form of continuous variables, which were the age of the aircraft and the age of the component, was checked by discretization of the variable and plotting coefficient values against their mean covariate value. In case of a relevant non-linear form of these covariates, the extension of the Cox model was needed and an appropriate step size for splitting up the intervals was determined. Apart from hazard ratios, the effect of a covariate was expressed via a ratio of the RMST, here called the restricted MTBR, were the time was restricted to the largest installation duration of the specific component in the data set. In order to compute this ratio, the validation data set was used in order to have a representative distribution of the covariates. For a specific categorical variable, all values in the validation set were set to its baseline value (e.g. temperate climate), and survival curves and RMST were predicted for each component in the validation set. Then, all values of this variable were set to one value (e.g. humid climate), and again survival curves and RMST were predicted for those component. The ratio between RMST's for each component, were only this one categorical variable had changed value, was taken. This process was repeated using the bootstrapping method, in order to compute confidence intervals of the RMST-ratio. Three risk groups were defined by placing cut points on the prognostic index at 0.25 and 0.75 percentiles of derivation set demeaned (mean of stratum) PI distribution. The Cox model was externally validated by applying the fitted Cox model to the validation set and performing a graphical comparison of the predicted mean survival curves and observed Kaplan-Meier point-wise survival probabilities wit 95% confidence intervals within each risk group. Python programming language (Python Software Foundation, `https://www.python.org/`) and R statistical computing language (R Foundation for Statistical Computing, Vienna, Austria) were used to perform the statistical analysis. For more information on the modelling steps and iterations that had to take place, the reader is referred to Appendix F.

IV. Results

The final Cox regression model results are given in Table III. The statistical validation results of the proportional hazards assumption are tabulated in Table IV. The MTBR-ratio results are presented in Table V. The results for the external validation are given in Figure 7.

TABLE III: Cox regression results for the three selected components based on the derivation data

Variable	ACM			ADIRU			Display unit		
	HR[a]	*95% CI*[b]	*p-value*[c]	*HR*	*95% CI*	*p-value*	*HR*	*95% CI*	*p-value*
Sqrt aircraft age (FH)									
1 increase	1.0055	(1.0031, 1.0079)	<0.001	1.0079	(1.0066, 1.0091)	<0.001	1.0079	(1.0065, 1.0092)	<0.001
Sqrt component age (FH)									
1 increase	-	-		1.0025	(1.0010, 1.0040)	<0.001	1.0022	(1.0003, 1.0040)	0.022
Repaired									
No (*Ref*)	1	-		1	-		-	-	
Yes	1.256	(0.948, 1.663)	0.1	1.262	(1.032, 1.543)	0.05	-	-	
Number of previous repairs									
1 increase	-	-		-	-		1.113	(0.996, 1.243)	0.05
Modification design									
0 (*Ref*)	-	-		1	-		*stratified*	-	
1	-	-		1.045	(0.796, 1.371)	0.75	-	-	
2	-	-		1.711	(1.240, 2.363)	0.001	-	-	
Natural climate, n (%)									
Temperate (*Ref*)	1	-		1	-		1	-	
Humid	2.003	(1.625, 2.470)	<0.001	0.979	(0.853, 1.124)	0.76	1.006	(0.886, 1.142)	0.93
Desert	1.616	(1.236, 2.113)	<0.001	0.801	(0.659, 0.974)	0.03	0.807	(0.678, 0.961)	0.02

Results in this table should be interpreted as: adjusting for, and holding constant, all other variables.

[a] HR: hazard ratio

[b] Confidence interval with 95% chance that it contains the true value of HR

[c] p-value of < 0.05 indicates strong evidence against $H_0 : HR = 1$.

TABLE IV: Proportional hazards statistical test results based on scaled Schoenfeld residuals

Variable	ACM			ADIRU			Display unit		
	ρ[a]	χ^2[b]	*p-value*[c]	ρ	χ^2	*p-value*	ρ	χ^2	*p-value*
Repaired	-0.075	2.806	0.09	-0.025	0.766	0.38	-	-	
Number of previous repairs	-	-		-	-		0.019	0.568	0.451
Modification design 1	-	-		0.025	0.707	0.40	-	-	
Modification design 2	-	-		-0.026	0.835	0.36	-	-	
Humid	0.066	2.266	0.13	0.063	4.732	0.03	0.109	16.84	4E-05
Desert	-0.030	0.459	0.50	0.015	0.259	0.61	0.020	0.586	0.444

[a] Correlation coefficient between transformed survival time and the scaled Schoenfeld residuals.

[b] Asymptotic chi-square test statistic on one degree of freedom to test $H_0 : \rho = 0$.

[c] p-value of > 0.05 indicates weak evidence against the $H_0 : \rho = 0$, meaning the PH assumption is valid.

TABLE V: MTBR-ratio results based on predicted survival curves of components in the validation set

Variable	ACM		ADIRU		Display unit	
	MTBR-ratio[a]	*95% CI*	*MTBR-ratio*	*95% CI*	*MTBR-ratio*	*95% CI*
Age aircraft (FH)						
Increase from 0 to 3,000	0.83	(0.76, 0.87)	0.76	(0.72, 0.80)	0.79	(0.72, 0.81)
Increase from 15,000 to 18,000	0.95	(0.92, 0.96)	0.91	(0.89, 0.93)	0.93	(0.88, 0.95)
Age component (FH)						
Increase from 0 to 3,000	-	-	0.88	(0.83, 0.92)	0.92	(0.82, 0.99)
Increase from 15,000 to 30,000	-	-	0.97	(0.94, 0.99)	0.98	(0.89, 1.00)
Repaired						
No (Ref)	1	-	1	-	-	-
Yes	0.86	(0.77, 1.00)	0.80	(0.68, 1.00)	-	-
Number of previous repairs						
1 increase	-	-	-	-	0.93	(0.81, 1.00)
Modification design						
0 (Ref)	-	-	1	-	*stratified*	-
1	-	-	0.96	(0.73, 1.24)	-	-
2	-	-	0.59	(0.47, 0.79)	-	-
Natural climate, n (%)						
Temperate (Ref)	1	-	1	-		
Humid	0.62	(0.50, 0.70)	1.02	(0.91, 1.17)	1.00	(0.92, 1.15)
Desert	0.73	(0.57, 0.80)	1.23	(1.02, 1.40)	1.16	(1.07, 1.31)

[a] Ratio of restricted mean survival times (RMST), restricted by the maximum survival time for all components in the validation set.

A. Air Cycle Machine

The square root of the aircraft age, a humid hub environment, and a hot desert hub environment, were independent significant predictors for the hazard rate of an air cycle machine (Table III). Although a previous repair (HR = 1.256, 95% CI = 0.948-1.663, $p < 0.1$) was not found to be a significant predictor at 95% confidence level, it did reduce the AIC and therefore was included in the model. All these variables satisfied the proportional hazards assumption (Table IV). The component age and the number of previous repairs were not found to be significant predictors. When the age of the aircraft in which the component was installed increased from 0 to 3,000 flight hours, the model predicted that the time-on-wing duration since installation reduced on average to 83%, ceteris paribus (Table V). This reduction was less for an age increase of the aircraft from 15,000 to 18,000. The predicted decrease in MTBR for a previously repaired component with respect to a new component equalled 14%, ceteris paribus. For a humid and desert environment with respect to a temperate environment, the estimated reduction in MTBR was 38% and 28% respectively. This can visually be verified by a plot of the mean predicted survival curves per hub climate class as is given in Figure 4. As example, the distribution of the MTBR-ratio for a desert hub environment derived from Cox model fits on 1,000 bootstrap samples from the derivation data is given in Figure 5.

Distributions of the prognostic index for both derivation and validation data sets were similar (Figure 6a), meaning external validation via risk group creation was possible. The Cox model separated the survival estimates between risk group well, and the calibration was almost perfect for risk group 1 (Figure 7a). The model slightly over-predicted for risk group 2 and 3, although survival estimates lay still in the 95% confidence interval of the observed estimates.

B. ADIRU

The square root of the aircraft age, the square root of the component age, a previous repair, a second modification design, and a hot desert hub environment were independent significant predictors for the hazard rate of an ADIRU (Table III). The effect of the number of previous repairs had no extra predictive power and was therefore excluded in the final model. The indicator variables for a first modification design and a humid hub climate were also found to be insignificant predictors, but were included in the model because the other classes of the main variable (i.e. modification design 2 and desert climate) were found to be very significant; leaving out one class results in a changed reference class. All time-dependent and significant predictors satisfied the proportional hazard assumption (Table IV). An aircraft age increase from 0 to 3,000 flight hours resulted in an estimated time-on-wing duration reduction of 24%, whereas an aircraft age increase from 15,000 to 18,000 flight hours yielded time-on-wing duration reduction of 9%, ceteris paribus (Table V). For a component age increase from 0 to 3,000 flight hours and from 15,000 to 18,000 flight hours, the mean time-on-wing duration reduction equalled 12% and 3% respectively, ceteris paribus. The predicted decrease in MTBR for a previously repaired component with respect to a new component equalled 20%, ceteris paribus. The second modification design performed significantly worse than the reference modification design: on average it was predicted to be installed only 59% of the average installation duration of the reference design. Finally, a 23% rise in MTBR was predicted for a component operated from a desert setting compared to a temperate setting. The spread of the log relative hazard for both derivation and validation data sets were similar (Figure 6b), and no obvious outliers were found. External validation was satisfactory as the model discriminated properly between risk groups and predicted accurately the survival curves per risk group on the validation data (Figure 7b).

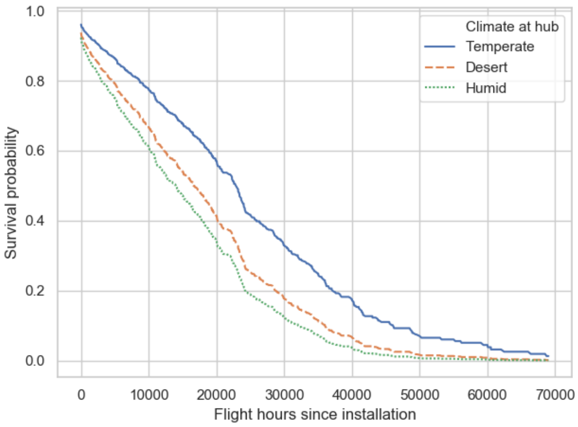


Fig. 4: Mean predicted survival curves per hub climate class for the ACM, based on altered values in validation set. The area under the curve represents an estimate of the restricted MTBR.

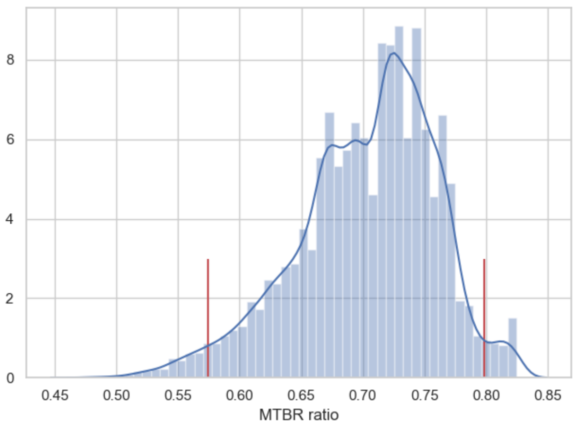
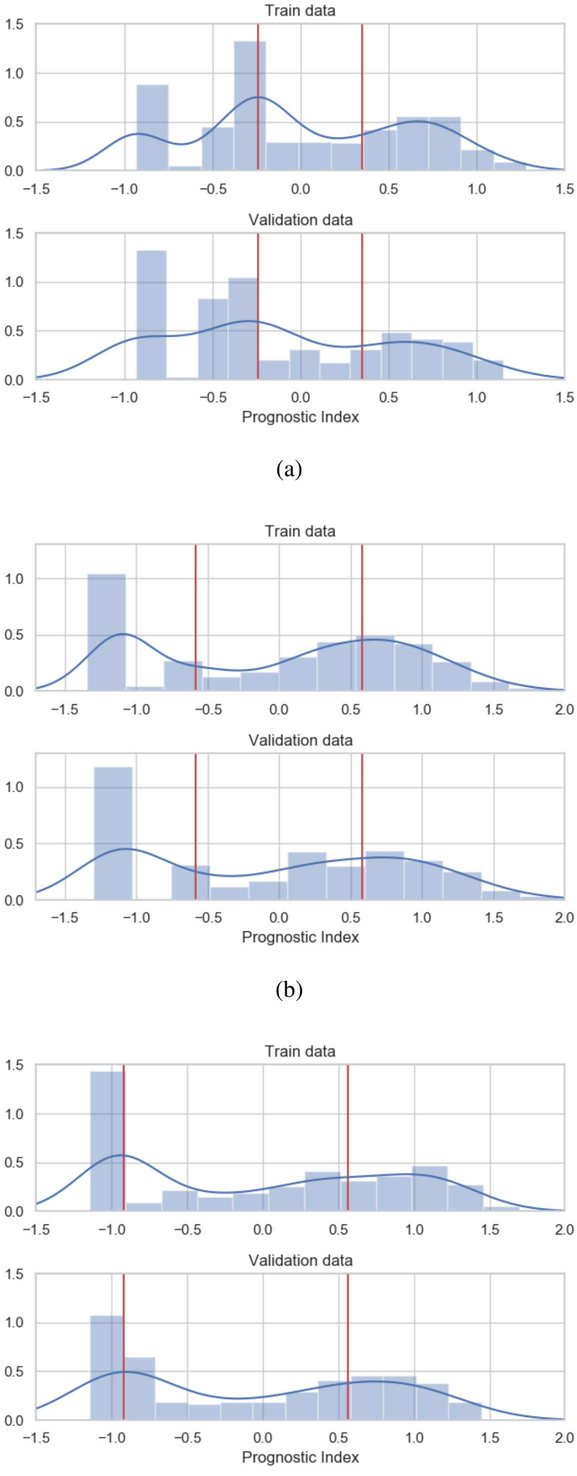


Fig. 5: Distributions of the RMST-ratio for covariate *desert* of the ACM, derived from Cox model fits on 1,000 bootstrap samples. The red vertical lines indicate the 2.5th and 97.5th centiles, which approximate the true 95% CI limits.

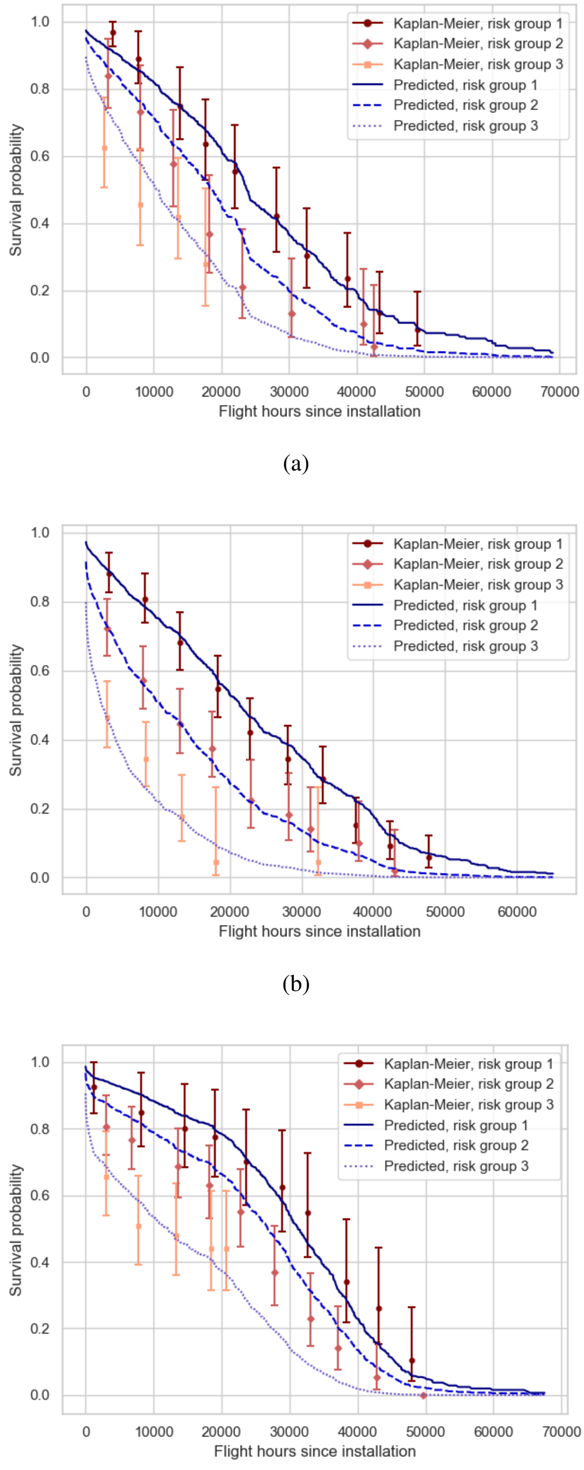


(a)

(b)

(c)

Fig. 6: Normalised histogram of the Prognostic Index for the ACM (a), ADIRU (b) and modification design 1 of the Display unit (c). The PI was centered on the mean in the train dataset. The blue line indicates the kernel density estimate, the vertical lines indicate the 33th and 66th centile risk bands.



(a)

(b)

(c)

Fig. 7: Discrimination and calibration of the Cox model in the validation dataset for the ACM (a), ADIRU (b) and modification design 1 of the Display unit (c). The vertical capped lines denote 95% confidence intervals of the Kaplan-Meier estimates.

C. Display Unit

The aircraft age and the component age in square root form, each extra previous repair, and a hot desert hub environment were independent significant predictors for the hazard rate of an ADIRU (Table III). Note that there were five different modification designs of the display unit in the data (Table II), however these did not satisfy the proportional hazards assumption with respect to the baseline design. Therefore, in order to satisfy the PH assumption, stratification was applied to this variable. The number of previous repairs, although slightly surpassing the $p \leq 0.05$ threshold for significance, still added significantly to the model performance based on AIC. All included significant variables satisfied the proportional hazards assumption (Table IV). An aircraft age increase from 0 to 3,000 flight hours resulted in an estimated time-on-wing duration reduction of 21%, whereas an aircraft age increase from 15,000 to 18,000 flight hours yielded time-on-wing duration reduction of 7%, ceteris paribus (Table V). For a component age increase from 0 to 3,000 flight hours and from 15,000 to 18,000 flight hours, the mean time-on-wing duration reduction equalled 8% and 2% respectively, ceteris paribus. The predicted decrease in MTBR for each extra previous repair of the component equalled 7%, ceteris paribus. Finally, a 16% rise in MTBR was predicted for a component operated from a desert setting compared to a temperate setting.

Because of the stratified modification design covariate, different baseline hazard rate and survival curves were present per stratum. This meant that the external validation process had to be performed for each stratum separately. Due to the low number of data points for most strata, the Kaplan-Meier survival estimates had to much uncertainty making validation impractical. Therefore, only validation results for the stratum with most data points was carried out, which was for modification design 2. The prognostic index distribution of this stratum for both derivation and validation data sets were similar (Figure 6c). External validation was satisfactory as the model discriminated properly between risk groups and predicted accurately the survival curves per risk group on the validation data (Figure 7c). The model slightly over-predicted for risk group 2, although survival estimates lay still in the 95% confidence interval of the observed estimates.

V. Discussion

By using an extended version of the Cox regression model, the effect of ageing aircraft has been estimated more precisely by making possible other forms of its functional form. The effect of ageing aircraft was found to be hazard increasing for all three analysed components. This effect however decreased over time, as evident from the square root form. Another remarkable result is that ageing effects of the aircraft seem to impact the reliability of the component more than ageing of the component itself. A possible explanation is the colinearity present between component age and previous repairs, which was also included as a covariate in the model and which might take away some of the predictive power of the component age. This also explains the relatively high p-value found for the estimates of the effect of a previous repair. The common assumption in reliability analysis that renewal takes place after failure and repair of a component is not valid for the analysed components, as previous repairs of a component were shown to impact the following time-on-wing duration. Repairable components have the characteristic that complete replacement does not take place after failures, resulting in sub-parts which are degraded and still present in the component. An example is the ADIRU, which consists of three ring laser gyroscopes. When one gyro breaks down, the repair shop might choose to only repair or replace that single gyro, while the other gyro's might soon fail as well.

Controlling for all other covariates, it was found that modification design 2 of the ADIRU performed significantly worse than the pre-modification design. Such information can be valuable for MRO companies, as oftentimes the same flight-hour based price is set for all designs within one component. It shows that a modification does not always lead to a higher reliability.

As expected, air cycle machines in aircraft operated from hubs with a desert or humid climate have much lower mean time-on-wing than those same components installed in aircraft operated in e.g. Western Europe. This information can directly be implemented by MRO service providers in order to improve estimates of the expected number of removals and repairs per customer. One remarkable result that is found in Table III is for the ADIRU and display unit, where a hot desert climate at the hub actually seems beneficial with respect to the reliability of those components. An explanation for this can be deducted from Figure 3; because the natural environment is related to the operator and other potential covariates like operator's skill and standards are not controlled for, the natural climate covariate might be confounded by these uncontrolled covariates, i.e. operators in this dataset with their hub in a desert might have high operation quality standards. This result directly shows one limitation of this study.

A unique aspect of this research with respect to available reliability research literature in the airline industry is the large data set used. Airline components have a very low failure rate, resulting in a small number of maintenance data points per aircraft. The size and variety of the available data made possible that the Cox model could identify differences in reliability, for example due to hub climate differences. Environmental factors are hard to couple to failure instances as aircraft are inherently non-stationary. To the best of the author's knowledge, using the Köppen climate dataset by geographic location coupled to the hub were the aircraft is based in order to quantify environmental impact on aircraft component reliability has not been done before in literature.

As mentioned in the introduction, the interpretation and practical usefulness for industry of a hazard ratio, which is the standard output of a Cox model, is not so straightforward. By quantifying the effect on time-on-wing performance via the MTBR-factor, a widely used measure for reliability, the results found can more easily be applied. Using the bootstrap method in order to compute confidence intervals was necessary, as uncertainty measures are of uttermost importance.

In literature, it is very uncommon to validate the Cox models on an external dataset. This research showed via the method of Royston [12] that the Cox model performed satisfactorily on the validation data. A problem ran into when using this method, was in case of the stratified Cox model for the display unit. Because a separate baseline hazard and survival function are generated per stratum, a multiple of the external data size is needed in order to be able to generate Kaplan-Meier estimates with respectable confidence intervals.

Limitations of this study include the retrospective setting, the inability to control for all hypothesised covariates, and the selection bias due to data availability limitations which only included components which have been removed at least once. Furthermore, no distinction between failure modes was made, as this information was not readily available on such a large scale. Finally, the author was not in the possession of a completely independent data set for external validation of the Cox model, and therefore had to use a randomised split of the original data in order to mimic out-of-sample predictive performance.

References

[1] E. Commission, "Keeping the aviation industry safe - safety intelligence and safety wisdom," *A Future Sky Safety White Paper*, 2016.

[2] I. M. C. T. Force, "Airline maintenance cost executive commentary - exclusive benchmark analysis (fy2017 data)," Nov. 2018.

[3] J. Kilpi and A. P. Vepsäläinen, "Pooling of spare components between airlines," *Journal of Air Transport Management*, vol. 10, no. 2, pp. 137–146, 2004.

[4] D. J. Smith, "Power-by-the-hour: The role of technology in reshaping business strategy at rolls-royce," *Technology Analysis & Strategic Management*, vol. 25, no. 8, pp. 987–1007, 2013.

[5] D. M. Louit, R. Pascual, and A. K. Jardine, "A practical procedure for the selection of time-to-failure models based on the assessment of trends in maintenance data," *Reliability Engineering & System Safety*, vol. 94, no. 10, pp. 1618–1628, 2009.

[6] W. J. Verhagen and L. W. de Boer, "Predictive maintenance for aircraft components using proportional hazard models," *Journal of Industrial Information Integration*, vol. 12, pp. 23–30, 2018.

[7] B. Pogačnik, J. Duhovnik, and J. Tavčar, "Aircraft fault forecasting at maintenance service on the basis of historic data and aircraft parameters," *Eksploatacja i Niezawodnosc – Maintenance and Reliability*, vol. 19, no. 4, pp. 624–633, 2017.

[8] D. Kumar and B. Klefsjö, "Proportional hazards model: A review," *Reliability Engineering & System Safety*, vol. 44, no. 2, pp. 177–188, 1994.

[9] D. R. Cox, "Regression models and life-tables," *Journal of the Royal Statistical Society: Series B (Methodological)*, vol. 34, no. 2, pp. 187–202, 1972.

[10] A. C. Mendes, "Proportional hazard model applications in reliability," *Northeastern University (doctoral dissertation)*, Apr. 2014.

[11] M. J. Bradburn, T. G. Clark, S. B. Love, and D. G. Altman, "Survival analysis part iii: Multivariate data analysis – choosing a model and assessing its adequacy and fit," *British Journal of Cancer*, vol. 89, no. 4, pp. 605–611, 2003.

[12] P. Royston and D. G. Altman, "External validation of a cox prognostic model: Principles and methods," *BMC Medical Research Methodology*, vol. 13, no. 1, 2013.

[13] K. Pak, H. Uno, D. H. Kim, L. Tian, R. C. Kane, M. Takeuchi, H. Fu, B. Claggett, and L. Wei, "Interpretability of cancer clinical trial results using restricted mean survival time as an alternative to the hazard ratio," *JAMA Oncology*, vol. 3, no. 12, p. 1692, 2017.

[14] J. D. Kalbfleisch and R. L. Prentice, *The Statistical Analysis of Failure time Data*. J. Wiley, 2003.

[15] J. P. Klein and M. L. Moeschberger, *Survival Analysis: Techniques for Censored and Truncated Data*. Springer, 2010.

[16] R. J. Cook and J. F. Lawless, *The statistical analysis of recurrent events*. Springer, 2007.

[17] D. G. Kleinbaum and M. Klein, *Survival Analysis: A Self-Learning Text*. Springer, 2012.

[18] T. M. Therneau and P. M. Grambsch, *Modeling survival data: extending the Cox model*. Springer, 2011.

[19] L. Moulton and M. Dibley, "Multivariate time-to-event models for studies of recurrent childhood diseases," *International Journal of Epidemiology*, vol. 26, no. 6, pp. 1334–1339, 1997.

[20] D. F. Moore, *Applied survival analysis using R*. Springer, 2016.

[21] K. Chamie, M. S. Litwin, J. C. Bassett, T. J. Daskivich, J. Lai, J. M. Hanley, B. R. Konety, and C. S. Saigal, "Recurrence of high-risk bladder cancer: A population-based analysis," *Cancer*, vol. 119, no. 17, pp. 3219–3227, 2013.

[22] E. A. Peña and M. Hollander, "Models for recurrent events in reliability and survival analysis," *Mathematical Reliability: An Expository Perspective*, pp. 105–123, 2004.

[23] P. J. Kelly and L. L.-Y. Lim, "Survival analysis for recurrent event data: An application to childhood infectious diseases," *Statistics in Medicine*, vol. 19, no. 1, pp. 13–33, 2000.

[24] K. Liang and S. L. Zeger, "Longitudinal data analysis using generalized linear models," *Biometrika*, vol. 73, no. 1, p. 13, 1986.

[25] D. Y. Lin and L. J. Wei, "The robust inference for the cox proportional hazards model," *Journal of the American Statistical Association*, vol. 84, no. 408, pp. 1074–1078, 1989.

[26] T. R. Fleming and D. P. Harrington, *Counting Processes and Survival Analysis*. Wiley, 2011.

[27] J. K. Lindsey and B. Jones, "Choosing among generalized linear models applied to medical data," *Statistics in Medicine*, vol. 17, no. 1, pp. 59–68, 1998.

[28] P. M. Grambsch and T. M. Therneau, "Proportional hazards tests and diagnostics based on weighted residuals," *Biometrika*, vol. 81, no. 3, pp. 515–526, 1994.

[29] D. Schoenfeld, "Partial residuals for the proportional hazards regression model," *Biometrika*, vol. 69, no. 1, p. 239, 1982.

[30] H. Uno, B. Claggett, L. Tian, E. Inoue, P. Gallo, T. Miyata, D. Schrag, M. Takeuchi, Y. Uyama, L. Zhao, and et al., "Moving beyond the hazard ratio in quantifying the between-group difference in survival analysis," *Journal of Clinical Oncology*, vol. 32, no. 22, pp. 2380–2385, 2014.

[31] L. Trinquart, J. Jacot, S. C. Conner, and R. Porcher, "Comparison of treatment effects measured by the hazard ratio and by the ratio of restricted mean survival times in oncology randomized controlled trials," *Journal of Clinical Oncology*, vol. 34, no. 15, pp. 1813–1819, 2016.

[32] B. Efron and R. Tibshirani, "Bootstrap methods for standard errors, confidence intervals, and other measures of statistical accuracy," *Statistical Science*, vol. 1, no. 1, pp. 54–75, 1986.

[33] P. Royston, "Tools for checking calibration of a cox model in external validation: Prediction of population-averaged survival curves based on risk groups," *The Stata Journal: Promoting communications on statistics and Stata*, vol. 15, no. 1, pp. 275–291, 2015.

[34] "737 model orders and deliveries data," *Boeing*, Apr. 2019.

[35] D. Chen and H. W. Chen, "Using the köppen classification to quantify climate variation and change: An example for 1901–2010," *Environmental Development*, vol. 6, pp. 69–79, 2013.

[36] *Airport, airline and route data*. [Online]. Available: `https://openflights.org/data.html`.

Appendices

Appendix A

Research Methodologies

Previously graded under AE4010

A.1 Abstract

KLM E&M provides component availability for numerous airline customers. To ensure on-time performance, inventory of spare parts is an indelible part of their operations. Demand forecasts form the basis for the planning of inventory levels, however, the demand tends to be random and is intermittent. Many forecasting methods for aircraft spare part demand exist in literature, none of which utilise the true drivers of demand. Information on the units of component class actually in use, also called installed base information, could be leveraged for spare part demand forecasting. The research will be about developing such a causal forecasting model, comparing its performance with state-of-the-art methods, with the potential of a significant cost decrease for KLM E&M and other Maintenance, Repair and Overhaul players worldwide.

A.2 Introduction

Demand forecasting is one of the most essential concerns of inventory management in the repair and overhaul industry. Forecasts form the basis for the planning of inventory levels, and the high cost of modern aircraft and the expense of such repairable spares as aircraft components and avionics constitute a large part of the total investment of many airline operators [1]. These parts are critical to operations and their unavailability can lead to excessive down time costs. However, forecasting spare parts demand is challenging as the demand tends to be irregular with a great amount of zero-demand periods.

Having spare part inventory allows for an immediate substitution in case of repair, postponing the repairing or buying activities only after having restored system's operations, minimizing the downtime of the aircraft. It relies on defining a stock quantity able to cover the demand within the re-supply time, minimizing the inventory costs. Nuclear plants, oil drilling, defence and transportation industry are some industries where spare parts management is gaining an increasing attention as characterized by low failure rate, high inventory and stock-out costs [2]. Hence, it makes sense to consider demand forecasting using all available information on the so-called installed base, and to find relevant explanatory variables from this installed base information.
In this light a MSc. graduation research project is performed at KLM Engineering & Maintenance. At the very start of the project a research objective had been formulated:

To analyse relationships of operational and maintenance data with spare part reliability and demand with the aim to increase forecasting accuracy with respect to state-of-the-art methods.

A.3 State of the art & Literature review

The issue of spare parts demand forecasting has been studied for many years, which has resulted in the development of numerous prediction methods and techniques [3]. Various categories of forecasting methods exist, however most papers concern quantitative forecasting methods [4]. These are approaches where numerical information about the past is used in order to predict future demand, and unlike qualitative methods, these methods are objective; once the underlying model or technique has been chosen, the corresponding forecasts are determined automatically and hence they are fully reproducible by any forecaster. The weakness however is the need for data and the danger of using unreliable and unclean data sets and sources. Furthermore, the assumption is made that the underlying model does not change over time. There are mainly two groups of quantitative forecasting techniques: causal and time-series methods. A time series is defined as a time-ordered sequence of observations taken at regular intervals (e.g., hourly, daily, weekly, monthly, quarterly, annually) [5]. It is based on the assumption that future values of the target variable can be estimated from past values of this same variable. By discovering patterns in the past values of the variable it extrapolates these into the future and uses it to predict future values of the variable of interest. Causal forecasting methods on the other hand are a way of estimating future demand by finding a relation between explanatory variables and spare parts demand [6]. The motivation for such causal model is the assumption that the variable to be forecast, the dependent variable, has cause-and-effect relationship with one or more other (independent) variables. Caution should be taken with classifying methods as *causal*, as often *correlation* is found between independent and dependent variables but correlation does not imply causation. Establishing a causal relationship between two variables is actually one of the biggest statistical challenges from both a theoretical and practical perspective [7].

Forecasting aircraft spare parts demand is challenging as the demand tends to be irregular with a great amount of zero-demand periods. The state-of-the-art method for classifying demand patterns as proposed by Syntetos et al. [8] and tested and validated in multiple succeeding studies [9] [10] classifies demand patterns into four categories: intermittent, slow moving, erratic and lumpy demand. It uses the average demand interval (ADI) and squared coefficient of variation of demand sizes (CV^2), which indicate the average time between occurrence of subsequent demands and the extent of demand size variability respectively. Intermittent or lumpy demand, as often seen for aircraft spare parts, is in literature predominantly forecasted using time-series methods [3], yet the methods differ from the classical time-series techniques due to the number of zero demand values and being data of counts [1] [11]. This is due to the fact that most state-of-the-art time-series forecasting techniques, like weighted averages and regression based forecasting methods, are not well capable of capturing the intermittency of the data [12]. Artificial Neural Networks, and especially the most simple 3-layer perceptron have shown potential in capturing intermittent patterns due to their flexible and non-linear nature [13]. They however require a vast amount of data for setting the estimator and for outperforming conventional statistical methods as Markham and Rakes [14] proofed, which is often not available for intermittent demand patterns. Furthermore, ANN's do not give insight in the demand generating process. The Syntetos-Boylan Approximation seems the most suitable time-series method for forecasting intermittent demand patterns, as it is theoretically more sound than Croston's method and simple and easy to implement [15] [16]. Adding to that, it has been shown in numerous studies to perform equal or better compared to other proposed method's [12] [1].

A different way to tackle the forecasting problem at hand is to make use of cross-sectional and/or temporal aggregation [17]. By aggregating the data on a different level in the hierarchy of the product or part, or in lower frequency time units, the number of zero demand observations will be less. Given the reduction of zero observations, a far richer arsenal of forecasting methods and models are available to be employed for time-series extrapolation. Furthermore, empirical studies show that temporal aggregation also generally reduces demand volatility meaning CV^2 is reduced [16] [18]. This is however not always the case and no theory exists in current literature which identifies when and when not temporal aggregation leads to decrease in CV^2 [17]. The applicability of aggregation depends on the situation, and expert opinion is often needed. Furthermore, there is no conclusive solution in literature with regard to the identification of the appropriate time aggregation level, and it remains to be analysed if the demand forecast improves [18].

In spare part demand forecasting literature, causal forecasting methods are very scarce and use of variables from the installed base information like age of fleet, age of components, and environmental impact is non-existent, which is identified as the main research gap. Auweraer *et al.* [3] recently reviewed the literature of using installed base information for aircraft spare part demand forecasting, however it lacks an overview of the potential installed base information and interrelations between the variables and mostly discusses installed base size as a variable. Even though the idea of causal forecasting methods for spare part demand sounds straightforward, it is however not easily realised in practice as it requires information on causal variables. This installed base information needs to be maintained which is frequently not the case, or the information is unreliable and is scattered throughout different legacy information systems [19]. The few practical applications that have been undertaken to include causal factors are mainly in forms of reliability and survival analysis [20] [21] [22] [19] and only include time-invariant factors like environmental impact. Although some of those studies found that installed base information is relevant for modelling the reliability of component, no direct link to forecasting future spare part demand is made.

The performance increase regarding reliability modelling, as found in the scarce literature available, is not translated into intermittent demand forecasting.
Although numerous comparative studies in the literature exist regarding performance between the various state-of-the-art intermittent demand forecasting methods, the performance criteria used differ and the results are often inclusive [23] [24]. The most commonly used per period forecast error is not informative when not combined with other measures for demand series that consist of many zeros and few positive demands. A key error metric which has been used extensively in literature since its introduction is the Mean Absolute Scaled Error, as it effectively scales the errors and does almost never give an undefined number [25]. Nevertheless, a better way of comparing forecasting methods for slow-moving items is to analyse their effect on inventory control parameters and to compare resulting inventory and service level or inventory costs [12]. This however requires a lot of simplifying assumptions or extensive simulations studies, which are time- and cost-inefficient.

Research gap The motivation of this research is to leverage installed base information in causal spare part demand forecasting models. As can be concluded from the extensive literature review, most studies in spare part demand forecasting (or intermittent pattern forecasting in general) only consider time-series forecasting methods. To the extent of the author's knowledge, no studies in academia or in industry have been performed on the applicability of causal forecasting methods for spare part demand in a real life aircraft maintenance and operations case using installed base information as age of fleet, age of components, environmental impact and airline operator differences. The aim of the research therefore is to fill this research gap.

A.4 Research question, aims and objectives

In various industries which constitute parts with low failure rate, high inventory and stock-out costs (e.g. transportation, defence, oil drilling), spare parts management is gaining an increasing attention. As a result, it makes sense to consider demand forecasting using all available information on the so-called installed base, and to find relevant explanatory variables from this installed base information. This is especially true since the values of underlying demand generating factors might change over time. To address the research gaps and add to the current state of the art, the following research objective has been formulated:

To develop a reliability model for aircraft repairables using operational and maintenance data with the goal to quantify the effects of maintenance, climate and ageing of aircraft and component on the time-on-wing duration.

In order to reach the objective, tangible sub-goals are defined to further structure the research process:

1. To retrieve reliable and clean (historical) data on component operational and maintenance information (e.g. age in flight hours and cycles, time since last installation, install and removal data).
2. To retrieve reliable and clean (historical) data on the operating environment (e.g. temperature, humidity, sand at hub).
3. To asses the statistical inference of variables retrieved from maintenance data on time-on-wing durations.
4. To verify an validate the reliability model's performance and to compare it with state-of-the-art methods' performance.

The achievability of the first four sub-goals directly relates to the feasibility of the study: if insufficient reliable data is available, the feasibility of validating forecasting performance using installed base variables decreases.

The main research question to be solved in reaching the project goal is formulated as:

How can operational and maintenance data be leveraged for reliability modelling of aircraft repairables?

This research question gives rise to some important elements of the research. First of all, variables from the available databases which have potential predictive power for spare part demand have to be identified. Furthermore, a choice of the reliability modelling technique has to be made, and the usage of the potential covariates in the reliability model has to be determined. Lastly, the question is defined as a feasibility study framed by the conditions of a real life use case. The research is conducted in collaboration with KLM Engineering & Maintenance, and the scope of the research will specifically be for Boeing 737-NG aircraft models. After specifying the forecasting model, it needs to be defined, with the knowledge from the literature study on performance, what the criteria for success are that will be used to answer the question. This includes a performance comparison with the current practice. All this can be summarised in the following research subquestions:

1. What operational and maintenance information can be leveraged?
   (a) Which data sources can be used for obtaining operational and maintenance information?
   (b) How can the operational and maintenance information be retrieved from these data sources?
   (c) How reliable are these data sources and the corresponding data?
   (d) How can erroneous data be recognised and cleaned correspondingly?
   (e) What information on component operational and maintenance history is available?
   (f) What information on operating environment can be leveraged?
   (g) How are these variables interrelated and related to the reliability of the component?
2. How is the scope regarding aircraft spare parts determined?
   (a) Which components receive extra attention by KLM E&M engineers and for which reasons?
   (b) How does the reliability, availability and variability of data impact the reliability model's performance?
3. How to use operational and maintenance covariates in a reliability model?
   (a) Which operational and maintenance covariates have predictive power for reliability of component?
   (b) Which reliability model is most suitable for assessing the effect of operational and maintenance covariates?
   (c) How to cope with censored data?
   (d) How to cope with multicollinearity between variables?
4. How well do reliability models with operational and maintenance information for aircraft repairables perform with respect to state of the art methods?
   (a) Which performance metrics are suitable for testing the reliability models?
   (b) How to verify and validate performance of the reliability model?
5. How can the newly acquired insights from the developed reliability model be used and implemented for KLM E&M?
   (a) What is the current reliability methodology of KLM E&M?
   (b) How can the current reliability methodology be adjusted or replaced by newly acquired insights from the developed reliability model with the operational and maintenance covariates?

A.5 Theoretical Content & Methodology

As the research objective and question as established in Section A.4 are now clear, the steps to be undertaken in order to achieve the objectives and corresponding theoretical basis can be described. Due to the nature of the problem at hand the methodology is similar to a classical data science process. The description of the proposed methodology is therefore based on the of the working principle is based on the work of Ali *et al.* [26] and of McKinney [27].

First and foremost, data needs to be collected. During the *data retrieval* process, which depends on the needed data as described in the first four sub-objectives, the existence of the data, quality and access to the data is checked. Data can be stored in-house, on the internet or be delivered by third-party companies. It can take many forms ranging from Excel spreadsheets to different types of databases.
Given that data is extracted from the various sources, *data preprocessing* needs to be performed. Data collection is an error-prone process and therefore the quality of the data needs to be enhanced and prepared for use in subsequent steps. This phase consists of three sub-phases: data cleaning removes false values and inconsistencies across data sources, data integration or aggregation enriches data sources by combining information from multiple data sources, and data transformation ensures that the data is in a suitable format for use in respective models.
The third step is *exploratory data analysis*, in which a deeper understanding of the data is build: relationships between variables, the distribution of the data, whether there are outliers etc. Methods used in order to achieve this understanding are mainly descriptive statistics, visual techniques, and simple modelling.
The next step consists of *building a predictive model*. In this phase different models, domain knowledge, and insights about the data found in the previous steps are used to answer the research question. A technique from the fields of statistics, machine learning, operations research or reliability modelling needs to be selected. A hybrid approach could also be taken, meaning various techniques are combined. Building a model is an iterative process that involves variable building and selection (a.k.a. feature engineering), executing the model, and model diagnostics.
The last phase entails *verification and validation* of the built model, which are connected to the fifth and sixth research sub-objective. This step makes sure the model is correct and works as expected. A technique regularly used in literature is cross-validation, which is a resampling procedure used to evaluate statistical or machine learning models on a limited data sample.

A.6 Experimental Set-up

Causal forecasting has shown great capabilities for demand prediction in general, but is limited by the availability and reliability of the causal variables. The academic contribution of this research is to assess the feasibility of techniques to forecast aircraft spare part demand with real life operational and maintenance data. For that reason a case study centred approach is chosen. The project will be performed at KLM Engineering & Maintenance, where the case study has been formulated based on the potential industrial benefit on one hand, and the potential academic value on the other. Using real life data from KLM E&M and computer programming, the forecasting model will be tested and validated. The case study will is therefore partly defined by component scope and partly by scenario analysis.
Component scope
As stated in Section A.4 as part of the research question, it is important to confine the research by means of scoping the components to analyse. This is necessary, due to the vast amount of different components KLM E&M has in their supply chain (2000+) and due to time and resource constraints. The components for which a case study is formulated will be selected based on various criteria. These factors are also included in the second research sub-question.

- **Availability and reliability of historical data** As is entailed in the first research sub-question, it is important to look at what installed base information can be leveraged. If for some components there is more data available and/or of a higher quality, this could help narrow the scope. This is also known as convenience sampling. Furthermore as described in the literature study in Section A.3, four different demand pattern classes exist, for which different model techniques are applicable and optimal. Components with a too lumpy or intermittent pattern could be discarded or delayed until results are booked on components with simpler (i.e. less intermittent) demand patterns. See research question 2.2.

- **Added value of performance increase over current method** This is linked to research question 2.1. Adding value is a broad term, and in this case study format it can be influenced by various factors. For example, a better forecast for components which are the biggest cost-drivers or bottlenecks for KLM E&M has more added value than a better forecast for components which are already forecasted well. Another example can be components which are already phased out or in the process of phasing out; there might be enough historical data available, however the forecast will be less useful for the future. A third example is the generalisation of the component to other components: if the findings can generalise to a lot of other components, the impact of the research and therefore also the added value is larger.

Scenario analysis
The case study could also entail scenario analysis, which are predetermined future scenarios and corresponding changes in values of the underlying variables. This solves the problem of having to forecast the underlying independent variables first before generating a demand forecast. Examples of such scenarios are: the introduction of a new client with a very old fleet, the introduction of a new client very new components, the introduction of a new client making high number of cycles per day, the introduction of a new client with it's hub airport in a dessert area, buying new or second-hand components etc. This is linked to validation part of the research as stated in research sub-question 4.2.

A.7 Results, Outcome and Relevance

Before the outcome of interest is projected, it is useful to describe the data and corresponding variables which will be analysed. The main data source will be install and removal data, as logged internally at KLM E&M. Furthermore, natural environment factors will have to be attained from online data sources or from third party companies. Variables from these data sources include:

Component specific variables are associated with information about the specific individual components installed on the plane. The expectation is that as the age of a component increases, the number of times it needs maintenance (and therefore the number of removals) also increases. Furthermore, KLM E&M engineers have an hypothesis that the time on wing of some components decreases after each maintenance operation on the specific components. Variables to include are therefore:

1. Age
   (a) Time Since New [flight hours, cycles, days]
   (b) Time Since Installation [flight hours, cycles, days]
2. Maintenance history [#previous repairs]
3. Hard-time limit [flight hours, cycles, days]

Operational exposure variables have to do with the number of components exposed as well as the amount of time these components are in operation. The expectation is that components degrade if in use or if time passes by (e.g. electronic wear over time), which is the main assumption in reliability and survival analysis as found in literature. Examples of variables considered are:

1. Installed base size [#components]
2. Effective fleet size [#tails]
3. QPA [#components/tail]

4. Operating time [flight hours/day]
5. Operating frequency [cycles/day]
6. Intensity of use [flight hours/cycles]
7. Seasonality [calendar date]

Operating environment variables include factors influencing the specific component from the outside. Apart from natural environment factors, aircraft factors also fall under this group as it ‘surrounds’ the component. The expectation is that if the operating environment is harsh (e.g. old aircraft in dessert area), it negatively impacts the reliability of the individual installed component as well.

1. Natural environment
   (a) Temperature at hub
   (b) Humidity at hub
   (c) Sandy at hub [yes/no]
   (d) Location [city/country/world region of hub] (which incorporates salinity, humidity, sand, temperature, air pollution, terrain, radiation)
   (e) Airline (which also incorporates intensity of (mis)use, technical education of users, storage conditions)
2. Aircraft
   (a) Age [flight hours, cycles, days]
   (b) Maintenance history [#previous repairs]

The relevance of establishing relationships with aforementioned variables and spare part demand is twofold. First, there is a potential in improvement of forecasting performance with respect to state of the art methods by including these explanatory variables. As already stated in the literature review in Section A.3 this would fill a huge research gap in literature. Second, for the first time the expected relationships between individual installed base features and spare part demand will be quantified in a real life maintenance aviation case study. It could show the potential and value of having and storing the installed base information to Maintenance, Repair and Overhaul players across the world. The value would come from improved forecasts in scenarios as will be tested and validated in the scenario analysis as described in Section A.6. Examples of scenarios are the introduction of a new client, operational changes of current clients; the added value could be a more optimal stock level, decreased borrows, more accurate flight hour based contract prices etc.

A.8 Project Planning and Gantt Chart

In order to reach the main research objective and answer the research question, a project plan is important because it ensures there is a proper plan for executing on research goals. From the research objectives and questions as mentioned in Section A.4 the intended work can be distributed into work packages and incorporate into a schedule of work through a Gantt Chart, which is shown in Figure A.1. The work is distributed in two phases: the initial phase and the final phase, which are separated by a successful mid-term meeting. The initial phase entails the first four steps from the methodology as described in Section A.5. As stated there, these steps are part of an iterative process meaning that one step is (potentially) not finished before the next step starts. This is made clear in the Gantt Chart in Figure A.1 by the overlap of the task periods. The final phase consists of verification and validation of the model, including the scenario analysis as previously explained. This is however also an iterative process with the predictive analysis and model building task, which therefore is part of both the initial and final phase. Key review points and deliverables are presented as milestones in the Gantt Chart, and if a milestone is not made on time, the interlinked tasks in the schedule will shift accordingly with the duration of the delay.

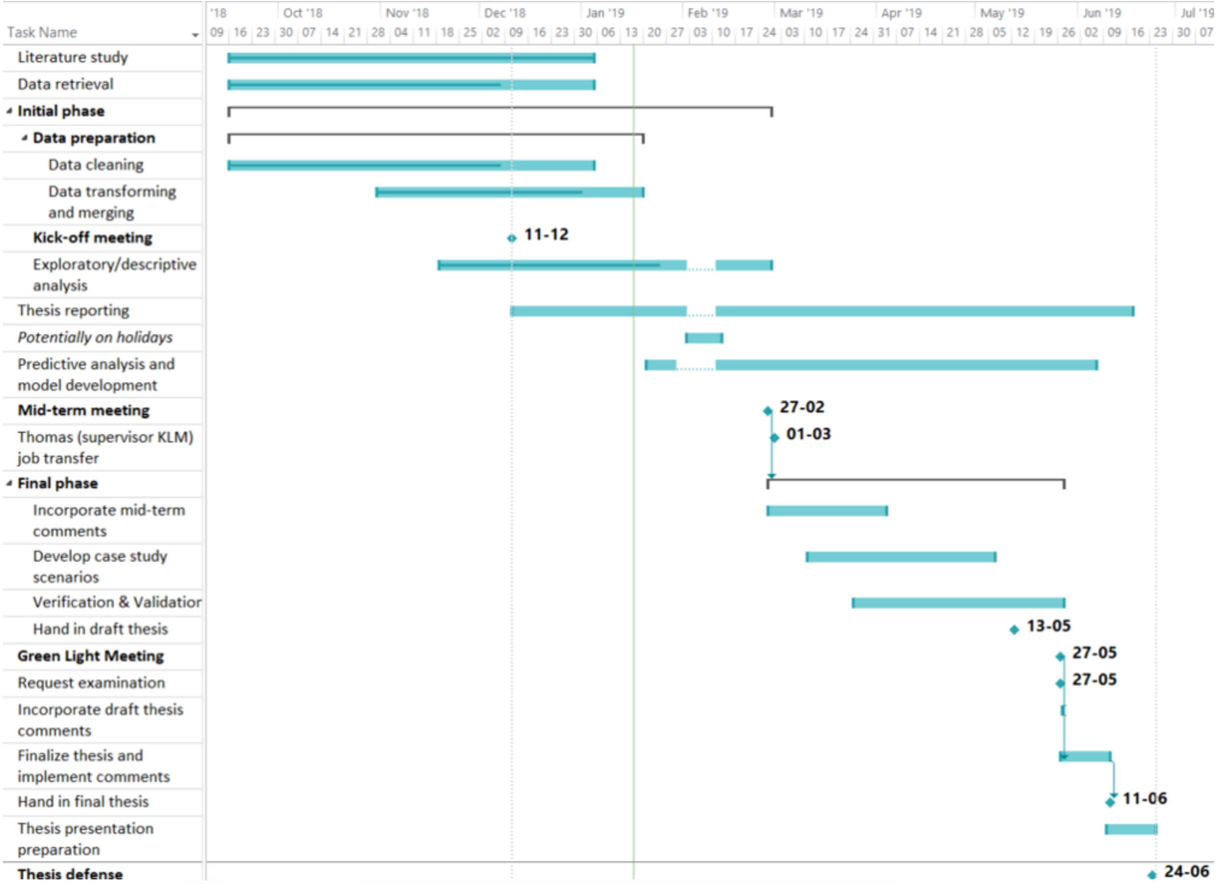


Figure A.1: Gantt Chart visualising the steps to be taken in order to reach the main research objective.

A.9 Conclusions

Demand forecasting is one of the most essential concerns of inventory management in the repair and overhaul industry. However, forecasting spare parts demand is challenging as the demand tends to be irregular with a great amount of zero-demand periods. As a result, it makes sense to consider demand forecasting using all available information on the so-called installed base, and to find relevant explanatory variables from this installed base information. However, a rigorous literature study has shown that research on causal forecasting of aircraft spare parts demand is very scarce, and the state-of-the-art method only makes use of time-series techniques. Use of variables from the installed base information like age of fleet, age of components, and environmental impact for spare part demand forecasting is very limited in the body of literature, which therefore has been identified as the main research gap and has led to the following research question: *How can installed base information be leveraged for spare part demand forecasting?* The steps to be undertaken in order to answer this question are a typical example of a data science process, with first data retrieval and preprocessing, followed by exploratory data analysis and predictive model building and ending with verification and validation. This is an iterative process, which has been taken into account in the project planning. A Gantt chart has also been constructed, indicating when the work packages are to be completed and when certain milestones are to be achieved. This will help guiding the researcher in successfully carrying out the project and accomplishing the research objective. The project will be performed at KLM Engineering & Maintenance, where a case study will be formulated of which the component scope is an important factor. Based on availability, reliability and added value of performance increase of the forecast, a convenient sample of the total pool of components will be extracted. Installation and removal data of the sample of components will be analysed, which include component specific variables like age and maintenance history, operation exposure factors and operating environment variables like natural environment and aircraft age. The relevance of establishing relationships with aforementioned variables and spare part demand is a potential in improvement of forecasting performance with respect to state of the art methods and a quantified relationship between individual factors and demand. This could lead to more optimal stock levels, decreased inventory costs, less borrowed parts and more fair and profitable flight hour based contract prices for new customers.

Appendix B

Literature Study

Previously graded under AE4020

B.1 Executive summary

Demand forecasting is one of the most essential concerns of inventory management in the repair and overhaul industry. Forecasts form the basis for the planning of inventory levels, and the high cost of modern aircraft and the expense of such repairable spares as aircraft components and avionics constitute a large part of the total investment of many airline operators. However, forecasting spare parts demand is challenging as the demand tends to be irregular with a great amount of zero-demand periods.

This so-called intermittent demand is mostly forecasted using solely historic demand numbers, yet the methods differ from the classical time-series techniques due to the number of zero demand values and being data of counts. Most state-of-the-art time-series forecasting techniques are not well capable of capturing the intermittency of the data, like weighted averages and regression based forecasting methods. Artificial Neural Networks, and especially the most simple 3-layer perceptron have shown potential in capturing intermittent patterns due to their flexible and non-linear nature. They however require a lot of data, which is often not available for intermittent demand patterns, and do not give insight in the demand generating process. The Syntetos-Boylan Approximation seems the most suitable time-series method for forecasting intermittent demand patterns, as it is theoretically more sound than Croston's method and simple and easy to implement. Adding to that, it has been shown in numerous studies to perform equal or better compared to other proposed method's.

A different way to tackle the forecasting problem at hand is to make use of cross-sectional and/or temporal aggregation. By aggregating the data on a different level in the hierarchy of the product or part, or in lower frequency time units, the number of zero demand observations will be less. Given the reduction of zero observations, a far richer arsenal of forecasting methods and models are available to be employed for time-series extrapolation. The applicability of aggregation depends on the situation, and expert opinion is often needed. Furthermore, there is no conclusive solution in literature with regard to the identification of the appropriate time aggregation level, and it remains to be analysed if the demand forecast improves. Therefore, this is identified as research gap.

In spare part demand forecasting literature, causal forecasting methods are very scarce and use of variables from the installed base information like age of fleet, age of components, and environmental impact is missing, which is identified as the main research gap. Even though the idea of causal forecasting methods for spare part demand sounds straightforward, it is however not easily realised in practice as it requires information on causal variables. This installed base information needs to be maintained which is frequently not the case, or the information is unreliable and is scattered throughout different legacy information systems. The few practical applications that have been undertaken to include causal factors are mainly in forms of reliability and survival analysis, and although various studies found that installed base information is relevant for modelling the reliability of component, no direct link to forecasting future spare part demand is made. The performance increase regarding reliability modelling, as found in the scarce literature available, is not translated into intermittent demand forecasting.

Although numerous comparative studies in the literature exist regarding performance between the various state-of-the-art intermittent demand forecasting methods, the performance criteria used differ and the results are often inclusive. The most commonly used per period forecast error is not informative when not combined with other measures for demand series that consist of many zeros and few positive demands. A key error metric which has been used extensively in literature since its introduction is the Mean Absolute Scaled Error, as it effectively scales the errors and does almost never give an undefined number. Nevertheless, a better way of comparing forecasting methods for slow-moving items is to analyse their effect on inventory control parameters and to compare resulting inventory and service level or inventory costs. This however requires a lot of simplifying assumptions or extensive simulations studies, which are time- and cost-inefficient.

Based on the literature study and identified research gaps the following research question has been formulated that will guide the research project.

How can installed base information be leveraged for spare part demand forecasting?

B.2 General Introduction

Demand forecasting is one of the most essential concerns of inventory management in the repair and overhaul industry. Forecasts form the basis for the planning of inventory levels, and the high cost of modern aircraft and the expense of such repairable spares as aircraft components and avionics constitute a large part of the total investment of many airline operators. These parts are critical to operations and their unavailability can lead to excessive down time costs. However, the demand tends to be variable with a great amount of zero values.

Having spare part inventory allows for an immediate substitution in case of repair, postponing the repairing or buying activities only after having restored system's operations, minimizing the downtime of the aircraft. It relies on defining a stock quantity able to cover the demand within the re-supply time, minimizing the inventory costs. Nuclear plants, oil drilling, defence and transportation industry are some industries where spare parts management is gaining an increasing attention as characterized by low failure rate, high inventory and stock-out costs. As a result, it makes sense to consider demand forecasting using all available information on the so-called installed base, and to find relevant explanatory variables from this installed base information.
In this light a MSc. graduation research project is performed at KLM Engineering & Maintenance. At the very start of the project a research objective had been formulated:

To develop a forecasting model for aircraft spare part demand using relevant explanatory variables from installed base information.

A literature study is performed at the start of the research project to assess the current state of the art in academic literature relevant to this objective and to define the research scope based on research gaps in this literature. The literature study will be structured according to the following questions which arise from the research objective.

1. What is the state of the art in time-series forecasting?
2. What is the state of the art in aircraft spare part demand forecasting?
3. What is the state of the art in causal forecasting for spare part demand?
4. How is the performance of state of the art methods compared?

These sub questions give rise to the structure of this work. The structure aims to guide the reader through the process of narrowing the scope and defining the research question, starting from a general forecasting perspective to the specific causal forecasting techniques that are most promising to improve the forecast for spare part demand.

After a small chapter with an introduction into forecasting, the second chapter will cover the different time-series forecasting models that exist. Next, the third chapter will cover the current state of the art in causal forecasting fir spare part demand and identifies the main research gaps. The final chapter will cover the performance criteria used as found in literature, regarding the various state of the art forecasting models. Based on the identified possibilities, the research question and scope are defined.

B.3 An introduction to spare part demand forecasting

The issue of spare parts demand forecasting has been studied for many years, which has resulted in the development of numerous prediction methods and techniques [3]. This chapter will introduce these forecasting methods, and identify their strengths, weaknesses and applicability.

A spare part is an interchangeable part that is kept in an inventory and used when the in-service part is replaced. Forecasting spare parts demand is a basic requirement of spare parts management. Because of the demand characteristics of spare parts, it is very difficult to accurately forecast demand in this area and is therefore a highly ranked challenge by companies in the airline sector [28] [29] [30]. This is due to the nature of demand pattern variation in the airline sector which has many time zero-demand periods and the demand appears at random intervals.

A demand forecast can be defined as company's best estimate of what demand will be in the future, given a set of assumptions [31]. Different forecasting distinctions are found in literature, e.g. Hu et al. [6] categorizes forecasting approaches into three groups according to the type of forecasting technique and where in the lifecycle process it can be used: time-series, reliability and judgmentally based forecasting. However these groups fall into two general forecasting categories: qualitative and quantitative [32]. An overview of the different forecasting categories can be found in Figure B.1.

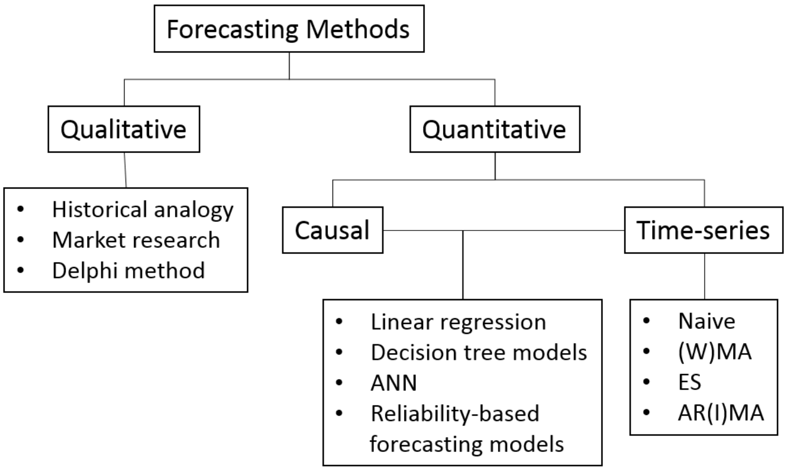


Figure B.1: Overview of the different forecasting categories and techniques. Adjusted figure from [33], see critical remark below.

A critical remark is deserved regarding the literature about causal forecasting methods. Often it is stated that (multiple) linear regression is a causal forecasting technique, however autoregressive models like ARIMA are in essence linear regression models which uses past values of the predictor as features, and are therefore classified as time-series forecasting. The difference between time-series forecasting and causal forecasting lies solely in the fact that time-series features do not *cause* demand to go up or down, whereas for causal features this is assumed to be the case. The same arguments hold for ANN, a machine learning technique which can be seen as a time-series or causal forecasting (or even as a classification [34] or unsupervised learning [35], linear or non-linear) method depending on the inputs to the model. More about causation will be explained in the section on quantitative forecasting methods.

Qualitative forecasting methods

A qualitative forecasting method is an approach to forecasting in which human judgement is used. Examples include market research and expertise and the Delphi Method. The underlying similarity is that it requires people with some knowledge of the products and markets developing the forecasts.

There are three general conditions in which judgemental forecasting is used: (i) there are no available data, so that statistical methods are not applicable and judgemental forecasting is the only feasible approach; (ii) data are available, statistical forecasts are generated, and these are then adjusted using judgement; and (iii) data are available and statistical and judgemental forecasts are generated independently and then combined [32].

Advantages of qualitative forecasting methods are that they can be used when data is either not available or scarce. Examples are sales of a new product or implications of long term changes in markets, environment and technology. The evident downside is the subjectivity of the experts at time and the inability to reproduce forecasts by other forecasters.

Since the research objective as stated in Chapter B.2 implies usage of data and the sufficient amount of data available, qualitative forecasting techniques will not be considered further in this study and elaboration on specific methods is omitted.

Quantitative forecasting methods

Quantitative forecasting methods are approaches where numerical information about the past is used in order to predict future demand. Unlike qualitative methods, these methods are objective; once the underlying model or technique has been chosen, the corresponding forecasts are determined automatically and hence they are fully reproducible by any forecaster. The weakness however is the need for data and the danger of using unreliable and unclean data sets and sources. Furthermore, the assumption is made that the underlying model does not change over time.

As depicted in Figure B.1 there are mainly two groups of quantitative forecasting techniques: causal and time-series. Each has a wide range of methods, often developed within specific disciplines for specific purposes. These will be elaborated upon in separate chapters.

Time-series methods

A time series is defined as a time-ordered sequence of observations taken at regular intervals (e.g., hourly, daily, weekly, monthly, quarterly, annually) [5]. In literature, with time-series forecasting is meant forecasting using a single time-series of the target variable, also known as a univariate time series. This is different from time-series where multiple variables are stored over time (multivariate time-series) and where other variables than the target variable are used for forecasting. It is based on the assumption that future values of the target variable can be estimated from past values of this same variable. By discovering patterns in the past values of the variable it extrapolates these into the future and uses it to predict future values of the variable of interest.

The main strength of time-series forecasting is that it only requires historical data of one variable. The methods are easy to implement and validate on historical data. Furthermore, substantial amount of established theory can be found in literature. It is particularly useful when there is a lack of a satisfactory explanatory model. The main weakness of time-series forecasting is that it heavily relies on the assumption that the discovered patterns in historical data will continue in the future. Nonetheless, time-series methods is the current state of the art for forecasting aircraft spare parts demand.

Causal methods

As explained at the start of this chapter, causal forecasting methods are the scope of this research. Causal forecasting methods are a way of estimating future demand by finding a relation between explanatory variables and spare parts demand [6]. The motivation for such causal or econometric models is the assumption that the variable to be forecast, the dependent variable, has cause-and-effect relationship with one or more other (independent) variables. The steps involved in generating and choosing such a causal forecasting method will be explained in detail in Chapter B.5. The function of this section is to compare its strengths and weaknesses with time-series methods.

The main strength of causal methods is that it has explanatory power; it is possible to evaluate impact of changes in other variables than the target variable itself. This results in a better understanding of the relationships among variables. The main weakness of causal methods is that it assumes that a historical relationship between the dependent and the independent variables will remain valid in the future. Furthermore, it requires historical data on all variables of the model. These other independent variables might also need to be predicted if future values of these variables are needed.

As mentioned at the beginning of this chapter, the distinction generally made in literature between causal forecasting methods and time-series forecasting methods is that time-series forecasting techniques only uses past data of the target variable as input for the forecast. Features created from this time-series can never have a direct cause-and-effect relation with future values. An example of the inter-dependencies between time-series features, causal features and predictor is shown in Figure B.2.

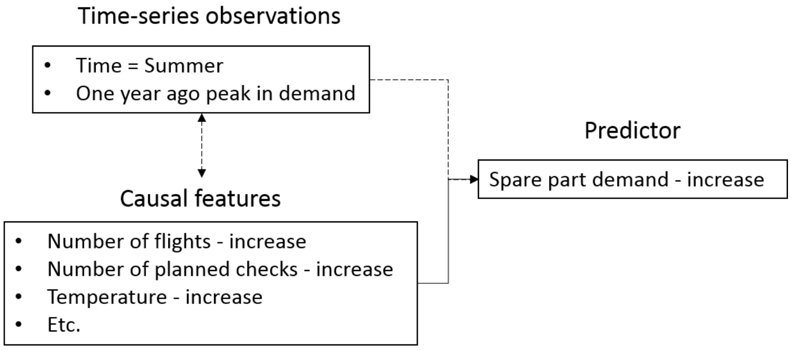

Time-series observations
• Time = Summer
• One year ago peak in demand
Predictor
Spare part demand - increase
Causal features
• Number of flights - increase
• Number of planned checks - increase
• Temperature - increase
• Etc.

Figure B.2: Example of inter-dependencies between time-series features, causal features and predictor. The full arrow indicates causation, the dashed arrow indicates correlation.

Caution should be taken with classifying methods as *causal*, as often *correlation* is found between independent and dependent variables but correlation does not imply causation. Establishing a causal relationship between two variables is actually one of the biggest statistical challenges from both a theoretical and practical perspective [7]. Guyin, Statnikov and Aliferis [36] argue that causal forecasting is to predict the consequences of given actions, also called interventions, manipulations or experiments instead of observations. They state that observations imply no manipulation on the system under study whereas actions introduce a disruption in the natural functioning of the system. This is why it is crucial to understand the data-generating methods and distribution.

Demand pattern classification

Different spare parts are associated with different underlying demand patterns. In literature, these demand pattern classes are mainly used for choosing the extrapolative time-series based demand forecasting method as it is found that the best time-series model choice is dependent on these patterns [1] [11]. There are 13 contributions about this kind of classification published to date and for the interested reader, a description of the evolution of these demand pattern classification methods from these 13 papers can be found in the paper by Boylan et al. [6].

The state-of-the-art method is the method defined by Syntetos et al. [8] and is tested and validated in multiple succeeding studies like [9] and [10]. This method classifies demand patterns into four categories: intermittent, slow moving, erratic and lumpy demand. It uses the average demand interval (*ADI*) and squared coefficient of variation of demand sizes (CV^2). The ADI indicates the average time between occurrence of subsequent demands in the historical demand data. CV^2 is the squared ratio of the standard deviation to the mean of the demand, and shows the extent of demand size variability. An overview of this demand pattern classification with corresponding cut-off values of the two variables ADI and CV^2 is shown in Figure B.3.

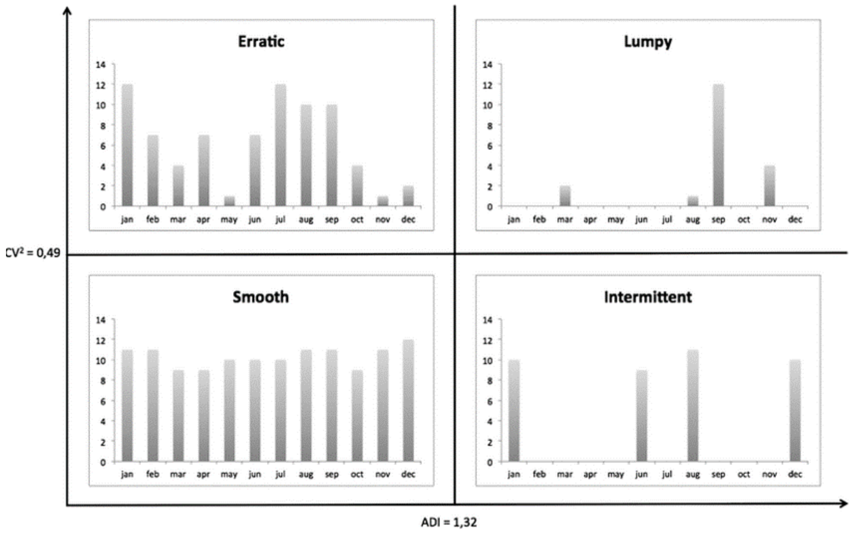


Figure B.3: Demand pattern classification scheme as introduced by Syntetos et al (2005) with examples of the categories. [37]

The benefit of using these demand pattern categories for time-series forecasting techniques has already been stated. For causal forecasting the differences in demand pattern categories can also influence the model decision. For example, a very intermittent demand pattern could result in a causal model which predicts if there is going to be demand or not at a certain period (binary) instead of a prediction of the demand amount itself. In Chapter B.5 these difference will be discussed more in detail.

In general, the classification and especially the variables ADI and CV^2 give the researcher a better understanding of the forecasting problem at hand and enables comparison with other methods per category. However, by manipulating the data and reformulating the forecasting model, the demand pattern changes and therefore also the ADI and CV^2 values. These techniques, some specific for aircraft spare parts demand, are discussed next.

Temporal Aggregation

Temporal aggregation is a method which aggregates demand in lower frequency time units, for example from daily to weekly. In this lower frequency time unit the number of zero demand observations will be less (or equal if there are no zero demand observations in higher frequency), which means the ADI will decrease. Given the reduction of zero observations, a far richer arsenal of forecasting methods and models are available to be employed for time-series extrapolation. Furthermore, empirical studies show that temporal aggregation also generally reduces demand volatility meaning CV^2 is reduced [16] [18]. This is however not always the case and no theory exists in current literature which identifies when and when not temporal aggregation leads to decrease in CV^2 [17].

There are two forms of temporal aggregation: non-overlapping and overlapping [17]. Non-overlapping temporal aggregation divides the historical information into consecutive non-overlapping blocks of equal length. In overlapping aggregation, some observations in the higher frequency time unit are used in multiple lower frequency time unit bins. This means the data are actually moving sub-totals of demand history.

The strength of temporal aggregation is the potentially improved forecast accuracy associated with the uncertainty reduction. This should be weighted against the weakness being a considerable reduction of the number of periods; e.g. from 21 daily demands to 3 weekly demands. This reduction in sample size can result in natural loss of information, especially for short demand histories and non-overlapping aggregation. In the contrary, working at a level that is too granular may present noisy data that is difficult to model. Therefore initial analysis and/or expert opinion is needed, as there is no conclusive solution in literature with regards to the identification of the appropriate time aggregation level. Nikolopoulos et al. [18] do recommend a heuristic that is meaningful for inventory management: aggregate to the level that corresponds to the lead time plus review period. Taking this all into account, temporal aggregation requires further research which is why it is identified as one of the most important areas in a service parts forecasting context [38].

Cross-sectional aggregation

Cross-sectional aggregation is a form of aggregation of data in which demands across items or customers (locations) is summed and all items are reported for the same time periods [17]. One example can be to forecast on a different level in the hierarchy of the product or part. There are multiple different hierarchy levels in literature and corresponding naming (piece parts, subcomponents, subassemblies, assemblies, product family). By aggregating specific items, the forecasting problem changes from forecasting an individual item to the group of items which therefore gives different values for ADI and CV^2.

Two commonly used approaches in practice and research start from opposite ends of this hierarchy to generate forecasts for all series: bottom-up forecasting and top-down forecasting [39]. In bottom-up forecasting, base forecasts are generated for product demand at the lowest level in the hierarchy [40]. These are then aggregated to determine forecasts at higher hierarchical levels. Top-down forecasting is the opposite, in which aggregated demand forecasts are disaggregated downwards to determine forecasts at lower levels in the hierarchy [41]. Discussions remain largely inconclusive however as to which of those two methods performs better under which situation and Babai et al. [17] argue that expanding the empirical knowledge base in this area would be of a great benefit for real world practices.

Component pooling is a natural example of cross-sectional aggregation over customers, as multiple airlines have one pool of components from which they can get spare parts. From the pool-manager's perspective, the total demand for a part is important as it determines inventory levels, however it might be beneficial to forecast demand for all customers separately and then sum up these demands. The benefit of pooling is in essence economy of scale; multiple operators are supported with less total capital than if each operator owned their own parts [42]. The downside however is a general increase in lead times.

Advantages of cross-sectional aggregation is a potential decrease in demand uncertainty due to increase in sample size. An example given by Babai et al. [17] is about ice cream: brands of ice cream will have a similar seasonality with a summer peak, which may not be easily detected for low-volume flavors but can be estimated at a group level and applied on the product level. Useful information is thus extracted from the aggregate series that would otherwise be potentially lost at the lower hierarchy level due to the shortness of data. The downside however is that the method often requires a qualitative analysis as experts in the field should make the decision if (dis)aggregation of product/customer is possible considering he (dis)similarities between the individuals. Similar to temporal aggregation or any aggregation method, it will reduce the sample size with potential loss of information regarding differences between individual items/products. These differences, in econometrics called heterogeneity, could contain valuable cause-and-effect relations regarding demand.

Conclusion

In this chapter the main categories of forecasting techniques are described and compared in terms of strengths and weaknesses. The research project aims at improving aircraft spare part demand by means of Installed Base information, which means both time-series and causal methods will have to be applied and analysed. A well-validated causal model could lead to higher demand prediction accuracy and therefore decrease inventory holding costs. Furthermore, evaluating and quantifying the impact of Installed Base variables could give a better understanding of the demand-drivers for spare parts. Finally, various aggregation methods have been discussed together with the possible benefits regarding forecasting. Therefore the research scope will include determining the most useful aggregating factors for the research problem at hand, by using expert opinion and iteration over aggregation-types and -levels.

B.4 Time-series based forecasting

This chapter will cover time-series forecasting methods found in literature, especially regarding spare part demand forecasting. First the principles regarding time-series analysis and forecasting are given in order to provide a fundamental background that is important for abstracting its applicability to spare part demand data and understanding the literature discussion. The brief explanation of the working principle is based on the work of Adhikari *et al.* [43]. For a more detailed description of the working principles behind time-series forecasting, one is referred to the textbook of Box *et al.* [44].

Principles & State of the art methods

A time series in need of forecasting is non-deterministic in nature, i.e. it is impossible to predict with certainty what will occur in the future. Generally a time series is assumed to follow a certain probability model which describes the joint probability distribution of the random variable. Thus the sequence of observations of the series is actually a sample realisation of the stochastic process that produced it. The mean value of the forecast probability distribution in literature is termed the point forecast of the target variable.

An important concept regarding a stochastic process is the concept of stationarity, which is a form of statistical equilibrium [44]. Often stationarity is assumed as the statistical properties such as mean, variance and autocorrelation structure of a stationary process are time invariant, which therefore reduces the mathematical complexity of the fitted model. As stated by Hipel and McLeod [45], the larger the time span of historical observations, the higher is the probability of non-stationary characteristics.

There are in general three main components which can make the time-series exhibit non-stationary properties and which can be separated from the observed data. These components are: trend, seasonality and cycles. A trend exists when there is a long-term increase or decrease in the data, which does not have to be linear. A seasonal pattern occurs when a time series is affected by seasonal factors such as the time of the year or the day of the week and therefore is of a fixed and known frequency [32]. Factors causing seasonal variations can be: climate and weather conditions, customs, traditional habits, etc. An example was given in Figure B.2 in the introductory chapter about differences between causal and time-series features. The cyclical variation in a time series occurs when the data exhibit rises and falls that are not of a fixed frequency, which repeat in cycles. The duration of a cycle extends over longer period of time, usually two or more years, and the magnitudes of cycles tend to be more variable than the magnitudes of seasonal patterns. Most of the economic and financial time series show some kind of cyclical variation, e.g. the four phases of a business cycle (Prosperity, Decline, Depression and Recovery) [43].

If time series show trend, seasonal or cyclical patterns the stationarity assumption is thus invalid. In such cases, differencing and power transformations are often used to remove the trend and to make the series stationary. Transformations such as logarithms can help to stabilise the variance of a time series. Differencing can help stabilise the mean of a time series by removing changes in the level of a time series, and therefore reducing trend and seasonality.

An other approach requires decomposing the time series into a trend, seasonal and residual component. Forecasts are thereafter made on these separate components and these are combined in order to end up with the forecast of the target variable.

Three main categories of time-series forecasting techniques have been identified by the author while studying the body of literature and are in line with the categories defined by Makridakis and Hibon [46]: weighted averages forecasting, regression-based forecasting and artificial neural networks forecasting. Most techniques are covered in these categories, however a few techniques exist (e.g. support-vector regression) which are not discussed here due to their very limited discussion and application in current literature. The goal of this chapter is to inform on current state-of-the-art time-series forecasting methods rather than to give a complete overview of all techniques available.

Weighted Averages forecasting

For stationary time-series the most common time-series methods found in literature make use of weighted averages over previous observations. Using the *average method*, all future forecasts are equal to a simple average of all the observed data. Hence, the average method assumes that all observations are of equal importance and gives them equal weights when generating forecasts. Using the *naive method*, all forecasts for the future are equal to the last observed value of the series. This means it assumes that the most recent observation is the only important one, and all previous observations provide no information for the future, i.e. all of the weight is given to the last observation. Often something between those two extremes is desired, e.g. it may be sensible to attach larger weights to more recent observations than to observations from the distant past. *Simple Moving Average (SMA)* is an example of such a method, which uses the last n periods of demand with equal weights $1/n$ as a forecast, the other periods have zero-weights. *Weighted Moving Average (WMA)* on the other hand allows for more emphasis to be placed on certain observations by using variable weight scores, where often more weight is put on the most recent data. *Single Exponential Smoothing (SES)* methods, proposed in the late 1950s by Brown [47] and Holt [48], produces forecasts which are weighted averages of past observations with the weights decaying exponentially as the observations come from further in the past. The rate at which the weights decrease is controlled by a smoothing parameter. It has proven through the years to be very useful in many forecasting situations as it generates reliable forecasts quickly and for a wide range of time series. On top of these weights having nice properties, it is not necessary to keep track of each of the weights. The only thing that is needed is the smoothing factor, last period's demand, and last period's forecast, as all past demand data is effectively "stored" in the last period's forecast.
Holt later offered a procedure that does handle trends, *Holt's methods* or also *Double Exponential Smoothing (DES)* as exponential smoothing is also aplied to the trend component. Winters generalised the method to include seasonality, hence the name *Holt-Winters Method* or *Triple Exponential Smoothing (TES)* [49]. A requirement however is at least one complete season's data to determine initial estimates of the seasonal indices.

Regression-based forecasting

A different common approach for modeling univariate time series makes use of linear regression. One example is the *Auto-Regressive (AR)* model, which is simply a linear regression of the current value of the series against one or more prior values of the series. The number of prior values used determines the order of the model. An other example is *Moving Average (MA)* model which, rather than using past values of the forecast variable, uses past forecast errors in a regression-like model. These white noise error terms are assumed to be mutually independent and to come from the same distribution, typically a normal distribution, with zero mean and constant variance. This model should not be mistaken with the previous explained SMA, which is not a regression but a weighted average. Fitting the MA estimates is more complicated than with AR models because the error terms are not observable, meaning that iterative non-linear fitting procedures need to be used in place of linear least squares. MA models also have a less obvious interpretation than AR models.

Combinations of AR and MA has been proposed in literature [44], namely the *Autoregressive Moving Average (ARMA)* and Autoregressive Integrated Moving Average (ARIMA) models. In ARIMA models a non-stationary time series is made stationary by applying finite differencing of the data points. The basic assumption made to implement this model is that the considered time series is linear and follows a particular known statistical distribution, such as the normal distribution. For seasonal time series forecasting, a variation of ARIMA termed *the Seasonal Autoregressive Integrated Moving Average (SARIMA)* model is used. ARIMA model and its different variations are based on the famous Box-Jenkins principle and are therefore also known as the *Box-Jenkins* models. The popularity of the ARIMA model is mainly due to its flexibility to represent several varieties of time series with simplicity and due to its straightforward interpretation. But the severe limitation of these models is the pre-assumed linear form of the associated time series which becomes inadequate in many practical situations. To overcome this drawback, various non-linear stochastic models have been proposed in literature. From an implementation point of view these non-linear models are not so straight-forward and simple as the ARIMA models. Another weakness is the higher probability of overfitting the data due to more parameters having to be trained on the data.

Artificial Neural Networks forecasting

Artificial neural networks (ANNs) approach has been suggested as an alternative technique to the classical time series forecasting techniques and it gained immense popularity in last decade, due to its ability of learning complex nonlinear relationships between the response variable and its predictors without the need for any distribution assumptions. An ANN imitates the intelligence of the human brain into machine: it tries to recognise regularities and patterns in the input data, learns from experience and then provides generalised results based on its previous knowledge [50]. A neural network can be thought of as a network of neurons or nodes which are organised in layers. The node is a computational unit that has one or more weighted input connections, a transfer (or activation) function that combines the inputs, and an output connection. The first layer is formed by the predictors as inputs and the last layer is formed by the forecasts as outputs. The simplest networks do not contain any intermediate or hidden layers which make them equivalent to a regression, where the forecasts are obtained by a combination of the inputs and the weights are the coefficients. The weights are selected in the neural network framework using a learning algorithm that minimises a cost function such as the mean-squared error.

The most widely used ANNs in time-series forecasting problems are *multi-layer perceptrons (MLPs)*, which use lagged values of the target variable as input and have at least one single hidden layer in a feed forward network [51] [52]. A simple three-layer feed forward architecture of an ANN model is diagrammatically depicted in Figure B.4. The feed forward ANN model can be seen as a non-linear autoregressive process in which the network structure and connection weights map the past observations of the time series to the future value.

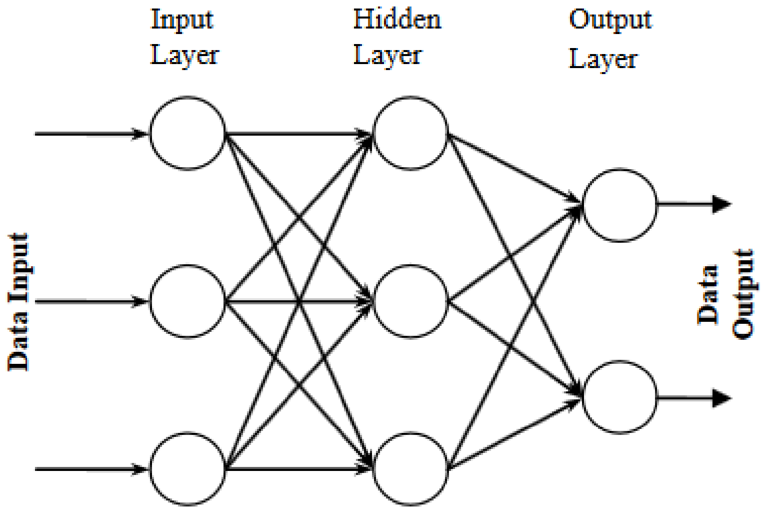


Figure B.4: Example of three-layer feed forward ANN architecture.

The main strength of ANNs is the fact that they are data-driven and self-adaptive in nature; there is no need to specify a particular model form or to make any a priori assumptions about the statistical distribution of the data. For many practical situations where no theoretical guidance is available for an appropriate data generation process this is very useful. A second strength is the inherently non-linear nature of ANNs, which make them more practical and accurate in modeling complex data patterns, as opposed to various traditional linear approaches such as ARIMA methods [52]. As mentioned by Hornik and Stinchcombe [53], ANNs are universal functional approximators, meaning they can approximate any continuous function to any desired accuracy. Another advantage is that ANNs can easily be implemented in parallel architectures (i.e. in multicore processors or systems with GPUs), which reduces drastically the processing time. Lastly, they have been shown to deal well with situations where the input data are erroneous, incomplete or fuzzy [50].

Weaknesses also exist for ANNs, one well-known being the complexity of the methodology; interpretation and understanding of the model is hard, which makes explaining how learning is done from input data difficult. In contrary, weights in a regression model have simple statistical meaning. Because neural networks are not based on a well-defined stochastic model, it is not straightforward to derive prediction intervals for the resultant forecasts. Prediction intervals have to be computed using simulation where future sample paths are generated using bootstrapped residuals. Training ANNs generally requires more data compared to training a regression model because of the higher number of parameters. Furthermore, there are many design decisions that have to be made, from the number of layers to the number of nodes in each layer to the activation functions. There are no generic rules for fine-tuning the architecture in order to achieve the best performance. An inadequate or large number of network parameters may lead to overtraining of data, which means cross-validation is a necessity as there is no theoretical guidance available.

Time-series forecasting methods for aircraft spare part demand

Focusing on time-series forecasting methods for aircraft spare part demand, the body of literature is smaller due to being only a niche market. First, the important differences with the classical time-series methods will be discussed after which the most important studies will be reviewed and analysed for their value for this research project.

The traditional time-series methods were designed for fast-moving items, meaning time-series with either smooth or erratic demand patterns. However, most aircraft spare parts have demand patterns which are intermittent or lumpy. This gives rise to two major differences. First of all, intermittent demand series, seen through a decomposition lens, consist of two stochastic time series instead of one: sizes and inter-demand intervals. The latter stochastic process is not taken into account in the traditional methods. For example, SES is known to perform poorly in forecasting intermittent demand due to an upward bias in the forecast in the period directly after a non-zero demand.
A second major difference is that forecasting demand means forecasting a time-series of counts, i.e. the sample space contains only non-negative integers. With the weighted average methods or the time-series regression models you generally end up with a fractional number. For fast-moving items this rarely matters provided the counts are sufficiently large, as then the difference between a continuous sample space and the discrete sample space has no perceivable effect on the forecasts and simply rounding to the nearest integer would do the trick. However, in the case of slow-moving items the demand data contains small counts (0,1,2,) and a more appropriate forecasting model for a sample space of non-negative integers might be needed.

Croston and Syntetos-Boylan

The standard forecasting method for intermittent demand items is considered to be *Croston's (CR)* method as presented in 1972 [54]. It separately smooths the inter-demand interval and nonzero demands via exponential smoothing with the same smoothing parameter being for both cases, but updates both only when there is nonzero demand. The forecast of demand is then the ratio of the forecasts of the non-zero demand and the inter-demand interval, which in fact is more a demand rate forecast. For example if the forecasted non-zero demand is 4 and the forecasted inter-demand interval is 10, the CR

forecast will be 4/10 = 0.4. This means eventually over the 10 periods of the forecasted interval a total of 4 units of demand is expected. The true timing of the demand event within the predicted interval is unknown and hard to predict, therefore generally the demand is evenly distributed over the expected interval periods. A weakness of the method is the strong assumption of stationary, identically, independently distributed series of demand sizes and demand intervals. The method therefore cannot deal with trend or seasonality. Furthermore, CR assumes that the demand size and the inter-demand intervals are independent, however as Hyndman and Shenstone [55] argue this has not been proven to be true and in many cases it is taken for granted without testing the assumption on the available data. Furthermore, Syntetos and Boylan [15] showed that Croston's estimator is biased. They proposed a modification by multiplying the mean demand estimator by a specific factor, and called the method the *Syntetos–Boylan approximation (SBA)* [16]. It demonstrated improved accuracy, however the strong assumptions of CR remain for SBA. Both methods are incapable of computing prediction intervals as both methods have no underlying stochastic model. Additionally, both CR and SBA do not address to problem of forecasting a time-series of counts, as both generate non-integer forecast values. This might not result in significant problems, as Petropoulos *et al* [56] empirically examined the impact of rounding the final point forecasts derived from these intermittent demand methods and found that rounding resulted in better accuracy levels (up to 2%) while at the same time no deterioration in terms of bias is recorded.
The strength of CR and SBA is the practical usefulness as they are relatively easy to implement and have been shown empirically to outperform conventional methods [15]. The comparison of methods however depends on the performance metrics and criteria used. The comparison of all the performance metrics found in literature deserves a separate chapter, and one is referred to Chapter B.6. This section will continue by describing spare parts demand forecasting methods with more emphasis on working principle differences.

Regression-based forecasting

As discussed before, for intermittent demand an ARMA process is an inappropriate model since it allows values that are not non-negative integers. An adjustment to the model is needed in order add this constraint. One idea quite commonly used is the *Discrete ARMA (DARMA)* models developed by Jacobs and Lewis [57], which take a random choice between autoregressive and moving-average terms using independent Bernoulli random variables instead of the weighted average of the two quantities. The main disadvantage of the DARMA models is that the process will generally contain many runs of a constant value, especially so when the serial correlation is high. Another regression-based class of models for count data called *integer-valued autoregressive moving average models (INARMA)* overcome these problems. These models were originally introduced in the 1980s [58] [59] and are analogous to to earlier discussed ARMA models. It replaces the scalar multiplication in usual ARMA models by the probabilistic operation of binomial thinning. A weakness of adding the probabilistic operation is the increased complexity of the method. It remains to be researched if this outweighs ARMA methods with forecast values rounded to nearest integers. Additionally, ARIMA requires a lot of data to be effective. An advantage of the regression-based models is the convenience and ease of adding other explanatory variables in case of multivariate time-series.

Bootstrap-based approach

Bootstrapping has been proposed by Willemain *et al.* [60] as a non-parametric method to forecast intermittent demand. Bootstrapping is a statistical technique involving random sampling with replacement, and the goal is to simulate an entire demand distribution during lead time. For demand forecasting it is applied on previous observations of non-zero demand and a jittering process is used in order to avoid making forecasts that can only take the same values as have previously occurred. In order to model autocorrelation that might be present in the demand, a two-stage Markov Chain model is used with the states corresponding to zero and non-zero demand observations. The main advantage of bootstrapping is that (the mean and variance of) the lead time demand distribution is forecasted directly by repeated sampling from realised demands, therefore also resulting in integer values for forecasts. It only assumes that demand is stationary. Because it is a non-parametric approach that does not rely upon any underlying distributional assumption, it might fit the true demand size distribution better compared to any standard theoretical distribution assumed by parametric methods, and therefore may improve stock control. Accordingly, Babai *et al.* [61] evaluated the effects of forecasting intermittent inventory demands via simple parametric methods and bootstrapping on stock control performance in more than 7,000 demand series. However, they concluded that simple parametric methods perform well and that it is questionable whether bootstrapping is worth the added complexity. Another weakness is the assumption that the underlying demand distribution is not changing over time, which of course does not have to be the case. Finally, there may be little non-zero data to sample from, meaning the method will have difficulties generating a distribution. The jittering however tries to counteract this problem.

Artificial Neural Networks forecasting

Another non-parametric forecasting model which can be used for intermittent demand are the feedforward multilayer perceptrons which have been discussed in Section B.4. Because they are universal approximators, in theory they are able to capture the data generating process of intermittent demand time series. The network allows for interaction between the demand size and the inter-demand intervals of demand events or their lags without the need for expert input.

ANNs have been explored by Gutierrez *et al.* [13] for lumpy demand forecasting applications. They propose a network with three hidden nodes in a single hidden layer and with solely two inputs. The first input is the last observed demand (showing the difference with CR and SBA which use last non-zero demand), the second input is the inter-demand interval between the last two non-zero demand occurrences. The models are trained using the standard back-propagation algorithm. They report that ANNs outperformed CR and SBA methods with different smoothing parameters on a set of 24 time series. These time series contained 967 daily observations, providing a substantial sample for ANNs to train effectively. This research therefore shows potential in using ANN for aircraft spare part demand forecasting. Moreover, the ANNs output a dynamic forecast due to their autoregressive nature. They are able to predict different values for different forecast horizons according to the time series dynamics, in contrast to Croston's method or SBA. Nevertheless, a consistent amount of data is required for setting the estimator and for outperforming conventional statistical methods as Markham and Rakes [14] proofed. Intermittent demand time series have very few observations, especially when only non-zero demand is modelled. The most granular level for aircraft spare part demand forecasting will probably be a monthly level given lead times being in that order of magnitude, meaning e.g. 5 years of data would only result in 60 datapoints. A potential solution could be regularization of the network which can decrease the number of weights, decreasing the problem of small sample size.

Conclusion

In the opinion of this author, the main contributor to the research regarding time-series is the investigation if state of the art time-series methods will be capable of achieving the same superiority with another large real life data set. Furthermore, some time-series methods might be more ideal for incorporating other explanatory variables. Depending on the time-series patterns a time-series method as baseline will be chosen for comparison with causal forecasting methods. Some conclusions can be drawn from studying the body of literature.

For smooth or erratic patterns, exponential smoothing methods have been shown to perform well and their simplicity and easiness of implementation make them ideal for creating quick forecasts. Regression-based models have shown to challenge weighted average methods in performance, but it entails more model parameters and choices in order to optimize the method. The benefit however is that other ex-

planatory variables can easily be added to an ARMA model.
For intermittent and lumpy demand patterns, bootstrap-based methods are disregarded as they have not been well developed in the literature and in comparative studies with the best benchmark methods their performance did not outweigh their complexity. ANNs, the other non-parametric forecasting method, might be useful if a simple multi-layer perceptron can be trained with enough data. Their flexible nature is advantageous for capturing the intermittent demand structure, and other explanatory variables can easily be added to the network. However, they do not reveal any insight on the demand process due to their complex nature. The Syntetos-Boylan Approximation seems the most suitable method for forecasting intermittent demand patterns, as it is theoretically more sound than Croston's method and simple and easy to implement. Adding to that, it has been shown in numerous studies to perform equal or better compared to other proposed method's. One should not forget that the aggregation methods of demand as discussed in Chapter B.2 could lead to different demand patterns and therefore different methods to be most useful.

As discussed in Chapter B.3 time-series methods simply approximate historical patterns and therefore do not aim to explain the structure of the underlying cause-and-effect mechanism in the data. The main research contribution will be about trying to find and use the cause-and-effect mechanism in multivariate time-series for forecasting as will be elaborated upon in the next chapter.

B.5 Causal forecasting methods

As discussed in the previous chapter, many forecasting methods for aircraft spare part demand exist in literature, none of which though utilise the true drivers of demand. In many industrial environments and in particular maintenance and aviation businesses, more diverse historical information is available from databases. This information, in literature also referred to as installed base information, could be useful in combination with causal forecasting methods. This chapter aims to explore literature on installed base driven forecasting methods; how to specifically define the installed base, which information on the installed base to take into account, and how to relate this information to forecasting future spare parts demand. The goal is to identify potential research gaps which can be filled in further research.

Installed base information

There are many definitions found in literature of installed base information [62, 63, 19], but in general it can be defined as all information on the installed base, i.e. the units of a particular component class actually in use at a specific time. The general idea is that components only deteriorate when installed. This information can vary from age and status information of the specific parts and aircraft, to scheduled maintenance information. This information could have a cause-and-effect relationship with demand for the spare parts. Auweraer et al. [3] recently reviewed the literature of such information. They make a distinction between on the one hand the size of the installed base and part failures, which produces the need for unscheduled maintenance actions, and on the other hand information on the maintenance policy, that determines spare part demand for scheduled maintenance. However, other installed base information like the operational use and operating environment are not discussed. They state that it seems intuitive and self-evident that the installed machines containing a certain part impact the future demand for that part, but an explanation of how to take this information into account is absent and no literature is discussed which takes this into account. Therefore, a thorough and critical literature study regarding forecasting spare part demand using installed base information is conducted and described below.

The author of this document has divided the installed base information in four main groups: operational exposure, operating environment, component (reliability), and maintenance policy information. These groups have been defined iteratively during research of available literature and common sense and are solely used for clarity and structure.

Operational exposure

The first and by far most discussed installed base information in literature is about the operational exposure. The idea behind it is simple: the more and the longer components are used, the more components are deteriorating, and a higher demand for spares can be expected. An often discussed example of this information is about the installed base size (also often referred to as just installed base), which is the number of units of a particular component class actually in use at a specific time. Reasons for this is that installed base size is a variable which is relatively easy to manage and forecast, and the obvious relation with spare part demand. Most literature considering installed base size take it into account via reliability base forecasting approaches. Hong et al. [64] take the installed base size into account by considering the sale rate and discard rate of the product. They use these rates in a stochastic model, together with the estimated failure rate and replacement probability of the product, in order to come up with spare part demand forecasts. Another research example is from Kim et al. [65] proposed a set of installed base concepts with associated empirical forecasting methodologies: depending on characteristics of the product, the spare part under concern, and the consumer market, the authors suggest a different installed base size development over time which is used to determine future spare parts demand. These methodologies however use reliability based forecasting and entail various simplifying assumptions, e.g. a constant hazard rate over time. Furthermore, these models serve the consumer-goods industry for which demand patterns are smooth and installed base information is hardly available, something which is less of a prob-

lem for airline MRO's due to their service contracts. The forecast results of the latter paper actually give negative coefficients for the installed base size and in the end these variables are simply removed due to this unexpected relationship.
Dekker et al. [19] discuss cases where installed base information is used in forecasting at four companies and list the issues involved. One of the company cases is about Fokker Services, and an issue described is the very long service period of a plane and consequently the many equipment changes in these planes. They conclude that setting up installed base forecasting system for an already existing installed base is quite challenging, but state that due to installed base forecasting Fokker was able to predict the consequences of changes in ownership of planes and number of planes with service contracts. They however do not mention if and how the real installed base size (part level) was taken into account, and how predictions were tested and validated.

Schraven [66] used installed base size in order to adjust time-series forecasting methods via a steering variable. He assumes that the trends in the evolution of the installed base are more useful for improving the demand forecasts rather than the values of the installed base, as positive or negative trends in installed base likely occur in the case of phase in or phase out of components. The steering variable was therefore calculated by dividing the step-to-step differences of the installed base over time by the largest observed difference. The time series method is then adjusted by multiplying the forecast error from the previous period by this steering variable. This adjusted method is tested on real spare part demand data of the Royal Netherlands air force. The author compared the adjusted forecasts with the actual demand by means of a simple stock level simulation. His results showed that for 75% of the considered items the forecasting accuracy could be improved with his methodology. There are however limitations to Schraven's research. For example, the installed base is determined by using install and removal data, however this means that components which are not removed are not visible to the author. These censored data are not separately taken into account in the research, it is only stated that the model requires an initialisation time in order to learn what the size of the installed base is. Furthermore, the author excludes records for parts in the maintenance database with a high inconsistency percentage, but does not state how many records are removed. This exclusion of data could have detrimental effect on forecasting results. However, his idea of adjusting time-series forecasting via these steering variables is a novelty and could have potential for further research.

A different way to include installed base size which to this author's knowledge has not been used for spare part demand forecasting is normalization of the demand by the installed base. Athanasopoulos and Hyndman [32] refer to this as population adjustment: any data that are affected by population (or installed base) changes can be adjusted to give per-capita data. By creating a model that is trained on the normalized demand, installed base size differences are taken into account and a forecast of the the total demand can be created by multiplying the normalized forecast with the expected installed base size. The requirement however is that the population changes, or installed base size in this case, is known beforehand or can be predicted with reasonable accuracy. Furthermore, if the installed base size would be the only explanatory variable used, the method would assume that the demand is directly proportional to the installed base size which would be only true if all the underlying causal variables remain constant. If other explanatory variables were added which were counts of the installed base, these should be normalized identically to the target variable.

Other factors regarding operational exposure of the installed base is the exposure time and frequency often expressed in flight hours and cycles. The idea is that the length and number of mechanical, thermal and environmental load conditions encountered in flight and during take-off/landing impact reliability of the part and therefore the spare part demand. Ghobbar and Friend [67] stated that the relationship between mean demand and flying hours/cycles is not well understood and that it can differ per component, e.g. for landing gear the number of landings is far more crucial to its reliability compared to the time the aircraft is in the air. Moreover, they state that flying hours understate the actual usage time for some parts, as a substantial amount of time they might be switched on and running while the aircraft is on the ground. Finally, they argue that there is no specific reason to assume that the relationship between flying hours and mean demand goes through the origin, as some failures would inevitably occur even if aircraft are simply kept in a hangar. In their research the causal factors behind intermittence and lumpiness of demand for aircraft spare parts are examined. They were able to show a significant positive correlation between the coefficient of variation of a demand pattern and the operational intensity in flight hours per day, and a significant negative correlation between the average demand interval and this same operational intensity. Thus, they concluded, if the planned flying hours programme increases, the estimation of demand is expected to increase. This assumes a strictly linear relationship between demand and flying hours. They however do not elaborate on how this found relationship can be used in order to increase spare part demand forecasting performance or accuracy.

Schraven [66] incorporated the number of flight hours, like the installed base size as discussed earlier, via a steering variable. He however did not generate a continuous time span sampled output of this variable, but only steady state total values for the time interval defined by the input data file due to limitations of his model complexity. This means the possibilities of analysing relation between component operational exposure and spare part demand is limited.

Operating Environment

Environmental conditions in which equipment is to be operated often have considerable influence on product reliability characteristics [68] [69]. Examples of environmental conditions are temperature, humidity and dust. Ghodrati and Kumar [20] incorporated these factors in a model based on reliability theory in which the failure rates of the items considered are modelled. They show that most research and articles on reliability consider operating time as the only variable when estimating reliability of a system. They use a binary feature for the climatic condition; it is assigned the value -1 when the climatic condition is bad (i.e. high temperature and very humid) and +1 for better conditions. The baseline hazard rate is calculated using the manufacturer's recommendation of mean time to failure, and the binary feature can influence this hazard rate in a proportional hazard model. Via Cox proportional hazard regression, which is a maximum likelihood estimation, an estimate of the effect of climatic condition on the hazard function is calculated. The authors however do not explain how the proportional hazard model is trained other than stating that the estimates were obtained via maximising the likelihood function, and just give the model with coefficients. If this model is trained on the same data as it is tested, then the resulting model performance might be overestimated due to overfitting of the data. They further state that after 1.5 years the company has less downtime regarding the unavailability of spare parts due to a stock increase, but no real performance comparison is done. The research does show the potential in including operating environmental factor in the analysis, as significant values for the operating environment variable is found. Similar research done on influence of environmental factors on the hazard rate via the proportional hazard rate model can be found in the papers of Ghodrati et al. [21] and Barabadi [22].
Han et al. [70] uses sequential association rules in order to discover interesting relations between variables in failure data of four types of aircraft. Examples of variables include mission, aircraft type, failure date and failure modes. Even though interesting scenarios are analysed for which strong correlation is found between failure modes and mission and season, these findings are not implemented in a causal forecasting model. Furthermore, the link to spare part demand is not discussed. Their results however serve as initial contribution to failure forecasting using installed base information.
Dekker et al. [19] also discussed the impact of the natural environment on spare parts usage and give an example of military tanks operated in the desert areas of Afghanistan and Iraq which needed many more engine filters compared to the same tanks operated in the United States. They also describe a company case for a shipyard for dredging ships in which a model was developed to predict the deterioration rate of certain parts as a function of the state of the ship and the soil to be dredged. They found out that the operational environment is very significantly correlated with the deterioration rate. This implies they used the operational environment in order to adjust the deterioration rate, which therefore gives different

demand forecasts in a reliability based forecasting methodology. The goal of the paper however is not to give information about the methodologies used at the different companies and therefore the reader is solely let with the bigger picture concepts and general problems encountered.

Component (reliability)

Multiple variables at component level are directly linked to the reliability of the component, e.g. the age of the component, the maintenance history of the component and monitoring information from sensors on the component. Monitoring information originates from the concept of condition monitoring, which is a part of condition-based maintenance. Peng et al. [71] define condition-based maintenance as a decision-making strategy to enable real-time diagnosis of impending failures and prognosis of future equipment health, where the decision to perform maintenance is reached by observing the condition of the system. The importance of this for spare parts demand is that one can anticipate demand for spare parts. For this research, this method is discarded as using this information for multiple components for a pool of customers is far too complex and the information is not readily available. For the interested reader, one is referred to [72], [73], [74] or [75].

Information on the age of components could improve the predictability of service parts demand, however, literature on this topic is scarce. One example is given by Deshpande et al. [76], who suggested generating a signal to the inventory system when a part reaches a certain age threshold. The inventory planner then knows the amount of parts that have crossed a certain age threshold. The authors argued that because the failure probability of a part increases with age, the part-age signal and the observed lead time demand are correlated. Consequently, they developed a Part-Age Based Advance Order Policy, where each installed part with an age higher than a certain threshold generates advance demand information and triggers an advance order. The use of these advance orders improved the availability of the spare parts, and reduced inventory costs significantly. Although their approach to match demand and maintenance information is valuable, their work is unfortunately not useful for pure forecasting purposes and does not allow comparison with state-of-the-art forecasting methods.

Maintenance policy

Maintenance information can be relevant when predicting parts needed for planned maintenance. From the planning of the maintenance one can directly derive the need for spare parts, as for most components maintenance packages including the necessary spare parts have been defined. Upgrades and modifications may change the part configuration of a system, resulting in demand for other parts. The maintenance policy applied can differ per airline and has a clear impact on the demand for spare parts. Dekker et al. [19] therefore advise to filter out the parts needed for preventive maintenance from the historical demand in case one has data on the preventive maintenance activities, and fit a separate forecasting model on them. However, they conclude that in practice the automated planning of preventive maintenance is scarce, and observations of parts demand do not show the regularity one would expect from preventive maintenance. Gu et al. [77] supported the view that maintenance activities are the key drivers of spare parts demand. They stress that corrective maintenance (unscheduled) demand differs from preventive maintenance (scheduled) demand, as the former originates from part failures and the latter from preventive maintenance actions. Their analysis only lacks an explanation of the underlying causal variable interrelations which would show if and how these two demand streams differ. For example, one would expect that both scheduled and unscheduled maintenance are correlated via the operational exposure, the age of the component and the maintenance history of the component, and therefore do not have to be considered fully separately when taking these factors into account.
Zhu et al. [78] estimated the demand distribution of a spare part using historical repair data. They determined the failure probability of a part, given that the component which contains the part needs repair. If periodic preventive maintenance is used, providing advance demand information on the component

showed to improve the forecasting of spare parts demand. This result is expected, as advance demand information reduces the stochasticity. They however do also stress the practical issues regarding availability of maintenance plans in some fields.

Another way of implementing maintenance policy information in spare part demand forecasting is described by Hua et al. [79]. They developed a method which forecasts the occurrence of non-zero demand using a logistic regression with explanatory variables, including variables related to the maintenance policy. If the time series of non-zero demands occurrences was strongly autocorrelated, they modelled the autocorrelation using a Markov process. After forecasting the occurrences of non-zero demands, they assigned the size of this demand by sampling from the nonzero values that had appeared in the past. A major weakness of their method is the complexity and the requirement of manual input, making it less suitable for implementation across a pool of components. Furthermore, random sampling from the past demand sizes seems awkward and will yield bad performance for lumpy demand patterns.
Romeijnders et al. [80] proposed a two-step forecasting method, in which they first separately forecast the number of repairs for each type of component and the average number of parts of the studied type needed per repair of that component using ES. In a second step, these forecasts are combined and summed over all components to forecast the total demand for this specific part. The benefit is that it can distinguish whether changes in demand intensity for a part are related to changes in the demand for components or changes in the number of parts needed per repair of a component. Their method however implies that a repair does not always lead to demand for a spare part, whereas at KLM E&M a repair always leads to demand because the unserviceable part is directly replaced by a serviceable unit so the aircraft is ready to fly as soon as possible.

Conclusion

Some conclusions can be drawn from studying the body of literature. Installed base forecasting seems straightforward, it is not that easily realised in practice as much information needs to be maintained and often companies do not have access to it. For this reason, the scientific research on installed base forecasting is limited and the notion is scarce in the airline maintenance and operations literature. Installed base information has mainly been used in reliability models for adjusting the hazard rate, and although various studies found that installed base information is relevant for modelling the reliability of component, no direct link to forecasting future spare part demand is made. The performance increase regarding the reliability model, as found in the scarce literature available, is not translated to intermittent demand forecasting. Various installed base features like the age of the fleet and the amount of cycles have not been studied, while many authors recognise the relationships with reliability and spare part demand. In the opinion of this author, the main contributor to the research could be identifying relevant explanatory variables from the installed base information and using these in a causal forecasting model for spare part demand forecasting. An example could be adding installed base features to ARIMA models.

B.6 Evaluating forecasting performance

This chapter will cover the performance evaluation of forecasting regarding spare part demand. First, an introduction to general performance metrics is given, which explains the state of the art metrics used in general time-series forecasting. The next section will continue with evaluating literature regarding forecasting performance for intermittent demand patterns, elaborates on the problems found, and identifies promising research gaps.

State-of-the-art performance evaluation

The performance of any forecasting method needs to be evaluated by some metric, to measure how closely the forecasted value matches the true value. Intermittent demand series turn out to be unusually tricky to evaluate as many observations of intermittent demand are zero. Typical forecasting accuracy metrics are often either inappropriate or even impossible to apply. This section will elaborate on state-of-the-art metrics regarding point forecast accuracy, inventory performance and prediction distribution.

Point forecast accuracy metrics

Typically, point forecasting methods are compared and assessed by means of an error measure or scoring function, such as the absolute error or the squared error. The individual scores are then averaged over forecast cases, to result in a summary measure of the predictive performance. The metrics presented here are classed as described by Hyndman [25]. Note that numerous other metrics exist, however those rarely occur in literature due to complexity or inferiority with respect to the following metrics.

Scale-dependent metrics Scale-dependent metrics work with the errors, i.e. the difference between the observed demand and the forecasted value. Example scale-dependent metrics include the *Mean Squared Error (MSE)* and the *Mean Absolute Error (MAE)*, or sometimes referred to as *Mean Absolute Deviation (MAD)*. Both of these are widely used in traditional forecasting as these methods are often easy to understand and compute. The main difference between the MSE and MAE is that the MSE is more sensitive to outliers, due to squaring the error instead of taking the norm. Disadvantages are the scale-dependency which means that comparing the MSE or MAE of multiple time-series is meaningless. In addition, if MAE is the only measure to be minimised, then for highly intermittent demand it can often be optimal to forecast simply zero for every period which is of course a biased forecast and of no use in a practical inventory control setting. Teunter and Duncan [12] drew the same conclusion in their study, and therefore advised to use a bias measure in evaluating forecasting performance.

Another scale-dependent metric is the *Geometric Mean Absolute Error (GMAE)* which takes the geometric mean of the absolute errors and is recommended by Syntetos and Boylan [16] for some specific intermittent demand scenarios. A weakness of this method is that any error term equal to zero will send the GMAE to zero, which could happen in an intermittent case quite easily if a zero forecast was made.

Percentage-error metrics Percentage errors measure the error for each period as a percentage of the period's observed demand. The removal of scale-dependency allows for comparison of forecasting methods across multiple data series. Example are the *Mean Absolute Percentage Error (MAPE)*, which is simply defined as the mean over all percentage errors, and the *median absolute percentage error (MdAPE)*, which takes the median. A disadvantage of these metrics is that they puts a heavier penalty on positive errors than on negative errors. This led to the proposal of the *symmetric MAPE (SMAPE)* by Makridakis and Hibon [46], which is computed by taking the mean over the ratio of absolute errors to the sum of the demand and forecast. Even though SMAPE penalises positive and negative forecast errors equally, it does not penalise errors in large forecasts and small forecasts equally.
In an intermittent demand setting, the demand is often zero, which would give undefined values of the percentage error. This alone disqualifies MAPE and MdAPE for use in forecasting intermittent demand.

If the actual demand is zero, the forecast is likely to be close to zero, giving the same problems for SMAPE.

Relative error metrics Relative errors are defined as the ratio of the forecasting error to the error obtained from some other chosen benchmark method. The idea is to compare the performance of the new method against this benchmark to get some measure on how much it improves upon that method. Examples include the *Median Relative Absolute Error (MdRAE)*, which takes the median of the absolute relative error, and the *Geometric Mean Relative Absolute Error (GMRAE)*, which takes the geometric mean instead of the median. The main issue with these methods is choosing the benchmark method. For intermittent demand the benchmark error can often be zero, making these relative error undefined.

Scale-free error metrics The key metric in this category is *Mean Absolute Scaled Error (MASE)*, which is estimated by the ratio of total forecast error divided by the in-sample MAE of the naive forecast method. The idea behind using the in-sample MAE in the denominator is because it is always available and it effectively scales the errors. Compared to relative errors in e.g. MdRAE, MASE is only undefined or infinite when all historical observations are equal. The metric was proposed by Hyndman and Koehler [25] and they recommend it as the measure to use when studying intermittent demand due to its robustness and being scale invariant; it can be used to compare forecasts across data sets with different scales. Furthermore, the mean absolute scaled error can be easily interpreted, as values greater than one indicate that in-sample one-step forecasts from the naive method perform better than the forecast values under consideration. One weakness however is that it seems to suffer the same problem as MAE in that a zero forecast for intermittent demand often proves best.

Percentage better metrics Another way to get a relative performance between two forecasting methods is, instead of quantifying differences in errors, quantifying the number of times one scores better than the other based on a specific accuracy metric. An example is *Percentage Better (PB)* metric, which describes the percentage of observations in which one forecasting method performs better than another forecast method based on the result of a selected accuracy measure. Disadvantages of this metric are that it only does a pairwise comparison, and it does not indicate how much better a method performs in comparison to another method. The *Percentage Best (PBt)* metric generalises all the pairwise PB results in order to make a more general conclusion, but still fails to indicate how much better a method performs in comparison to another method. Also, both metrics still depend on one of the aforementioned accuracy metrics including their weaknesses.

Inventory performance measures

Given that the main purpose behind forecasting intermittent demands is to plan inventory levels, a more compelling analysis examines the forecasting results on the inventory performance. The fact that a particular forecasting method or approach may perform better than one other in terms of forecast accuracy does not necessarily imply that such benefits carry over to the inventory performance. Some well-used inventory performance metric are the realised *customer service level (CSL)*, which is the fraction of replenishment cycles that end with all customer demand being met, and the realised *fill rate (fr)*, which is the fraction of customer demands that is met from stock. The lower both ratios, the poorer the inventory performance. Both metrics can be compared with their target levels. Another way to evaluate inventory performance is looking at the cost-side, consisting mainly of ordering costs, holding costs and backlog costs.

It is hard to get analytic solutions to the implications of forecasting demand on inventory performance, and would require simplifying assumptions. This sometimes makes the models unrepresentative of the real world, affecting their application. In this case, simulation models may be preferred. The downside however is that developing simulation models is time-consuming and costly. The simulation should take into account the stock order policy

Prediction distribution based scores

Since forecasts are often wrong, an estimate of the inaccuracy of the forecast can be just as helpful as the forecast of the expected demand. A more general forecasting metric compared to inventory performance are metrics about the prediction distribution. These measures help in understanding the uncertainty of the forecast and allows for decision-making incorporating variability that is present. Chatfield [81] back in 1992 already argued that, prediction intervals, which can be derived from prediction distributions, deserve much greater attention in forecasting applications. Snyder et al. [82] analyse two distribution metrics. The first measure is the *Prediction Likelihood Score (PLS)* which gives the likelihood that the test set target values come from the model under consideration. The second measure is the *Discrete Rank Probability Score (DRPS)*, which uses the L2-norm to measure the distance between two probability distributions. Both the location and spread of the forecast distribution are taken into account in judging how close the distribution is to the observed value.

Evaluation of overall model fit

While performance evaluation metrics help determine how close the fitted values are to the actual ones, they do not evaluate whether the model properly fits the time series data. This section discusses some metrics and tools regarding the quality of the overall fit of the model.

Residual diagnostics A frequently used way to assess how well the model is able to capture patterns makes use of the residuals or error terms. For an ordinary least squares regression model, one would expect the errors to behave as white noise as they represent what cannot be captured by the model. This means the residuals are uncorrelated and follow a normal distribution, with zero mean (unbiased) and constant variance. If either of these properties are not present, it means that there is room for improvement in the model. If however these improvements can be made depends on the data and the extraction of relevant features.
A quick and easy to implement manner of checking these properties is by visually showing relationships between target variable, predictors and error terms. A common visual inspection makes use of a plot of residuals against predictors or fitted values. If a model is properly fitted, there should be no correlation between residuals and predictors and fitted values. Ideally, the trend is a horizontal straight line without curvature [83]. The plot can help to identify non-linearity, unequal error variances and outliers. Correlation of the error terms can be visually checked by plotting the auto-correlation function and checking if the values of the autocorrelation function for the lags are inside or outside their 95% confidence intervals [44]. A straightforward visual check for the zero-mean property, normality and constant variance properties of the residuals is by evaluating the histogram of the residuals. Normality however is more often checked visually by making use of a *quantile-quantile plot (Q-Q plot)*. In this plot the ordered values of a variable (i.e. the residuals) are compared with quantiles of a specific theoretical distribution (i.e. the normal distribution). If the two distributions match, the points on the plot will form a linear pattern passing through the origin with a unit slope.

Although visually appealing, these graphical methods do not provide objective criteria to test the statistical properties as interpretations are a matter of judgements. Fortunately, statistical tests exists to test the aforementioned properties. Examples are the *lack-of-fit F-test* to see if a variable has relationship with residual [84], *Portmanteau test* to check the hypothesis that residuals are uncorrelated [85] and *normality test* for detecting violation of normality assumption [86].

Bias metrics The presented error measures in Section B.6 do not have the ability to reveal if there is a systematic error present, also known as bias. A common measurement of bias is *cumulated forecast error (CFE)*, which is the cumulated sum of all forecast errors. If the forecast is unbiased the CFE value should be close to zero. Another metric, which is also an inventory performance metric, is the *periods in stock (PIS)*. It measures the total number of periods the forecasted items has spent in fictitious stock or the number of fictitious stock-out periods, where a period is equal to the length of the used time period. A positive number is a sign that the forecasting method tends to overestimate the demand, and vice versa.

Performance evaluation of forecasting methods in literature

As the state-of-the-art metrics regarding point forecast accuracy, inventory performance, prediction distribution and residual diagnostics have been discussed, this section will continue by reviewing how performance measures are used in literature regarding intermittent time-series forecasting.

Eaves and Kingsman [23] examined various forecasting techniques with demand data from the Royal Air Force. The evaluation was done with forecast errors and stock-holding consequences using the the following metrics: MAD, RMSE, MAPE and inventory costs. The results on the evaluation of forecast errors were indefinite, as for different metrics different models were best. Wallstrom and Segerstedt [24] also argued that a single measure of forecast errors mostly does not present the different dimensions of the errors and therefore complementary error measures should be used.
Kourentzes [87] argued that measuring forecasting accuracy for Croston's and its variants is not straightforward, as those methods do not provide an expected demand as a forecast, but rather a demand rate. Therefore, measuring the difference of such a demand rate forecast from the time series data is not meaningful, as they have different units. Instead of comparing this demand rate with the realised demand, he proposes a method which compares it with the in-sample mean demand over time, thus including both zero and non-zero demand periods. This way the demand probability and size are considered, while the timing of the non-zero demand does have an impact on the errors.
The paper of Ghobbar and Friend [1] only reported performance using the MAPE metric. They state that this is due to space limitations and its advantageous performance with intermittent demand, however they do mention that other measures of forecast error may give different results with respect to best performing model. Hong *et al.* [88] took the advantages and disadvantages of each metric into consideration and used in their study MASE and MAE as the major measurements: MAE is used to optimize model parameters for each individual time-series, and MASE is leveraged to compare the overall performance of each type of model across series.

Conclusion

Although numerous comparative studies in the literature exist regarding performance between the various state-of-the-art intermittent demand forecasting methods, the performance criteria used differ and the results are often inclusive. The most commonly used per period forecast error is not informative when not combined with other measures for demand series that consist of many zeros and few positive demands. This chapter has discussed the various forecast error metrics found in literature, and the weaknesses some of those have regarding intermittent demand patterns. A key error metric which has been used extensively in literature since its introduction is the Mean Absolute Scaled Error, as it effectively scales the errors and does almost never give an undefined number. Nevertheless, a better way of comparing forecasting methods for slow-moving items is to analyse their effect on inventory control parameters and to compare resulting inventory and service level or inventory costs. This however requires a lot of simplifying assumptions or extensive simulations studies, which are time- and cost-inefficient.

B.7 Conclusions

Based on the literature study presented in this work, some main conclusions are drawn and some major research gaps are identified. These are used to formulate a research question and define a research scope for the research project.

Different forecasting methods exist, of which in literature two main classes are described: quantitative and qualitative methods. Quantitative methods have the most potential for the initial research question at hand, as those are objective, automatic (after model establishment) and reproducible. Two sub-classes of quantitative methods, namely time-series and causal forecasting methods, appear in literature. The main strength of causal methods is that it has explanatory power; it is possible to evaluate impact of changes in other variables than the target variable itself. This results in a better understanding of the relationships among variables. It however requires the assumption that a historical relationship between the dependent and the independent variables will remain valid in the future. Furthermore, it requires reliable historical data on all variables of the model. Time-series techniques however use only historical data of the forecast variable, and therefore can only mimic the past demand patterns. Furthermore, the ideal time-series technique depends on the demand pattern class. Luckily, using aggregation-tools, this pattern can be adjusted in order to end up with a potentially easier forecast problem to solve.

Tackling intermittent spare part demand patterns differs from the classical time-series techniques due to the number of zero values and being data of counts. Most state-of-the-art forecasting techniques are not well capable of capturing the intermittency of the data, like weighted averages and regression based forecasting methods. Artificial Neural Networks, and especially the most simple 3-layer perceptron have shown potential in capturing intermittent patterns due to their flexible and non-linear nature. They however require a lot of data, which is often not available for intermittent demand patterns, and do not give insight in the demand generating process. The Syntetos-Boylan Approximation seems the most suitable time-series method for forecasting intermittent demand patterns, as it is theoretically more sound than Croston's method and simple and easy to implement. Adding to that, it has been shown in numerous studies to perform equal or better compared to other proposed method's.

The idea of causal forecasting methods for spare part demand sounds straightforward, it is however not easily realised in practice as it requires information on causal variables. This installed base information needs to be maintained and a thorough literature study in the third chapter of this report has shown that installed base forecasting is limited and the notion is scarce in the airline maintenance and operations literature. Practical applications mainly covered reliability models for adjusting the hazard rate, and although various studies found that installed base information is relevant for modelling the reliability of component, no direct link to forecasting future spare part demand is made. The performance increase regarding reliability modelling, as found in the scarce literature available, is not translated into intermittent demand forecasting. Various installed base features like the age of the fleet and the amount of cycles have not been studied, while many authors recognise the relationships with reliability and spare part demand. The lack of the studies that use installed base features for increasing forecasting performance is identified as a major research gap.

Although numerous comparative studies in the literature exist regarding performance between the various state-of-the-art intermittent demand forecasting methods, the performance criteria used differ and the results are often inclusive. The most commonly used per period forecast error is not informative when not combined with other measures for demand series that consist of many zeros and few positive demands. The fourth chapter has elaborated on the various forecast error metrics found in literature, and the weaknesses some of those have regarding intermittent demand patterns. A key error metric which has been used extensively in literature since its introduction is the Mean Absolute Scaled Error, as it effectively scales the errors and does almost never give an undefined number. Nevertheless, a better way of comparing forecasting methods for slow-moving items is to analyse their effect on inventory control parameters and to compare resulting inventory and service level or inventory costs. This however requires a lot of simplifying assumptions or extensive simulations studies, which are time- and cost-inefficient.

To address the research gaps and add to the current state of the art, the following research question has been formulated:

How can installed base information be leveraged for spare part demand forecasting?

This research question gives rise to some important elements of the research. First of all, installed base variables which have potential predictive power for spare part demand have to be identified. Furthermore, a choice of the causal forecasting technique has to be made, and the usage of the installed base features in the prediction model has to be determined. Lastly, the question is defined as a feasibility study framed by the conditions of a real life use case. The research is conducted in collaboration with KLM Engineering & Maintenance, and the scope of the research will specifically be for Boeing 737 aircraft models. After specifying the forecasting model, it needs to be defined, with the knowledge from the literature study on performance, what the criteria for success are that will be used to answer the question. This includes a performance comparison with the current practice.

Appendix C

Data Preprocessing

(Not graded yet)

C.1 Data sources and datasets

Four datasets were used during this research study:

- **stmkrtur_20190331.txt**, a large dataset of installation and removal data of components of the B737, logged for maintenance administration and extracted out of software program dating back to 1977. Contains information on airline, registration number of aircraft in which component has been installed, aircraft model, identification codes of parts, installation and removal dates, installation and removal TSN aircraft, manufacturing date of the part and specification on whether a removal was scheduled or unscheduled.
- **koppen_interannual_1901-2010.tsv**, a Köppen climate classification data set with the Köppen climate type per geographical location grid box of size 0.5° longitude x 0.5° latitude [89].
- **airports.dat**, an airport dataset from the OpenFlights Airports Database. Contains information on location of airports (city, country and geographic coordinates) as well as IATA and ICAO airport codes [90].
- **List_of_hub_airports**, a list of the world's airports with airlines using the airport as hub [91].

The data entries in this report from the installation and removal dataset will be partly anonymised due to confidentiality agreements.

C.2 Cleaning the maintenance dataset

This section will cover the data cleansing of the installation and removal dataset, *stmkrtur_20190331.txt*. Data collection and data entry are error-prone processes. This dataset requires human input, and therefore is prone to sloppiness and typos. Furthermore, errors may originate from the machine mining the data out of program for which the system is not build. Cleaning and preparing the data for use in the modelling phase is extremely important because models will perform better and it costs ample time to fix strange output. As thought in every data science book or course: ”garbage in equals garbage out”.

C.2.1 Standard data cleansing

The following conventional data cleansing methods are generally applicable to datasets.

Duplicates

Duplicate data points are defined as datapoints with exact same entries for `CN`, `VN`, `tail`, `PN`, `SN`, `TSN_IN_FH`, `TSN_OUT_FH`, `date_in`, `date_out` and `reason`. Duplicates percentage equals 4.3%.

Redundant whitespace

Whitespaces tend to be hard to detect but cause errors like other redundant characters would, e.g. mismatch of keys during filters and data merges [26]. Fixing redundant whitespaces is luckily easy enough in most programming languages if you now where and what spaces to clean. They all provide string functions that will remove the leading and trailing whitespaces. For instance, in Python you can use the strip() function to remove leading and trailing spaces. Important variables to remove redundant whitespace are variables identifying a certain object. For example for this research variables for the registration number of the aircraft and the part number and serial number of the component are checked for redundant whitespace.

Data type conversion

Data types are a classification of data that tells the program language interpreter how to use the data. The type defines the operations that can be done on the data and the structure in which the data is stored. Changing the data to the right type results in easier manipulation and programming. Examples found in the maintenance dataset are:

- White spaces or dots (.) indicating empty value: change from string to NaN data type.
- Numbers indicating an object: change from number to string as it is a categorical variable. Exampler are the part numbers of serial numbers of a component.
- Calendar dates as text: change from string to datetime type.

Capital letter conversion

When working with different datasets containing some overlapping categorical variables, it should always be checked if the both datasets have the same capital letter convention for the naming. Often arised problems when neglecting this difference are a mismatch of keys during filters and data merges. Examples applicable to the installation and removal data set are aircraft registration numbers and operator names.

Missing values/NaN-analysis

Many real-world datasets contain missing values, often encoded as NaNs (Not a Number), for various reasons. In case of improper handling of the missing values, inaccurate inference about the data might be drawn. Simply getting rid of the observations that have missing data risks losing data points with valuable information. A better way to handle missing data is to infer those missing values from the existing part of the data. Examples of missing data in the installation and removal dataset, including how they are dealt with, are:

- Missing installation and removal dates, including unknown exposure duration. Solution: remove datapoint (17% of datapoints removed)
- Missing installation date, known removal date: nan for 'Date-ins' maar niet voor date-out 40962 data points, or 6%. Solution: changed by using 8.5 FH/day average and computing number of days using FH from time-on-wing duration. Substract those days from removal date.
- Missing removal date, known installation date. Solution: from data it was clear that these components were still installed on date of data extraction. Fill in date of datadump.
- Missing time-on-wing in flight hours on date of extraction, components still installed. Solution: calculate average flight-hours per day flown of specific aircraft. Use time-on-wing in days and this average in order to estimate time-on-wing in flight-hours.
- Missing operator, known registration number aircraft. Solution: use other datapoints of same registration number in order to create dictionary with operator and corresponding fleet.

C.2.2 Impossible Values and sanity checks

Two very important variables from the install and removal dataset are `AC_TSN_IN` and `AC_TSN_OUT`, as those two variables together also determine the survival time `duration_FH = AC_TSN_OUT - AC_TSN_IN` and the independent variable `PSN_TSN` (sum of all previous `duration_FH` of PSN. Furthermore, `AC_TSN_IN` is also an independent variable in the model. In order to clean the data, outliers are first detected using plots of `AC_TSN_IN` against `date_in` and `AC_TSN_OUT` against `date_out` for a certain aircraft registration number. In case of no errors, all datapoint should approximately lie in a straight line with a positive slope. An example is given for an aircraft in the dataset in Figure C.1. These outliers can be due to an error in the date and/or an error in the registered flight hours.

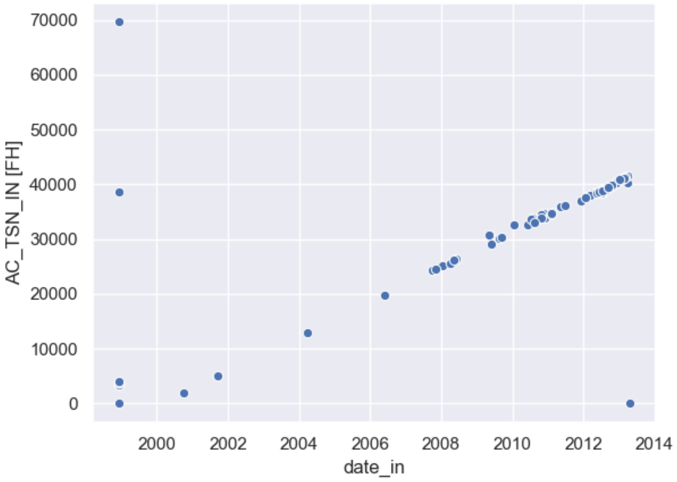


Figure C.1: Outlier detection from the install and removal data regarding flight hour registrations of a specific aircraft

Errors in date entries

There are a few sanity checks which indicate erroneous date values, note that these checks are performed in order of appearance which therefore obviate other sanity checks:

1. The removal date is earlier than the install date.
   - Explanation: Clearly, a part cannot be removed before it is installed. One or both date entries are faulty.
   - Cause: Data entry error and data collection error.
   - Solution: From inspecting examples of this error and looking at the duration of installation in FH, it was found that in most cases the install and removal dates should be switched.
2. The install date is earlier than the manufacturing date of the aircraft.
   - Explanation: the manufacturing date entry of the aircraft is very reliable as it does not have to be filled in manually during installation of removal. Clearly, a part cannot be installed or removed before the existence of the aircraft.
   - Cause: Data entry error.
   - Solution: Use average FH per day and duration of installation to recalculate install date from removal date, or use manufacturing date as installation date.
3. The install and/or removal date is earlier than the manufacturing date of the removed component.
   - Explanation: the manufacturing date entry of the component, although less reliable than the manufacturing date of the aircraft, is still more reliable as it does not have to be filled in manually during installation of removal. A part cannot be installed or removed before it is manufactures.
   - Cause: Data entry error.
   - Solution: Use average FH per day and duration of installation in flight hours to recalculate install date from removal date, or use manufacturing date as installation date.
4. The install and/or removal date is later than the date on which the data extraction has taken place (dump date).
   - Explanation: The dump date is known, and data entries are manually performed during removal of the specific component. A part cannot be installed or removed in the future.

- Cause: Data entry error.
- Solution: mostly already tackled in previous sanity checks. Individual cases of wrong year number entries (e.g. 2068 instead of 1968) are identified using outlier analysis as in Figure C.1, and tackled using average FH per day and duration of installation in flight hours.

Errors in flight hour registrations

There are a few sanity checks which indicate erroneous flight hour values `AC_TSN_IN` or `AC_TSN_OUT`:

1. Negative values of `AC_TSN_IN` or `AC_TSN_OUT`.
   - Explanation: age cannot be negative.
   - Cause: Data entry error and data collection error.
   - Solution: From inspecting examples of this error and looking at the install and removal dates, it was found that the absolute value should be taken as this number of flight hours seemed correct for the duration in days.
2. The installation age of the aircraft is higher than the removal age of the aircraft (`AC_TSN_IN` > `AC_TSN_OUT`).
   - Explanation: Obviously, the age of an object cannot decrease.
   - Cause: data entry error (in opposite fields).
   - Solution: From inspecting examples of this error and looking at the duration of installation in days and cycles, it was found that the install and removal ages should be switched.
3. The install/removal age of the aircraft in flight hours equals 0, while the install/removal age of the aircraft in cycles is positive.
   - Explanation: Obviously, both should be either zero or positive.
   - Cause: data entry error.
   - Solution: Often, the zero-entry is the wrong data entry. Use average flight hours per cycle in order to compute the age in flight hours.
4. The installation age of the aircraft `AC_TSN_IN` equals 0 while the installation date is (several months) later than the manufacturing date of the aircraft.
   - Explanation: In the first place, it is very exceptional to have a component removed which has been installed since the aircraft's existence. It is even more unlikely that the aircraft did not fly for a long period of time after the manufacturing date.
   - Cause: data entry error.
   - Solution: Using other data entries for this aircraft by filtering on its registration number, it can be checked if earlier and later data entries do have a positive value. Use average flight hours per day and install duration in days in order to compute `AC_TSN_IN`.
5. The removal age of the aircraft `AC_TSN_OUT` equals 0 while the removal date is (several months) later than the installation date.
   - Explanation: In the first place, it is very exceptional to have a component removed which has been installed since the aircraft's existence. It is even more unlikely that the aircraft did not fly during the installation period (of several weeks to months).
   - Cause: data entry error.
   - Solution: Using other data entries for this aircraft by filtering on its registration number, it can be checked if earlier and later data entries do have a positive value. Use average flight hours per day and install duration in days in order to compute `AC_TSN_OUT`.
6. The installation age of the aircraft in flight hours equals the removal age of the aircraft in flight hours (`AC_TSN_IN_FH` = `AC_TSN_OUT_FH`), while the installation age of the aircraft in cycles does not equal the removal age of the aircraft in cycles (`AC_TSN_IN_cyc` = `AC_TSN_OUT_cyc`).

  - Explanation: The installation durations expressed in both time units should be either zero or positive.
  - Cause: data entry error.
  - Solution: Often, the zero-duration entry is the wrong data entry. Use average flight hours per cycle in order to compute the age in flight hours `AC_TSN_OUT_FH`.

7. The installation age of the aircraft equals the removal age of the aircraft (`AC_TSN_IN` = `AC_TSN_OUT`), while the removal date is much later than the install date (`date_out` >> `date_in`).

   - Explanation: It is very unlikely that the aircraft did not fly during a period of several weeks to months. Could also be due to an erroneous date entry!
   - Cause: data entry error.
   - Solution: Using other data entries for this aircraft by filtering on its registration number, it can be checked if earlier and later data entries do have a positive value. Use average flight hours per day and install duration in days in order to compute `AC_TSN_OUT`.

8. The removal age of the aircraft in flight hours is much higher than the installation age of the aircraft in flight hours (`AC_TSN_OUT_FH` >> `AC_TSN_IN_FH`), while the install and removal ages in cycles are not far apart. This implies a very high average flight duration ($FH/cycle$).

   - Explanation: There is a max range an aircraft can fly, and therefore also a max flight time. For a B737, a estimated flight duration limit is $9FH/cycle$.
   - Cause: data entry error.
   - Solution: Using other data entries for this aircraft by filtering on its registration number, it can be checked if earlier and later data entries do have a similar values for flight hours and cycles. One of the two is faulty, meaning one of the two ages should be updated using average flight hours per cycle.

9. The removal age of the aircraft is much higher than the installation age of the aircraft (`AC_TSN_OUT` >> `AC_TSN_IN`), while the install and removal dates are not far apart. This implies a very high daily average operational use (FH/day).

   - Explanation: There is a max in average daily operational use; there are 24 hours in a day, and an aircraft needs to be on ground for maintenance meaning the daily operational max usage will be even lower.
   - Cause: data entry error.
   - Solution: Using other data entries for this aircraft by filtering on its registration number, it can be checked if earlier and later data entries do have a positive value. Use average flight hours per day and install duration in days in order to compute `AC_TSN_OUT`.

Note that the same checks hold for installation time and age in cycles.

C.3 Köppen climate classification data set

The Köppen Climate Classification System is the most widely used system for classifying the world's climates. Its categories are based on the annual and monthly averages of temperature and precipitation. The Köppen system recognizes six major climatic types; each type is designated by a capital letter. In addition to the major climate types, each category is further sub-divided based on temperature and precipitation [89]. An overview of the difference climate codes per region is given in Figure C.2.
Two climate classes which are hypothesised to affect component reliability are hot, dusty deserts and humid tropical climates. The former climate is identified by Köppen as codes BWh, BWk and Cwb. The humid climate comprises Köppen codes: Af, Am, Aw, As, Cfa, Cwa, Dsa, Dsb, Dwa, Dwb, Dfa and Dfb. Binary covariates were introduced indicating these two groups.

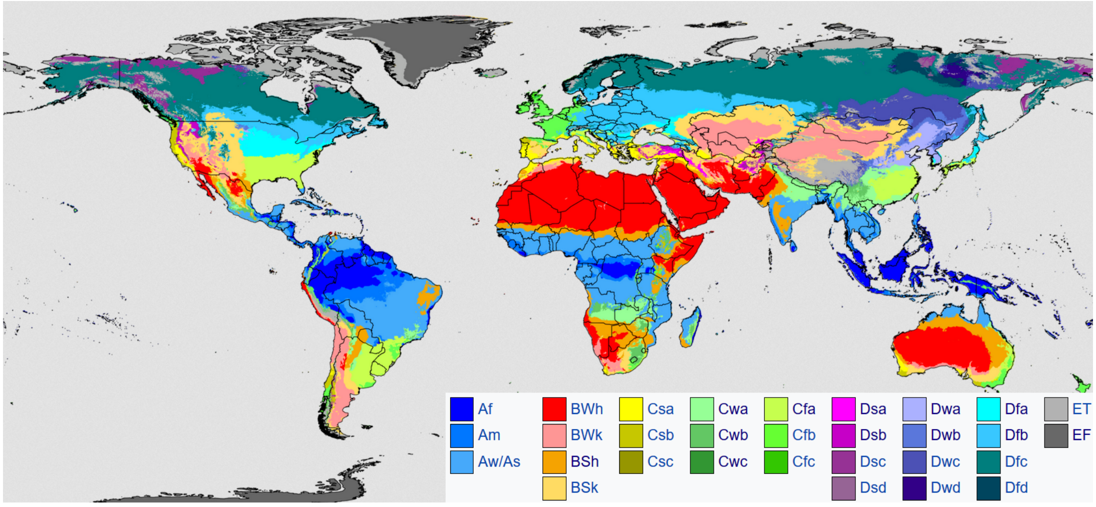


Figure C.2: A Köppen climate classification map (1980-2016) [92]

C.4 Combining datasets

In order to get all covariates in one dataset in order to perform survival analysis regression, the datasets from Section C.1 have to be combined. One merge worth mentioning is the one between the airport location data and the Köppen climate class data. Both sets contain the geographical location expressed in a longitude coordinate and a latitude coordinate. Because the precision differs between sets and the location of an airport does not correspond perfectly to one location of a Köppen class, the nearest Köppen class had to be acquired. This is possible via the Haversine formula, which determines the great-circle distance between two points on a sphere given their longitudes and latitudes [93] via

$$d = 2r \arcsin\left(\sqrt{\sin^2\left(\frac{\phi_2 - \phi_1}{2}\right) + \cos(\phi_1)\cos(\phi_2)\sin^2\left(\frac{\lambda_2 - \lambda_1}{2}\right)}\right) \tag{C.1}$$

where r is the radius of the earth, ϕ_1, ϕ_2 are the latitude coordinates and λ_1, λ_2 are the longitude coordinates of point 1 and 2. The Köppen class with the smallest distance d to the airport location is chosen as the climate class for that specific airport.

Appendix D

Component Scope Trade-off

(Not graded yet)

D.1 Introduction

KLM E&M provides component availability and maintenance for more than two thousand different components for the Boeing 737-NG alone. Because modelling and reporting on all those components is impractical and unclear, the scope of the evaluated components needs to be narrowed. This section explains the reasons why certain components have been included in the research.

D.2 Scoping factors

First and foremost, it is important to identify factors regarding component selection which result in added value for the research study performed. Three main factors are identified: data availability, variability in covariate values and economic impact. Below follows a division of each factor into sub-factors, including reason for importance and priority level.

Data availability

- No engine parts: due to the way how engine parts are logged in the maintenance program, important covariates can not be extracted for these components, leading to biased analysis. An important covariate which misses is the registration number of the aircraft in which the component was installed, leading to unknown operator and unknown climate class.
  Priority: high.
- Number of removals in the dataset: more removals means more datapoints, leading to more precise estimates in the model and more valition samples.
  Priority: high.
- Fast-movers, meaning components with a low MTBR: lower percentage of truncated data as components of new customers are relatively fast removed and logged.
  Priority: medium.

Variability in covariate values

- Joint pool parts: for some components KLM E&M shares a pooled inventory with Boeing from which both KLM E&M and Boeing customers are provided with spare part components. This means a higher number and variety of customers, fleet and natural climate.
  Priority: medium.
- Diverse component types via ATA chapter code: components within the same ATA chapter are hypothesised to be affected similarly by environmental factors and to have similar failure modes. While scoping down the components in the study, it is desirable to still have diversity in component types for illustrating differences in impact by covariates.
  Priority: high.
- Single B737 applicable components: some components can be installed in multiple Boeing aircraft models (747, 787 etc.). This results in a few datapoints with a different aircraft model, too little for estimation model effect.
  Priority: low.
- No component with a soft or hard-time limit: some components have a limit on the amount of flight hours or cycles they may be used before maintenance has to be performed. These removals are scheduled, and therefore produce noise with respect to time-on-wing durations.
  Priority: high.

Economic impact

- Average maintenance cost per flight-hour: because KLM E&M has an income cash flow based on flight hours while costs are made per repair, high cost-drivers are defined by having a high

average maintenance cost per flight-hour. Value for a specific component is computed via division of average shop visit costs by MTBR.
Priority: high.

- Inventory value: components which have a high net value are important due to higher risk of stock obsolescence.
Priority: medium.

D.3 Trade-off & selection

Having the important factors regarding component scope identified, the following step is to select components by means of a trade-off. The first selection is based on the average maintenance cost per flight hour. The top 23 components with the highest cost are selected and shown in Table D.1. Due to the importance of having all covariates available, the scope is narrowed down to non-engine only components. The top twelve components remain, of which Battery (#7) has a hardtime, the ASM (#4) and the Optical Digital Flight Datamanagement Unit (#9) have only a low number of different operators, and the Engine Driven Pump (#10) and the Start Power Unit (#12) have very low number of removals in the dataset. Seven components remain, divided over three different ATA chapters: Navigation, Air Conditioning and Indicating / Recording System. One of each chapter is chosen for diversity considerations, and based on total removals and average repair cost these components are:

- Air Data Inertial Reference Unit (ADIRU)
- Air Cycle Machine (ACM)
- Display Unit

Table D.1: Trade-off table used for component scope

#	Name	Fleet size[a]	Operators[a]	ATA chapter	Hardtime	MTBR	Total removals	Average[a] repair cost/FH	Inventory value	Aircraft models	Engine part
1	Air Data Inertial Reference Unit (ADIRU)	high	high	Navigation	no	18200	1722	$-	$451,000	737-NG	no
2	Air Cycle Machine (ACM)	high	high	Air Conditioning	no	37800	821	$-	$425,000	737-NG	no
3	Pack Temp Control Valve	high	high	Air Conditioning	no	12900	3590	$-	$36,000	737-NG, 737, 737-MAX	no
4	Air Seperation Module (ASM)	low	low	Inert Gas System	no	13300	121	$-	$122,000	737-NG	no
5	Ram Air Actuator	high	high	Air Conditioning	no	11800	2172	$-	$28,000	737-NG, 737	no
6	Multi Purpose Control and Display Unit (MCDU)	high	high	Navigation	no	13300	1743	$-	$125,000	737-NG, 767	no
7	Display Electronic Unit (DEU-I)	medium	high	Indicating / Recording System	no	24100	692	$-	$648,000	737-NG, 787	no
8	Battery	medium	medium	Electrical Power	yes	1600	3640	$-	$19,000	737-NG, 737-MAX, 744	no
9	Optical Digital Flight Datamanagement Unit	low	low	Indicating / Recording System	no	5300	582	$-	$140,000	737-NG, 767	no
10	Engine - Driven Pump	medium	medium	Hydraulic Power	no	999999	390	$-	$72,000	737-NG	no
11	Display Unit	high	high	Indicating / Recording System	no	78500	2235	$-	$489,000	737-NG, 787	no
12	Start Power Unit	medium	medium	Airborne Auxiliary Power	no	999999	208	$-	$403,000	737-NG	no
13	Integrated Drive Generator (IDG)	high	high	Electrical Power	no	16000	1849	$-	$642,000	737-NG	yes
14	Air Turbine Starter	medium	medium	Starting (Power Plant)	no	29700	742	$-	$78,000	737-NG	yes
15	Precooler Control Valve	high	high	Pneumatic	no	5600	3966	$-	$56,000	737-NG, 737, 737-MAX, 787	yes
16	Starter Generator (APU)	medium	high	Airborne Auxiliary Power	no	10700	1391	$-	$437,000	737-NG, 737-MAX, 787	yes
17	Turbine Clearance Control Valve	medium	medium	Power Plant	no	35300	597	$-	$255,000	737-NG, 787	yes
18	Electronic Control Unit	medium	medium	Power Plant	no	29600	1379	$-	$524,000	737-NG, 787	yes
19	Pressure Reg & Shutoff Valve	high	high	Pneumatic	no	13000	2120	$-	$54,000	737-NG, 737	yes
20	Hydro Mechanical Unit (HMU)	medium	medium	Engine Fuel and Control	yes	35500	753	$-	$400,000	737-NG, 787	yes
21	High Stage Valve	low	low	Pneumatic	no	999999	599	$-	$54,000	737-NG, 737	yes
22	High Stage Valve	high	high	Pneumatic	no	19400	1637	$-	$54,000	737-NG, 737, 787	yes
23	Inlet Cowl Termal Anti Icing Valve (TAI)	high	high	Ice And Rain Protection	no	18300	1915	$-	unknown	737-NG	yes

[a] Numbers are left out due to confidentiality restrictions.

Appendix E

Variables affecting Component Reliability

(Not graded yet)

E.1 Introduction

Aircraft operations are very complex in nature, and the stochastic failure process of aircraft component may be influenced by various operational and maintenance factors. This section will describe the hypothesised interrelations between variables from operational and maintenance data and the reliability of the component.

E.2 Variable interrelations

An overview of the hypothesised interrelations between variables from operational and maintenance data and the reliability of the component can be found in Figure E.1. These considered interrelations are based on literature [1, 94, 95, 96] and expert opinion inside KLM E&M. Due to the complexity of aircraft operations, systems and components, it is impossible to catch all interrelations in one diagram and Figure E.1 is therefore not all-encompassing. Its function is to aid the author in the modelling process, and the reader in understanding the line of thoughts and choices made by the author.
The arrows in E.1 indicate the direction of explanation. They do not mean direct cause-and-effect relations, as proving a causal relationship between two variables is actually one of the biggest statistical challenges from both a theoretical and practical perspective [7]. The variables in Figure E.1 are grouped in sets: operator factors, component factors, aircraft factors, preventive maintenance factors, condition based maintenance. The following sections will describe the reasoning behind the links per group.

E.2.1 Operator factors

The operator factors consist of factors which are dependent on the operator of the aircraft in which the considered component is installed. The airline operates from a hub, which geographical location has specific average temperature and humidity levels throughout the year, as well as air pollution and salinity levels. The monthly average temperature and precipitation determine the Köppen climate class, which are used to identify temperate, humid and hot desert climates. Desert climates contain concentrations of dust and sand damaging key components such as compressor blades, air conditioning units and nozzle guide vanes, leading to reductions in component efficiency and reliability [97].
Humidity impacts the reliability of components through steel and metal corrosion, increased damaging micro-organism activity, and worsened composite hygroscopic material features. In the electronic industry, printed wirings get corroded due to presence of high humidity [98]. Airborne salinity, the salt content in the atmosphere, accelerates the corrosion process. The same holds for air pollutants like sulphur dioxide and nitrogen dioxide which play a crucial role in the atmospheric degradation of copper, zinc and aluminium [99].
Factors of the operator that are difficult to quantify are operation skills & standards, technical education of users, storage conditions etc. These factors relate to the intensity of (mis)use of the component.

E.2.2 Component factors

Component factors are those factors related to the physical component identified by the part-serial number. The age of the component, often expressed in operating time (FH or cycles), is related to (electronic) wear over time, which due to imperfect repair might still affect the component reliability. Because in the analysis components are aggregated up to a certain aggregation level, different designs of the same component are present in the data. For example, pre-modification and post-modification designs could differ in reliability characteristics due to physical or software updates as provided by the original equipment manufacturer (OEM).
The maintenance history of the component evidently affects component reliability in case of imperfect repair; a new component will on average have a larger time-on-wing duration than the imperfectly repaired component, ceteris paribus. The arrows between component age and repaired indicate the positive correlation between the two: a previously repaired component has on average been operated more flight

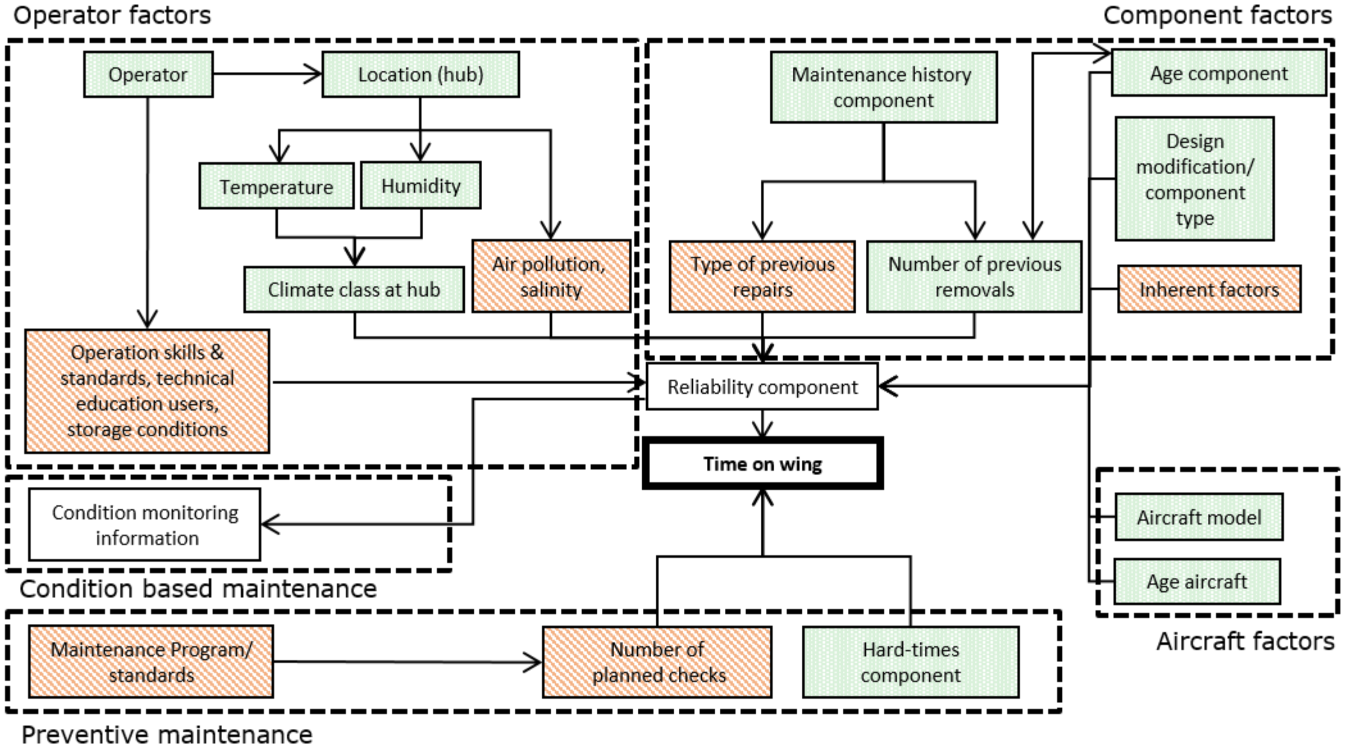


Figure E.1: Predictors for time-on-wing target variable, and their interrelation. The green (full) fields indicate included predictors, the red (downward diagonal) fields indicate covariates which are not controlled for.

hours than a component not yet repaired. Differences in previous repair types, i.e. overhauled vs repaired, affect component reliability differently as more sub-parts of the component are replaced during a more rigorous repair.
There are also inherent factors to the specific component which could affect its reliability characteristics, e.g. the strength and condition of raw materials of which each copy was made, the manufacturing and assembling processes, the quality control to which each copy was exposed, the initial packaging, transportation and storage of each copy prior to operation etc. Unfortunately, it is impossible to control for these factors.

E.2.3 Aircraft factors

The hypothesised factors related to the physical aircraft in which the component is installed are twofold; the age of the aircraft and the specific aircraft model. If the aircraft in which component is mounded is very old, surrounding components may be degraded, there might be more vibration and corrosion which due to imperfect repair over the aircraft's lifetime is still present. Some components can be installed in various aircraft models, for example in the Boeing 737 and Boeing 747. Even though the components are the same, the systems and subsystems in which the component is installed might differ, resulting in different reliability characteristics. One criterion used for determining the component scope was the number of different aircraft model's in which the component could be installed. A single aircraft model (Boeing 737 Next Generation) was preferred in order to leave out this factor.

E.2.4 Preventive maintenance factors

Apart from hard times specified by the OEM, some operators might utilise their own time limit before which the component needs to be removed and checked. These factors do not affect component reliability directly, but however affect the time-on-wing duration of components as most components are preventively removed before failure occurrence. The maintenance policy may specify various maintenance actions depending on states of other components, e.g. some component failure can trigger maintenance necessities on other components. The scope of the components for this research was chosen such that no components had a hard or soft time limit, excluding these preventive maintenance factors for this research.

E.2.5 Condition based maintenance

This block is solely added for giving the reader an idea where condition based maintenance comes into play with respect to component reliability and this research. The condition monitoring data provided by all on-board sensors give information on the reliability of the component (see direction of the arrow in Figure E.1), and in an ideal situation is used in order predict upcoming failure so maintenance can be proactively scheduled.

Appendix F

Modelling Iterations

(Not graded yet)

F.1 Introduction

Building a Cox survival regression model and making it usable in practice means compliance with numerous assumptions. This section will discuss the modelling steps and iterations which were needed in order to get the best possible and valid Cox regression model. Furthermore, it will elaborate on how the Cox model was used in order to compute restricted MTBR-ratios and their confidence intervals. The steps are illustrated for the Air Cycle Machine (ACM).

F.2 Model covariate selection

The performance of the model was evaluated and optimised with the AIC, due to the ability to compare unnested models made for the same outcome on the same data due to its correction for the amount of parameters used. Lower AIC values correspond to a better fit. A stepwise backward elimination procedure was used to drive the covariate selection. This procedure began with the full model and tests the elimination of each covariate using AIC, deleting the variable whose loss gives the lowest AIC score. This is repeated until no further covariates can be deleted without a loss of fit. The results of the first and second step from the full model of the ACM data are given in Table F.1. Note that the ACM in this dataset had only one unique design, and therefore there were no covariates defining the modification design.

Table F.1: Backward elimination scores for covariates of ACM, first step (left) and second step (right)

Covariates	**Df**	**AIC**
PSN_TSN_IN_FH	1	7186.9
repairs	1	7187.4
current model		7188
AC_TSN_IN_FH	1	7191.5
desert	1	7198.4
repaired	1	7199.7
humid	1	7237

Covariates	**Df**	**AIC**
repairs	1	7186.6
current model		7186.9
AC_TSN_IN_FH	1	7189.9
desert	1	7197.3
repaired	1	7197.8
humid	1	7236.4

PSN_TSN_IN_FH and *AC_TSN_IN_FH* denote the time-since-new in flight hours of the part-serial number and aircraft respectively. The terms are listed in order from the greatest AIC reduction to the smallest reduction when deleted. Thus, *PSN_TSN_IN_FH* was deleted in the first step, and *repairs* in the second step. The next iteration did not include any covariate whose elimination would lead to a better fit. The results of the ’final’ model from the model selection procedure are displayed in a forest plot in Figure F.1. In this plot the coefficient estimates together with their 95% confidence interval are illustrated. Note that in Figure F.1 the hazard ratio of the age of the aircraft looks insignificant, however its size is just very small due to the fact that it was based on a one flight hour increase in age.

A common sense check of the hazard ratios was performed in order to see if the covariates indeed show the expected hazard increase or decrease. As visible in Figure F.1 all covariates increased the hazard with respect to the baseline. Ageing of aircraft was hypothesised for components in general to have a negative influence on the time-on-wing. Furthermore, *desert* and *humid* both increased the hazard, which is expected for an air conditioning component. The positive hazard ratio for *repaired* indicates imperfect repair was performed on those components, a very plausible idea.

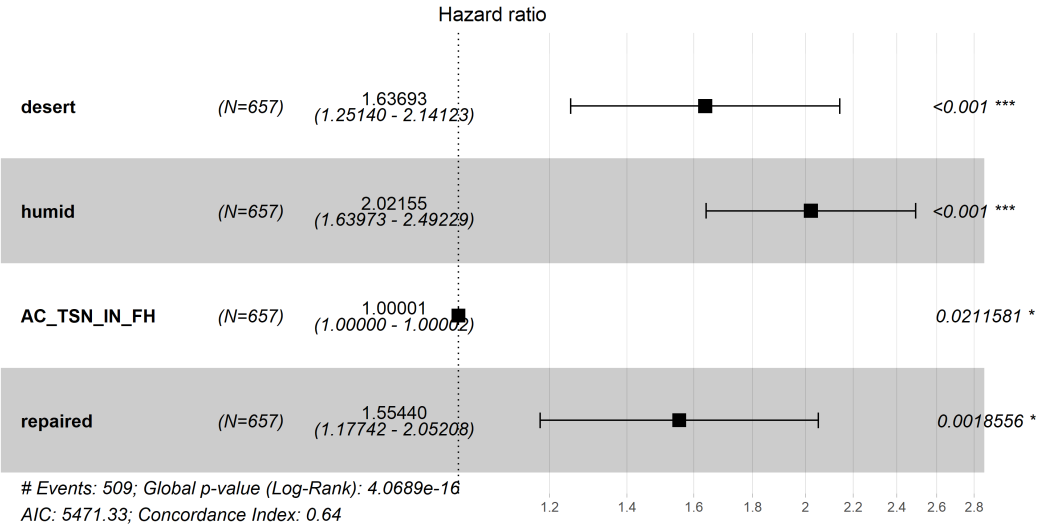


Figure F.1: Forest plot of estimates of hazard ratios for the model after covariate selection, for the ACM data

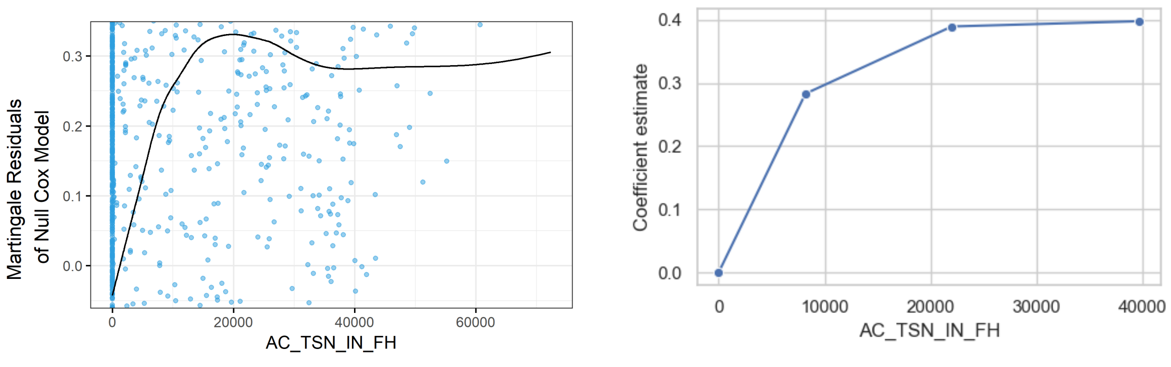
F.3 Functional form of continuous covariates

The next step involved checking if the log-linear relationship assumption between the hazard and continuous variables was valid, and if not, choosing the proper functional form. One way is via plots of martingale residuals of the null model against covariate value. Fitted lines with a lowess curve should be linear to satisfy the assumption. For the aircraft age as covariate for the ACM this plot is given in Figure F.2a.

Another way of checking the proper functional form, which is less prone to multicollinearity between covariates and dependence due to recurrent events, is by discretization of the continuous covariate. Dummy variables were created for non-overlapping intervals of the continuous variable and the model was fitted with these dummy variables. During this step for the ACM, the Cox model was formulated as:

$$h(t|\mathbf{x}) = h_0(t) \cdot e^{\beta_1 \cdot AGE_{0-15000} + \beta_2 \cdot AGE_{15000-30000} + \beta_3 \cdot AGE_{30000-\infty} + \beta_4 \cdot HUMID + \beta_5 \cdot DESERT + \beta_6 \cdot REPAIRED} \tag{F.1}$$

where $AGE_{0-15000}$ denotes a binary indicator with a value of 1 if the age of the aircraft at installation of the component lay between 0 and 15,000 flight hours. $HUMID$, $DESERT$ and $REPAIRED$ are binary variables for a humid hub environment, desert hub environment, and a previously repaired component respectively. A plot of the coefficient estimates against the mean covariate value of the interval should give a linear pattern in order for the assumption to be valid. Figure F.2b gives this plot for the ACM data.



(a) Graph of martingale residuals of null model against aircraft age in flight hours for the ACM

(b) Graph of coefficient estimates of discrete intervals of aircraft age against the mean covariate value of the interval for the ACM

Both plots in Figure C.2 show a linear increase up to an aircraft age of 20,000 flight hours, after which the increase stagnated. A square root or logarithmic form of the age aircraft therefore looked more appropriate. A problem one runs into by changing the functional form is that the variable now does not increase linearly with time anymore. This meant that the time dependent part of the variable did no longer cancel out in the partial likelihood. This can be shown mathematically by comparing the partial likelihood terms in both situations. From the partial likelihood term in the initial situation of a linear form of the aircraft age:

$$\frac{e^{\beta\cdot(AGE+t)}}{\sum_{j\in R_i} e^{\beta\cdot(AGE_j+t)}} = \frac{e^{\beta\cdot AGE}\cdot e^{\beta\cdot t}}{\sum_{j\in R_i} e^{\beta\cdot AGE_j}\cdot e^{\beta\cdot t}} = \frac{e^{\beta\cdot AGE}}{\sum_{j\in R_i} e^{\beta\cdot AGE_j}} \tag{F.2}$$

Changing the functional form of the aircraft age to square root, yields:

$$\frac{e^{\beta\cdot\sqrt{AGE+t}}}{\sum_{j\in R_i} e^{\beta\cdot\sqrt{AGE_j+t}}} \neq \frac{e^{\beta\cdot\sqrt{AGE}}\cdot e^{\beta\cdot\sqrt{t}}}{\sum_{j\in R_i} e^{\beta\cdot\sqrt{AGE_J}}\cdot e^{\beta\cdot\sqrt{t}}} \tag{F.3}$$

Since age changes continuously, to completely capture the effect a very large data set would be needed with one interval per flight hour to match the usual resolution for removal times. In practice this level of resolution was not necessary as the risk did not increase so quickly. Therefore an expanded data set was created with a coarser time grid. This time grid precision was chosen by running the model for various time grid lengths and determining at which length the model estimates stopped changing significantly. Table F.2 gives an overview of the coefficient estimates for five different time grid resolutions when the square root of the aircraft age was used as covariate. The most significant changes in estimates were for *sqrt_AC_TSN_FH* and *repaired*, increasing by more than 35%. The estimates for *desert* and *humid* did only change slightly, up to 5% with respect to no extension of the cox model. From Table F.2 it was chosen to fix the time grid size to 500 flight hours.

Table F.2: Coefficient estimates for five different time grid resolutions for the ACM data

Time grid size	**0**		**3000**		**1000**		**500**		**300**	
Covariates	*coef*	*p*	*coef*	*p*	*coef*	*p*	*coef*	*p*	*coef*	*p*
desert	0.428	4.98E-04	0.408	7.64E-04	0.408	7.60E-04	0.408	7.58E-04	0.408	7.57E-04
humid	0.649	1.77E-10	0.646	2.32E-10	0.647	2.08E-10	0.647	2.06E-10	0.648	2.02E-10
sqrt_AC_TSN_FH	2.46E-03	5.84E-03	5.31E-03	4.95E-07	5.25E-03	2.46E-06	5.27E-03	2.83E-06	5.24E-03	3.82E-06
repaired	0.427	5.58E-03	0.276	5.37E-02	0.302	3.55E-02	0.305	3.37E-02	0.309	3.12E-02

Figure F.3 shows that for the square root form of the aircraft age, the plot of the coefficient estimates against the mean covariate value of newly formed intervals gave a more linear pattern. AIC scores also clearly showed the difference in fit; the extended model with *sqrt_AC_TSN_FH* yields 7169, the model with *AC_TSN_FH* results in an AIC of 7187. The logarithmic form of the aircraft age was also tested, however the square root form resulted in the better fit.

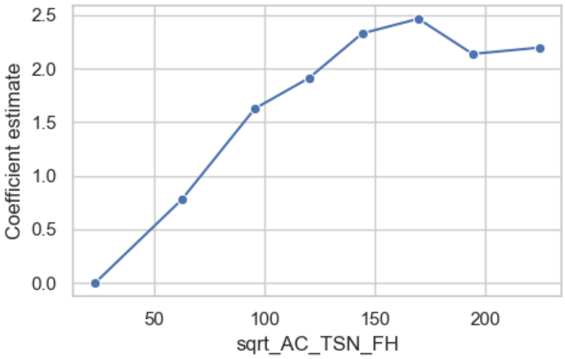


Figure F.3: Coefficient estimates of discrete intervals of square root of aircraft age against the mean covariate value of the interval for the ACM

F.4 Proportional hazards assumption

The other important assumption which had to be checked is the one of proportional hazards. For time-dependent covariates, here *AC_TSN_FH*, this was disregarded due to the fact that the PH assumption is inherently invalid. For the other covariates however this could be tested via use of the scaled Schoenfeld residuals. These residuals could be used in order to approximate $\beta(t)$; if the proportional hazards assumption is true, a fit through the scaled Schoenfeld residuals should form a straight horizontal line. Plots of the scaled Schoenfeld residuals against the transformed time for predictors *humid*, *desert* and *repaired* can be found in Figure F.4.

From graphical inspection of Figure F.4, the assumption of proportional hazards appeared to be supported for all three covariates. However, a statistical test to check if the slope of the fitted line trough the Schoenfeld residuals is significantly different from zero yields a more objective measure. The results of this test are given in Table F.3.

Table F.3: Correlation coefficients between transformed survival time and the scaled Schoenfeld residuals, a chi-square, and the two-sided p-value.

Covariates	ρ	χ^2	**p**
humid	0.066	2.266	0.13
desert	-0.030	0.459	0.50
repaired	-0.075	2.806	0.09

Table F.3 shows that there was weak evidence against the hypothesis that the slope coefficients were equal to zero, meaning the PH assumption was satisfied for all variables.

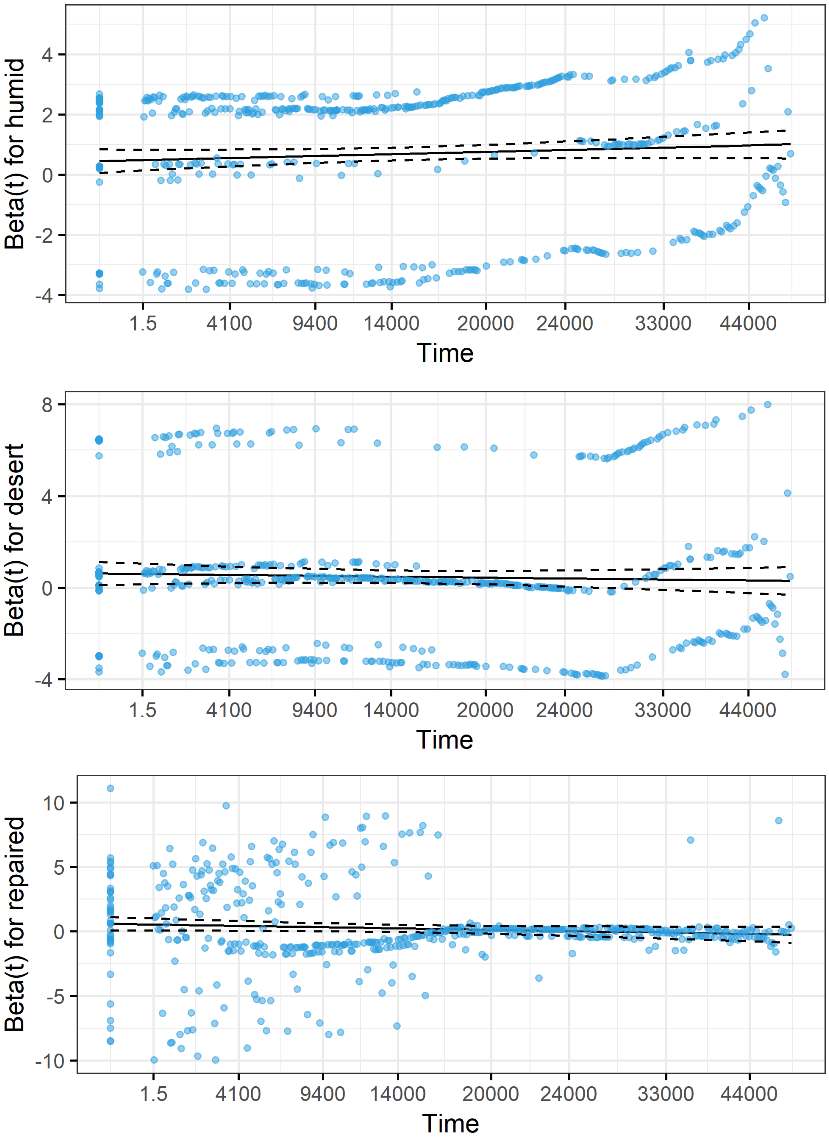


Figure F.4: Graphical test of Proportional Hazards assumption for predictors humid (top), desert (middle) and repaired (bottom) of the ACM data. The blue scatter are the scaled Schoenfeld residuals, the (full) black line is the smoothing spline fit of two degrees to the plot, and the dashed lines represent confidence bands at two standard errors around the fit.

F.5 Restricted MTBR ratio

The finals step of the modelling phase entailed obtaining restricted MTBR ratios for the covariate in the model. In order to compute these ratios, the validation data set was used in order to have a representative distribution of the covariates. The reason for not using the derivation data itself is to decrease computation time and memory requirements, as its size is only 25% of the derivation data. First, for a specific categorical variable, all values in the validation set were set to its baseline value (e.g. temperate hub climate), and survival curves and RMST (restricted by t=70,000) were predicted for each component in the set. Then, all values of this variable were set to one value (e.g. humid hub climate), and again survival curves and RMST were predicted for those component. The ratio between RMST's for each component, were only this one categorical variable had changed value, was taken.

For example, for the *repaired* and *climate* covariates, the mean over the survival curves per covariate value is given in Figure F.5. The ratio of the areas under the curves for two values of the same covariate would result in the MTBR-ratio for that covariate.

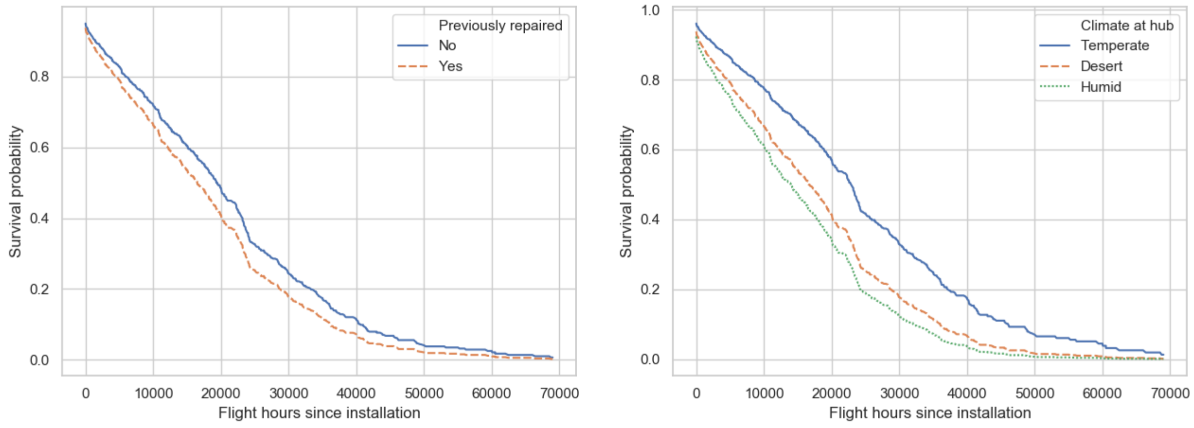


Figure F.5: Adjusted survival curves of the ACM data stratified by covariate repaired (left) and climate (right), based on hypothetical population

Even though standard errors for the survival curves and RMST *for a specific component* can be derived from the Cox model fit, the standard errors when averaging survival curves *over a group of components* is complicated. This is due to the fact that the predicted survival curves are correlated due to their common dependence on the model's coefficient vector β. One feasible solution to derive the the distribution of the ratio of RMST is using the bootstrap method, a resampling method which independently samples with replacement from an existing sample data and performs inference among these resampled data [100]. In this context, it meant fitting the Cox model on hundreds of samples from the original survival data, and using those different fits to compute the RMST-ratio. A histogram of all RMST-ratios approximates the real distribution of this ratio. For predictors *repaired* and *desert*, the distributions of the RMST-ratio derived from Cox model fits on 1,000 bootstrap samples from the derivation data are given in Figure F.6.

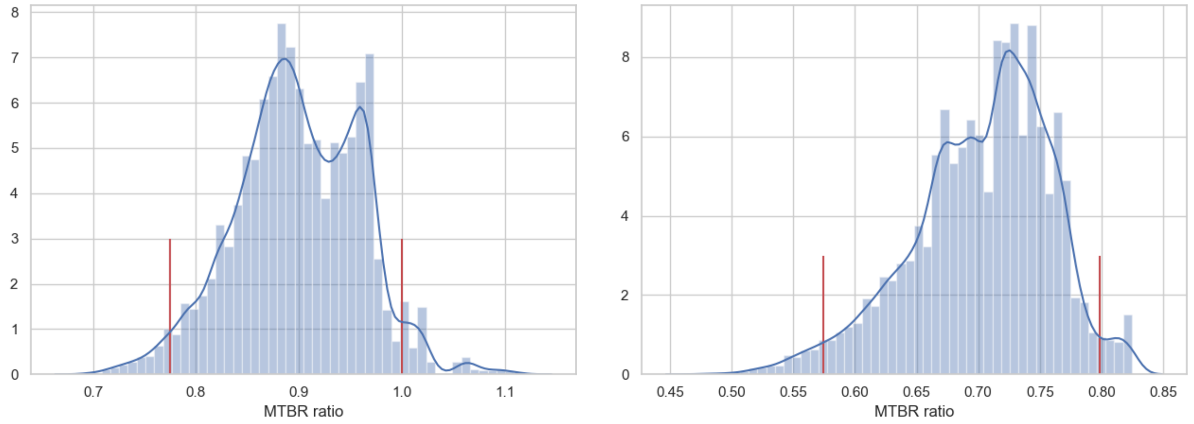


Figure F.6: Distributions of the RMST-ratio derived from Cox model fits on 1,000 bootstrap samples for covariate repaired (left) and desert (right). The red vertical lines indicate the 2.5th and 97.5th centiles, which approximate the true 95% confidence interval limits.

From the distributions of the RMST-ratio, estimates of the true 95% confidence interval limits can be computed using the 2.5th and 97.5th centiles. Concluding, the RMST-ratios and confidence intervals per covariate are given in Table F.4.

Table F.4: Restricted mean survival time ratios per covariate based on hypothetical population

Variable	*RMST-ratio*	*95% CI*
Age aircraft (FH)		
Increase from 0 to 3,000	0.83	(0.76, 0.87)
Increase from 15,000 to 18,000	0.95	(0.92, 0.96)
Repaired		
No (Ref)	1	-
Yes	0.86	(0.77, 1.00)
Natural climate, n (%)		
Temperate (Ref)	1	-
Humid	0.62	(0.50, 0.70)
Desert	0.73	(0.57, 0.80)

Bibliography

[1] A. A. Ghobbar and C. H. Friend, “Evaluation of forecasting methods for intermittent parts demand in the field of aviation: a predictive model,” *Computers & Operations Research*, vol. 30, no. 14, p. 2097–2114, 2003.

[2] F. Costantino, G. D. Gravio, R. Patriarca, and L. Petrella, “Spare parts management for irregular demand items,” *Omega*, vol. 81, p. 57–66, 2018.

[3] S. v. d. Auweraer, R. N. Boute, and A. Syntetos, “Forecasting spare part demand with installed base information: A review,” *SSRN Electronic Journal*, 2017.

[4] Q. Hu, J. E. Boylan, H. Chen, and A. Labib, “Or in maintenance spare parts management: A review,” *European Journal of Operational Research*, vol. 266, no. 2, p. 395–414, 2018.

[5] W. Stevenson, *Operations Management*. McGraw-Hill College, 2017.

[6] H. Qiwei, J. E. Boylan, C. Huijing, and A. Labib, “Or in spare parts management: A review,” *European Journal of Operational Research*, p. 395–414, 2018.

[7] C. O’Neil and R. Schutt, *Doing Data Science: Straight Talk from the Frontline*. O’Reilly, 2013.

[8] A. A. Syntetos, A. Boylan, and J. Croston, “On the categorisation of demand patterns,” *Journal of the Operational Research Society*, p. 495–503, 2005.

[9] J. E. Boylan, A. A. Syntetos, and G. C. Karakostas, “Classification for forecasting and stock control: a case study,” *Journal of the Operational Research Society*, vol. 59, no. 4, p. 473–481, 2008.

[10] E. V. Wingerden, R. Basten, R. Dekker, and W. Rustenburg, “More grip on inventory control through improved forecasting: A comparative study at three companies,” *International Journal of Production Economics*, vol. 157, p. 220–237, 2014.

[11] G. Heinecke, A. A. Syntetos, and W. Wang, “Forecasting-based sku classification,” *International Journal of Production Economics*, vol. 143, no. 2, p. 455–462, 2013.

[12] R. Teunter and L. Duncan, “Forecasting intermittent demand: a comparative study,” *Journal of the Operational Research Society*, vol. 60, p. 321–329, 2009.

[13] R. S. Gutierrez, A. O. Solis, and S. Mukhopadhyay, “Lumpy demand forecasting using neural networks,” *International Journal of Production Economics*, vol. 111, no. 2, p. 409–420, 2008.

[14] I. S. Markham and T. R. Rakes, "The effect of sample size and variability of data on the comparative performance of artificial neural networks and regression," *Computers & Operations Research*, vol. 25, no. 4, p. 251–263, 1998.

[15] A. A. Syntetos and J. E. Boylan, "On the bias of intermittent demand estimates," *International Journal of Production Economics*, vol. 71, no. 1-3, p. 457–466, 2001.

[16] ——, "The accuracy of intermittent demand estimates," *International Journal of Forecasting*, vol. 21, no. 2, p. 303–314, 2005.

[17] A. Syntetos, Z. Babai, J. Boylan, and S. Kolassa, "Supply chain forecasting: Theory, practice, their gap and the future," *European Journal of Operational Research*, no. 252, 2016.

[18] K. Nikolopoulos, A. A. Syntetos, J. E. Boylan, F. Petropoulos, and V. Assimakopoulos, "An aggregate–disaggregate intermittent demand approach (adida) to forecasting: an empirical proposition and analysis," *Journal of the Operational Research Society*, vol. 62, no. 3, p. 544–554, 2011.

[19] R. Dekker, C. Prince, R. Zuidwijk, and M. N. Jalil, "On the use of installed base information for spare parts logistics: A review of ideas and industry practice." *International Journal of Production Economics,*, vol. 143, no. 2, p. 536–545, 2013.

[20] B. Ghodrati and U. Kumar, "Operating environment-based spare parts forecasting and logistics: a case study," *International Journal of Logistics Research and Applications*, vol. 8, no. 2, p. 95–105, 2005.

[21] B. Ghodrati, A. Ahmadi, and D. Galar, "Spare parts estimation for machine availability improvement addressing its reliability and operating environment — case study," *International Journal of Reliability, Quality and Safety Engineering*, vol. 20, no. 03, p. 1340005, 2013.

[22] A. Barabadi, "Reliability and spare part provision considering operational environment. a case study," *International Journal of Performability Engineering*, vol. 8, no. 5, p. 497–506, 2012.

[23] A. H. C. Eaves and B. G. Kingsman, "Forecasting for the ordering and stock-holding of spare parts," *Journal of the Operational Research Society*, vol. 55, no. 4, p. 431–437, 2004.

[24] P. Wallström and A. Segerstedt, "Evaluation of forecasting error measurements and techniques for intermittent demand," *International Journal of Production Economics*, vol. 128, no. 2, p. 625–636, 2010.

[25] R. Hyndman, "Another look at forecast accuracy metrics for intermittent demand," *Foresight: the International Journal of Applied Forecasting*, vol. 4, no. 4, p. 43–46, 2006.

[26] D. Cielen, A. D. B. Meysman, and M. Ali, *Introducing data science big data, machine learning, and more, using Python tools*. Manning, 2016.

[27] W. McKinney, *Python for data analysis: data wrangling with Pandas, numPy, and IPython*. OReilly, 2013.

[28] C. A. Boone, C. W. Craighead, and J. B. Hanna, "Critical challenges of inventory management in service parts supply: A delphi study," *Operations Management Research*, vol. 1, no. 1, p. 31–39, 2008.

[29] C. H. Friend and A. A. Ghobbar, "Aircraft maintenance and inventory control: Using the material requirements planning system-can it reduce costs and increase efficiency?" *SAE Technical Paper Series*, Jan 1996.

[30] A. A. Ghobbar and C. H. Friend, "Aircraft maintenance and inventory control using the reorder point system," *International Journal of Production Research*, vol. 34, no. 10, p. 2863–2878, 1996.

[31] M. A. Moon, *Demand and supply integration the key to world-class demand forecasting*. Walter de Gruyter Inc., 2018.

[32] R. J. Hyndman and G. Athanasopoulos, *Forecasting: principles and practice*. OTexts, 2018.

[33] B. Malakooti, *Operations and production systems with multiple objectives*. John Wiley & Sons Inc., 2014.

[34] R. Bala and D. Kumar, "Classification using ann: A review," *International Journal of Computational Intelligence Research*, vol. 13, no. 7, p. 1811–1820, 2017.

[35] A. F. Atiya, "An unsupervised learning technique for artificial neural networks," *Neural Networks*, vol. 3, no. 6, p. 707–711, 1990.

[36] I. Guyin, A. Statnikov, and C. Aliferis, "Time series analysis with the causality workbench," *Challenges in Machine Learning*, vol. 5, 2011.

[37] R. Jiang and Z. Chen, "Modelling spare part demand data using transmuted or exponentiated poisson distribution," *Proceedings of the 10th IMA International Conference on Modelling in Industrial Maintenance and Reliability*, 2018.

[38] E. S. Gardner, "Forecasting for operations," *Keynote paper presented at the 31st international symposium on forecasting*, p. 27–29.

[39] H. Widiarta, S. Viswanathan, and R. Piplani, "Forecasting aggregate demand: An analytical evaluation of top-down versus bottom-up forecasting in a production planning framework," *International Journal of Production Economics*, vol. 118, no. 1, p. 87–94, 2009.

[40] T. P. Gordon, J. S. Morris, and B. J. Dangerfield, "Top-down or bottom-up: which is the best approach to forecasting?" *The Journal of Business Forecasting Methods & Systems,*, vol. 16, no. 3, p. 13–16, 1997.

[41] K. B. Kahn, "Revisiting top-down versus bottom-up forecasting," *The Journal of Business Forecasting Methods & Systems*, vol. 17, no. 2, p. 14–19, 1998.

[42] *Best Practices for Component Maintenance Cost Management*, 2nd ed. IATA, 2015.

[43] R. Adhikari and R. K. Agrawal, *An Introductory Study on Time series Modeling and Forecasting*. LAP Lambert Academic Publishing, 2013.

[44] G. E. P. Box, G. M. Jenkins, G. C. Reinsel, and G. M. Ljung, *Time series analysis: forecasting and control*. John Wiley & Sons, 2015.

[45] K. W. Hipel and A. I. MacLeod, *Time series modelling of water resources and environmental systems*. Elsevier, 1996.

[46] S. Makridakis and M. Hibon, "The m3-competition: results, conclusions and implications," *International Journal of Forecasting*, vol. 16, no. 4, p. 451–476, 2000.

[47] R. G. Brown, *Exponential smoothing for predicting demand*. Little, 1956.

[48] C. C. Holt, "Forecasting trends and seasonal by exponentially weighted averages," *Office of Naval Research Memorandum*, vol. 52, 1957.

[49] H. Pham, *Springer Handbook of Engineering Statistics*. Springer, 2006.

[50] J. Kamruzzaman, R. Begg, and R. A. Sarker, *Artificial Neural Networks in Finance and Manufacturing*. IGI Global, 2006.

[51] G. Zhang, B. E. Patuwo, and M. Y. Hu, “Forecasting with artificial neural networks: The state of the art,” *International Journal of Forecasting*, vol. 14, p. 35–62, 1998.

[52] G. P. Zhang, “Time series forecasting using a hybrid arima and neural network model,” *Neurocomputing*, vol. 50, p. 159–175, 2003.

[53] K. Hornik, M. Stinchcombe, and H. White, “Multilayer feedforward networks are universal approximators,” *Neural Networks*, vol. 2, no. 5, p. 359–366, 1989.

[54] J. D. Croston, “Forecasting and stock control for intermittent demand,” *Operations Research Q.*, vol. 23, p. 289–303, 1972.

[55] L. Shenstone and R. J. Hyndman, “Stochastic models underlying crostons method for intermittent demand forecasting,” *Journal of Forecasting*, vol. 24, no. 6, p. 389–402, 2005.

[56] F. Petropoulos, K. Nikolopoulos, G. P. Spithourakis, and V. Assimakopoulos, “Empirical heuristics for improving intermittent demand forecasting,” *Industrial Management & Data Systems*, vol. 113, no. 5, p. 683–696, 2013.

[57] P. A. Jacobs and P. A. W. Lewis, “Discrete time series generated by mixtures. i: Correlational and runs properties,” *Journal of the Royal Statistical Society. Series B (Methodological)*, vol. 40, no. 1, p. 94–105, 1978.

[58] E. Mckenzie, “Some simple models for discrete variate time series,” *Journal of the American Water Resources Association*, vol. 21, no. 4, p. 645–650, 1985.

[59] M. A. Al-Osh and A. A. Alzaid, “First-order integer-valued autoregressive (inar(1)) process,” *Journal of Time Series Analysis*, vol. 8, no. 3, p. 261–275, 1987.

[60] T. R. Willemain, C. N. Smart, and H. F. Schwarz, “A new approach to forecasting intermittent demand for service parts inventories,” *International Journal of Forecasting*, vol. 20, p. 375–387, 2004.

[61] A. A. Syntetos, M. Z. Babai, and E. S. Gardner, “Forecasting intermittent inventory demands: simple parametric methods vs. bootstrapping,” *Journal of Business Research*, vol. 68, no. 8, p. 1746–1752, 2015.

[62] J. Auramo and T. Ala-Risku, “Challenges for going downstream,” *International Journal of Logistics Research and Applications*, vol. 8, no. 4, p. 333–345, 2005.

[63] S. Minner, “Forecasting and inventory management for spare parts: An installed base approach,” *Service Parts Management*, p. 157–169, 2011.

[64] J. S. Hong, H. Koo, C. Lee, and J. Ahn, “Forecasting service parts demand for a discontinued product,” *IIE Transactions*, vol. 40, no. 7, p. 640–649, 2008.

[65] T. Y. Kim, R. Dekker, and C. Heij, “Spare part demand forecasting for consumer goods using installed base information,” *Computers & Industrial Engineering*, vol. 103, p. 201–215, 2017.

[66] M. Schraven, “Database driven forecasting of spare parts demand at the royal netherlands airforce,” *MSc. thesis*, 2015.

[67] A. A. Ghobbar and C. H. Friend, “Sources of intermittent demand for aircraft spare parts within airline operations,” *Journal of Air Transport Management*, vol. 8, no. 4, p. 221–231, 2002.

[68] D. Kumar, B. Klefsjö, and U. Kumar, “Reliability analysis of power transmission cables of electric mine loaders using the proportional hazards model,” *Reliability Engineering & System Safety*, vol. 37, no. 3, p. 217–222, 1992.

[69] W. R. Blischke and D. N. P. Murthy, *Reliability: Modeling, Prediction and Optimization.* Wiley, 2011.

[70] H. K. Han, H. S. Kim, and S. Y. Sohn, "Sequential association rules for forecasting failure patterns of aircrafts in korean airforce," *Expert Systems with Applications*, vol. 36, no. 2, p. 1129–1133, 2009.

[71] Y. Peng, M. Dong, and M. J. Zuo, "Current status of machine prognostics in condition-based maintenance: a review," *The International Journal of AdvancedManufacturing Technology*, vol. 50, no. 297, 2010.

[72] K. Tracht, G. Goch, P. Schuh, M. Sorg, and J. F. Westerkamp, "Failure probability prediction based on condition monitoring data of wind energy systems for spare parts supply," *CIRP Annals-Manufacturing Technology*, vol. 62, no. 1, p. 127–130, 2013.

[73] S. Letourneau, F. Famili, and S. Matwin, "Data mining to predict aircraft component replacement," *IEEE Intelligent Systems*, vol. 14, no. 6, p. 59–66, 1999.

[74] S. R. Hunt and I. G. Hebden, "Validation of the eurofighter typhoon structural health and usage monitoring system," *Smart Materials and Structures*, vol. 10, no. 3, p. 497–503, Jan 2001.

[75] S. Orhan, N. Aktürk, and V. Çelik, "Vibration monitoring for defect diagnosis of rolling element bearings as a predictive maintenance tool: Comprehensive case studies," *NDT & E International*, vol. 39, no. 4, p. 293–298, 2006.

[76] V. Deshpande, A. V. Iyer, and R. Cho, "Efficient supply chain management at the u.s. coast guard using part-age dependent supply replenishment policies," *Operations Research*, vol. 54, no. 6, p. 1028–1040, 2006.

[77] J. Gu, G. Zhang, and K. W. Li, "Efficient aircraft spare parts inventory management under demand uncertainty," *Journal of Air Transport Management*, vol. 42, p. 101–109, 2015.

[78] S. Zhu, W. v. Jaarsveld, and R. Dekker, "Using maintenance plan in spare part demand forecasting and inventory control." [Online]. Available: https://www.lancaster.ac.uk/media/lancaster-university/content-assets/documents/lums/forecasting/presentations/scfo/0203_Zhu.pdf

[79] Z. Hua, B. Zhang, J. Yang, and D. S. Tan, "A new approach of forecasting intermittent demand for spare parts inventories in the process industries," *The Journal of the Operational Research Society*, vol. 58, no. 1, p. 52–61, 2007.

[80] W. Romeijnders, R. Teunter, and W. v. Jaarsveld, "A two-step method for forecasting spare parts demand using information on component repairs," *European Journal of Operational Research*, vol. 220, no. 2, p. 386–393, 2012.

[81] C. Chatfield, "Calculating interval forecasts," *Journal of Business and Economic Statistics*, vol. 11, p. 121–144, 1992.

[82] R. D. Snyder, J. K. Ord, and A. Beaumont, "Forecasting the intermittent demand for slow-moving inventories: A modelling approach," *International Journal of Forecasting*, vol. 28, no. 2, p. 485–496, 2012.

[83] Z. Zhang, "Residuals and regression diagnostics: focusing on logistic regression," *Annals of Translational Medicine*, vol. 4, no. 10, p. 195–198, 2016.

[84] M. H. Kutner, C. Nachtsheim, J. Neter, and W. Li, *Applied linear statistical models.* McGraw-Hill, 1996.

[85] G. E. P. Box and D. A. Pierce, “Distribution of residual autocorrelations in autoregressive-integrated moving average time series models,” *Journal of the American Statistical Association*, vol. 65, no. 332, p. 1509–1526, 1970.

[86] C. M. Jarque and A. K. Bera, “Efficient tests for normality, homoscedasticity and serial independence of regression residuals,” *Economics Letters*, vol. 6, no. 3, p. 255–259, 1980.

[87] N. Kourentzes, “On intermittent demand model optimisation and selection,” *International Journal of Production Economics*, vol. 156, p. 180–190, 2014.

[88] Y. Hong, J. Zhou, and M. A. Lanham, “Forecasting intermittent demand patterns with time series and machine learning methodologies,” *Purdue University, Department of Management*.

[89] D. Chen and H. W. Chen, “Using the köppen classification to quantify climate variation and change: An example for 1901–2010,” *Environmental Development*, vol. 6, p. 69–79, 2013.

[90] “Airport, airline and route data.” [Online]. Available: https://openflights.org/data.html

[91] “List of hub airports,” Mar 2019. [Online]. Available: https://en.wikipedia.org/wiki/List_of_hub_airports

[92] H. E. Beck, N. E. Zimmermann, T. R. Mcvicar, N. Vergopolan, A. Berg, and E. F. Wood, “Present and future köppen-geiger climate classification maps at 1-km resolution,” *Scientific Data*, vol. 5, p. 180214, 2018.

[93] G. van Brummelen, *Heavenly mathematics: the forgotten art of spherical trigonometry*. Princeton University Press, 2013.

[94] T. Tinga, *Principles of loads and failure mechanisms: applications in maintenance, reliability and design*. Springer, 2013.

[95] T. Jin, *Reliability engineering and services*. Wiley Blackwell, 2019.

[96] J. Knezevic, *Reliability, maintainability and supportability: a probabilistic approach*. McGraw-Hill, 1993.

[97] A. Szczepankowski, J. Szymczak, and R. Przysowa, “The effect of a dusty environment upon performance and operating parameters of aircraft gas turbine engines,” 05 2017.

[98] C. Chang, “Study on the correlation between humidity and material strains in separable micro humidity sensor design,” *Sensors*, vol. 17, no. 5, p. 1066, 2017.

[99] S. Oesch and M. Faller, “Environmental effects on materials: The effect of the air pollutants so2, no2, no and o3 on the corrosion of copper, zinc and aluminium. a short literature survey and results of laboratory exposures,” *Corrosion Science*, vol. 39, no. 9, p. 1505–1530, 1997.

[100] B. Efron and R. Tibshirani, “Bootstrap methods for standard errors, confidence intervals, and other measures of statistical accuracy,” *Statistical Science*, vol. 1, no. 1, p. 54–75, 1986.