FAKULTÄT
FÜR INFORMATIK

Faculty of Informatics

Perfomance of Suvival Models on Predictive Maintenance of Construction Machines

Semiparametric evaluation of Predictive Models

MASTERARBEIT

zur Erlangung des akademischen Grades

Master of Science

im Rahmen des Studiums

MSc Logic and Computation

eingereicht von

Bakk.-techn. Daniel Guimarães, BEng.
Matrikelnummer 11831513

an der Fakultät für Informatik

der Technischen Universität Wien

Betreuung: Univ.-Prof., Dipl.-Ing., Dr.techn. Peter Filzmoser
Mitwirkung: Univ.Ass., Dipl.-Ing. Marco Huymajer, Bsc.

Wien, 15. November 2019

Daniel Guimarães

Peter Filzmoser

FAKULTÄT
FÜR INFORMATIK

Faculty of Informatics

Performance of Survival Models on Predictive Maintenance of Construction Machines

Semiparametric evaluation of Predictive Models

MASTER'S THESIS

submitted in partial fulfillment of the requirements for the degree of

Master of Science

in

Computational Logic

by

Bakk.-techn. Daniel Guimarães, BEng.
Registration Number 11831513

to the Faculty of Informatics

at the TU Wien

Advisor: Univ.-Prof., Dipl.-Ing., Dr.techn. Peter Filzmoser
Assistance: Univ.Ass., Dipl.-Ing. Marco Huymajer, Bsc.

Vienna, 15th November, 2019

Daniel Guimarães

Peter Filzmoser

Erklärung zur Verfassung der Arbeit

Bakk.-techn. Daniel Guimarães, BEng.

Hiermit erkläre ich, dass ich diese Arbeit selbständig verfasst habe, dass ich die verwendeten Quellen und Hilfsmittel vollständig angegeben habe und dass ich die Stellen der Arbeit – einschließlich Tabellen, Karten und Abbildungen –, die anderen Werken oder dem Internet im Wortlaut oder dem Sinn nach entnommen sind, auf jeden Fall unter Angabe der Quelle als Entlehnung kenntlich gemacht habe.

Wien, 15. November 2019

Daniel Guimarães

Abstract

Manufacturers of construction machines nowadays often provide tools to contractors to monitor through a telecommunication device the state of machines. Besides telematics data, repair information are sometimes collected by mechanics with the intention of improving the mainenance strategy. These two sources of data allow the construction of statistical models capable of condition-based maintenance. Nevertheless, either human effort need to be systematized or a cross-disciplinary effort must emerge to train statistical methods for use in production. This study demonstrates the process of building, testing and validating the Cox proportional hazards predictive model in the context of construction machines. The final model still provide poor discrimination and accuracy to be put in production, however, it presents the basic framework and validation process which potential production models must abide. Moreover, it provides insights on what can be done to improve data extraction.

Contents

CHAPTER 1

Introduction

Access to digital technology has shown a noteworthy accession over the last years with lower computing, storage, bandwidth, and sensor costs [HMRM]. Technological progress has at the same time increased efficiency and popularized a particular maintenance strategy once only accessible to industry leaders. This technology dissemination has made viable a powerful maintenance program that put organizations under constant pressure to remain competitive. Today, condition-based maintenance is possible due to smart, connected technologies that unite digital and physical assets, as explained in a Delloite study [CDCD].

Condition-based maintenance, commonly known as predictive maintenance is defined by ISO 13372:2012(en) as "maintenance performed as governed by condition monitoring programs". Moreover, condition monitoring (CM) is the "acquisition and processing of information and data that indicate the state of a machine over time". Although not a new concept, it still is the most promising category of maintenance to alleviate the impact of downtime (the period during which a machine is not functional). In an overview [CDCD], other existing categories of maintenance, namely, reactive, planned and proactive [CDCD] are compared to condition-based monitoring. Estimating the length from the current time to the end of the useful life of an asset, the residual useful life (RUL), is of major importance to schedule maintenance activities. The proposed category of maintenance aims at maximizing the RUL of machines avoiding unplanned downtime thereby saving costs.

Analysts have found that 5 to 20% of productive capacity can be affected by poor maintenance strategies, three-fourths of facilities are not even able to accurately estimate their total downtime cost [Wol]. In automotive manufacturing, downtime can cost 1.3 million dollars per hour [Wol]. It has been estimated [top] unplanned downtime costs industrial manufacturers an estimated 50 billion dollars annually, where equipment failure represents 42% of the total unplanned downtime. Strategies such as waiting for a component to fail or replacing a perfectly good component results not only in excessive

maintenance, repair, and equipment replacement but also an increase of inventory and delays.

In a comprehensive study on RUL estimation [SWHZ], two categories of observed condition monitoring are proposed, event data and CM data. Event data means simply recorded failure data from targeted assets, where failure can happen either to the whole machine or to a machine's part. In real scenarios, however, it may be burdensome and costly to run machines to failure for the sake of collecting and storing event data. CM data provides important information that may have a connection with the estimation of the RUL of machines, such as monitored CM information, operational, performance, environmental information, and degradation signals [SWHZ]. The study distinguishes, furthermore, CM data between direct and indirect CM. In 2011, Si, Wang, Hu, and Zhou defined as direct CM data, "the data which can describe the underlying state of the system directly so that the prediction of the RUL is the prediction of the CM data to reach a predefined threshold level". On the other hand, indirect CM data "can only indirectly or partially indicate the underlying state of the system so failure event data may be needed in addition to CM data for an RUL estimation purpose". After these definitions, it is possible to distinguish models based on directly observed state processes and based on indirectly observed state processes.

In this study, a model based on data describing indirectly observed state processes will be presented, most specifically, the Cox proportional hazards model. The Cox model was firstly introduced by Sir David Roxbee Cox in 1972 [Cox72], and is still nowadays a popular covariate-based model for lifetime analysis. Supplied with event data, this covariate-based model aims at predicting when a particular failure event is likely to occur. This analysis can also be applied to the domain of construction machines, nevertheless not the whole of the machine underwent modeling but only a very specific part within. The goal of this work is not to provide predictive maintenance, or RUL, given the model and manufacturer of a machine used in construction, nor to provide precise time points informing when the machine will fail. Instead, the focus lays on the process of building, testing and validating predictive proportional hazards models. Such predictive models, after verified, could then assist the decision of when to perform maintenance.

The thesis is structured in the following form, chapter 2 presents the methodology with theoretical detail; in chapter 3 the overall process will be demonstrated by a use case targetting a single minor event; lastly, the conclusions from the case study are presented along with further study in 4. Nevertheless, firstly a literature review elicits problems faced nowadays when using Cox proportional hazards for prediction and studies with similar data and domain.

1.1 Literature Review

As mentioned before, neither Cox proportional hazards nor the field of condition-based maintenance are new topics. Already in the 90s there were already researchers using Cox proportional hazards to define maintenance scheduling of replacement of pressure gauges [KW97], modelling plant maintenance [MJ92], and replacement policies

for system subject to stochastic deterioration [LG91]. Nonetheless, in the last decade, Cox proportional hazards is still regarded as a relevant method for condition-based maintenance. For instance, in 2016 it was used to estimate cutting tools lifetime, with cutting speed as the explanatory variable [ELSD16]; in 2015 time-varying hazard functions were used to analyse successive life intervals of construction heavy equipment with use of telematics data [SN15] and this year a hybrid between neural networks and Cox Regression has been proposed [KBS19].

Despite being used in industrial contexts, the bulk of proportional hazards uses come from the medical field. After a famous commentary from C. M. Parkes on the paper [CL00], published in 2000, where doctor's expertise were used to make point prediction estimates of patient's survival: "Prognoses should be based on proved indices not intuition", which motivated researchers to use statistical models to estimate patient's survival time. In 2001, Henderson, Jones and Stare [HJS01] using Weibull and Cox proportional hazards models for survival time prediction of patients diagnosed with non-small cell lung cancer concluded: "It seems that Parkes suggestion that clinicians should stop guessing and make more use of statistical models is unlikely to lead to much improvement in accuracy in point predictions", motivating the use of some reliability measure, such as prediction intervals or the associated probability of error. Later in 2005, Henderson and Keiding on the same data stated: "statistical indices provide poor discriminatory power at the individual level" [HK05] motivating the use of such models at the group or population level. Despite the poor results estimating when exactly patients would decease, Cox regression models are still used in medical domains. In 1991, it was used to measure the effect of repeated transcatheter arterial embolization on the survival time patients with hepatocellular carcinoma [IKS+91]; in 1999, to measure the association between cognitive function at ages 8, 11, 15, 26, and 43 years and menopause timing [MKRM99]; in 2010, to identify predictors of heart failure mortality on patients with pulmonary arterial hypertension [TTT+10]; in 2011, to evaluate the influence of various clinicopathological and biochemical factors on the survival of patients with epithelial ovarian cancer (EOC) after radical resection [WHSS11] and in 2014, it was used on two high-dimensional, massive sample-size data sets, concearning pediatric trauma and breast cancer gene expression [MMBS14].

Concerning construction machines, few studies were found on a similar setting, namely, using telematics data, however, with very small sample size. The first condition is met by two studies done by Said and Nicoletti, [SN15] and another including Perez-Hernandez [SNPH15]. In both works, time-varying periods of the survival function are defined to represent the dynamically changing equipment's failure hazard over time, whereas in this study a single model is built to represent the machine's life until the first failure event; only the baseline hazard reflects time-dependency. Regarding scarce data, [WZ04] performs oil based monitoring on a set of censored life data of 30 aircraft engines. This last paper has contributed with insights such as the use of principal component analysis to mitigate the estimation of multiple coefficients.

CHAPTER 2

Methodology

This chapter consists of the theoretical framework used throughout this study. The scope of the framework was defined by the topic's purpose in the implementation, presented in the upcoming Case Study chapter 3. Thus, all investigated topics should be further considered in the Case Study, unless, it is explicitly mentioned that it will not be taken into consideration.

2.1 Introduction to Survival Analysis

Generally, survival analysis is a collection of statistical procedures for data analysis for which the outcome variable of interest is time until an event occurs [KK12].

In survival analysis, interest centers on a group or groups of individuals for whom there is a defined event, occurring after a length of time often referred to as the failure time. The definition of a time point of failure was specified by Cox and Oakes [CO84]. According to them in order to determine failure time precisely, there are three requirements: a time origin must be unambiguously defined, a scale for measuring the passage of time must be agreed and finally, the meaning of failure must be entirely clear. In some applications, there is little or no arbitrariness in the definition of failure, for example in some industrial contexts when a tractor breakdown and it is prevented from locomotion due to a general or specific cause. In others, failure is defined as the first instance at which performance falls below an acceptable level. Then there will be some arbitrariness in the definition of failure and it will be for consideration whether to concentrate on modelling the time to failure event or whether to analyze the performance measure as a function of time.

Some statistical methods such as linear regression, are not adequate to perform survival analysis. Linear regression could yield negative survival times, and most importantly it does not offer support for censoring [Des].

2.1.1 Censoring of Observations

Survival analysis uses both censored and uncensored information to estimate important model parameters. The outcome in survival analysis is twofold, one quantitative and another qualitative, the former is simply the time to an event; the latter is the status of the observation, which records if the event has occurred or not. In censored observations, the quantitative part do not define a failure time. For instance, in a medical study observing heart attacks, if one died from another cause; or simply did not suffer a heart attack during the established duration of the study, this observation is said to be censored. An observation is called right-censored if it did not suffer a heart attack during the duration of the study or decided to quit the experiment. In general, an observation is said to be right censored if it was alive at study termination or was lost to follow-up at any time during the study . By right censoring, it is meant that the survival time is only known to exceed a certain value [LEA97]. Differently, left censoring happens when the event occurred before the start of the experiment. Throughout this work, the assumption of non-informative censoring holds, meaning, subjects exit due to reasons unrelated to the study. Oppositely, in informative censoring subject leave due to reasons related to the study.

2.1.2 The Survival and Hazard Functions

In the context of survival analysis, two time-dependent functions are important when describing the time for an event. Let T be a nonnegative continuous random variable representing failure time. The survival function is the probability of an event happening after given time $t > 0$.

$$S(t) = P(T > t) \tag{2.1}$$

Assuming that T has a probability density function $f(t)$ and cumulative distribution function $F(t)$.

$$S(t) = 1 - F(t) = \int_t^\infty f(x)dx \tag{2.2}$$

The survival times $t_1, t_2, ..., t_n$ arise in an independent and identically distributed fashion from density and survival functions $f(t)$ and $S(t)$ [MMBS14].

The hazard function, gives the instantaneous risk rate per unit of time. Let Δt be a short interval of time, then the hazard can be defined as the conditional probability of the event happening between the interval from t to $t + \Delta t$, assuming the observation has survived up to time t. The hazard is the instantaneous potential of an individual failing as Δt tends to zero.

$$h(t) = \lim_{\Delta t \to 0^+} \frac{P(t \leq T < t + \Delta t | T \geq t)}{\Delta t} \tag{2.3}$$

Although the hazard's numerator is a probability function, in the denominator it has a value which tends to zero, so it varies from 0 to infinity. By applying the definition of conditional probability to the numerator of equation 2.3, an interesting relationship can be found.

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

$$= \lim_{\Delta t \to 0^+} \frac{P(t \leq T < t + \Delta t)}{\Delta t S(t)}$$

$$= \lim_{\Delta t \to 0^+} \frac{f(t)\Delta t}{\Delta t S(t)} = \frac{f(t)}{S(t)} \tag{2.4}$$

The instantaneous risk rate of an event at t is the density at t divided by the probability of survival exceeding t. Despite the interesting result of 2.4, an assumption on the distribution of time to event T must be made to assess the density function $f(t)$, motivation to find a relationship solely between the hazard and the survival function. Deriving equation 2.2.

$$\frac{\delta S(t)}{\delta t} = 0 - f(t) = -f(t) \tag{2.5}$$

Substituting the density function in equation 2.4, by the density from equation 2.5,

$$h(t) = -\left(\frac{\delta S(t)}{\delta t}\right)\frac{1}{S(t)}$$

and using the following chain rule,

$$\frac{\delta ln(g(y))}{\delta y} = \frac{1}{g(y)}\frac{\delta g(y)}{\delta y}$$

the hazard can be written as,

$$h(t) = -\frac{\delta ln S(t)}{\delta t} \tag{2.6}$$

The result above could for instance be found on Rodríguez lecture notes [Rod07]. Inversely, one could also obtain the survival function as a function of the hazard.

$$S(t) = e^{-\int_0^t h(s)ds} \tag{2.7}$$

Equations 2.6 and 2.7 enables to convert between the hazard and survival functions without making any assumptions about the cumulative distribution function. The last relationship worth being aware of is between the cumulative hazard function and the survival function. The cumulative hazard function is the aggregated hazards from beginning to a given time t.

$$H(t) = \int_0^t h(s)ds \tag{2.8}$$

From equation 2.7 we can obtain the survival function as a function of the cumulative hazard.

$$S(t) = e^{-H(t)} \tag{2.9}$$

In this work, we have used the random variable approach to specify the framework of survival analysis. Alternatively, one might be interested in formulating the defined concepts using the counting process approach. The benefit of using counting processes is the connection with martingale residual theory, which may be useful to prove the distributional characteristics of survival analysis. For the interested reader, further details can be found in the book by Therneau and Grambsch (2000) [TG00] or Hosmer, Lemeshow and May "An Introduction to the Counting Process Approach to Survival Analysis. Appendix 2" on [HLS08] for a brief introduction of counting processes. However, these theoretical results are beyond the scope of this work.

For the sake of building intuition it is useful to look at a simpler concept of survival, that does not account for covariates for the computation of survival curves.

2.1.3 Kaplan-Meier Survival Estimate

The Kaplan-Meier estimator generates the survival probability from observed survival data. The Kaplan-Meier is a nonparametric estimate of the survival function from observed and possibly right-censored survival data. Assuming that n independent and possibly right-censored events happened at discrete time points $t_1, t_2, ..t_n$. Let e_i be the number of events at t_i; t_1 and t_n be the first and last observed event, respectively; n_i be the number of patients at risk (i.e. either alive or uncensored) up until time t_{i-}, moment just before time t_i. The Product-Limit estimator or the estimated probability of not experiencing an event up until (and including) time t_i [HLS08].

$$\hat{S}(t_i) = \hat{S}(t_{i-1})\left(1 - \frac{e_i}{n_i}\right) \tag{2.10}$$

Note e_i/n_i represents the rate of patients affected by an event at t_i, if no event happened e_i will be 0. Therefore this estimator gives a step function that is modified

at every unique t_i. The Product-Limit estimator is only well-defined until the last observation time t_n [KM03]. At the other boundary, $\hat{S}(t)$ is 1 for any time before the first event, $t < t_1$. Equation 2.10 can also be written in a nonrecurrent format.

$$\hat{S}(t) = \prod_{i:t_i<t} \left(1 - \frac{e_i}{n_i}\right) \tag{2.11}$$

The Product-Limit estimator can also be used to estimate the cumulative hazard function, using the inverse of equation 2.9. The Nelson-Aalen estimator defined up until the last observation time t_n.

$$\hat{H}(t) = \sum_{i:t_i<t} \left(\frac{e_i}{n_i}\right) \tag{2.12}$$

The construction of the survival and hazard functions with this estimator requires only censoring and event information. Because other context information is ignored, this method is said to provide a nonparametric statistic or a univariate analysis. To account for covariates that may influence the time to an event, other methods will be investigated.

2.1.4 Kaplan-Meier example

Going back to the heart attack example, imagine 12 patients were observed during a study period of 15 months, any incidence of a heart attack was reported on the annotation represented by the table bellow.

t_i	n_i	e_i	c_i	$1 - e_i/n_i$	$\hat{S}(t)$
0	12	0	0	1	1
2	12	2	0	10/12	0.833
4	10	1	1	9/10	0.75
5	8	1	0	7/8	0.656
8	7	1	2	6/7	0.562
12	4	1	0	3/4	0.421
13	3	1	0	2/3	0.28
15	2	1	1	1/2	0.14

Table 2.1: Example of Kaplan Meier survival function computation. Table of incidence of heart attack on 12 patients.

The column t_i represents the month the event occurred; n_i the number of subjects at risk at just before t_i; e_i the number of failure events at t_i and c_i the respective number

of censored events. Note that between months 4 and 5 one individual was censored, but there is no known time of event for this individual since only entries of actual failure times are registered. Moreover on month 8, three subjects are registered 1 failed and 2 were censored. At month 15 only a single subject neither failed nor was censored during the study, hence the survival in the figure below remains 0.14 for $t > 15$. The survival at respective failure times can be computed using equation 2.11.

The survival step function drops at every failure time. The height of the drop depends on the ratio between the amount of individuals that suffered the event at time t and the individuals at risk shortly prior to t. More details on Kaplan-Meier is provided by Lee and Go [sur97].

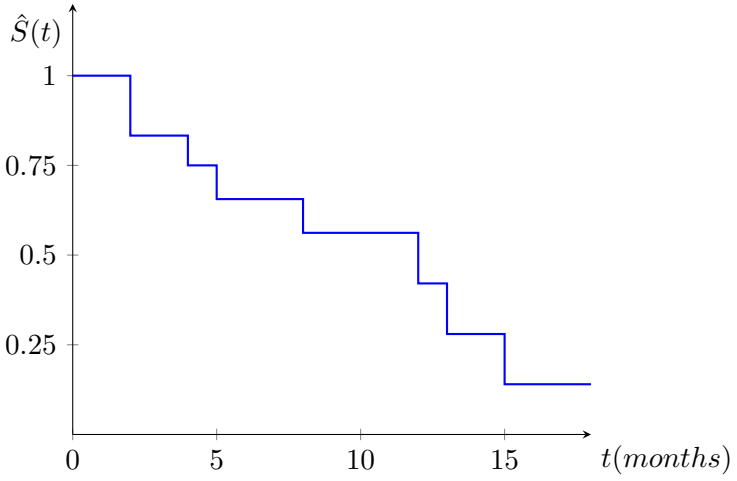


Figure 2.1: Kaplan-Meier survival probability estimates on 12 patients example.

2.2 The Cox Proportional Hazards model

The Cox proportional hazards (PH) models is a widely known and used procedure that makes use of covariates to model the time of an event. It was the main type of model in this study, and provides multiple extensions to model problems of various natures.

Let X_{ij} be a covariate value where i indicates a subject and j the selected covariate, notation which is only adequate if the covariates are fixed over time. Regardless if the observed time t_i was censored or not, it is conditionally independent given X_i and, as mentioned in section 2.1.1, the censoring mechanism is non-informative. The hazard function for subject i is given by the following expression.

$$h_i(t) = h_0(t)e^{X_i^T \beta} \tag{2.13}$$

This hazard function is often seen as a combination of a parametric piece, $e^{X_i^T \beta}$ (a function of the coefficients β); and a nonparametric piece $h_0(t)$, for this reason, the model

is referred as semi-parametric. The baseline hazard $h_0(t)$ is an unspecified nonnegative function when the covariate vector is equal zero, i.e.

$$X_j = 0 \quad \forall j \in \{1, ...p\} \tag{2.14}$$

The exponential function in the parametric piece and the nonnegative characteristic of the baseline guarantees that the hazard will never be negative. Additionally, when the hazard is multiplied by an interval of time Δt, it should represent the probability of an event over the interval, which cannot be negative.

The semi-parametric Cox model should closely approximate the underlying parametric model, this is because we make no assumptions on the baseline or distribution, hence the results obtained through Cox should be comparable to results obtained with the correct parametric model. Therefore, Cox was a "safe" choice [KK12] because although the distribution is not known a priori if we would have chosen a parametric algorithm, an assumption on the underlying distribution would need to be made. In the time-independent approach, each subject is assigned a single vector corresponding to its covariates X_i. For example, estimated time for an event could be attempted using the information of the age of a patient upon entry of study, and its gender.

Regardless of the selected covariates, only the first failure event is of interest, therefore on the case of multiple events, often times all the information after the first will not be used. The defining characteristic of the PH model is the PH assumption, which states the following [KK12].

> *The PH assumption requires that the Hazard Ratio is constant over time, or equivalently, that the hazard for one individual is proportional to the hazard for any other individual; where the proportionality constant is independent of time.*

The hazard ratio mentioned above, is the ratio of the hazards of two individuals, let X_1 and X_2 be the covariate vector of two subjects respectively.

$$HR = \frac{h_1(t)}{h_2(t)} = \frac{h_0(t)e^{X_1^T\beta}}{h_0(t)e^{X_2^T\beta}}$$

The baseline function is common to all subjects, it can be thought as the hazard for an individual with all covariates equal zero.

$$HR = \frac{e^{X_1^T\beta}}{e^{X_2^T\beta}} = exp((X_1 - X_2)^T\beta) \tag{2.15}$$

Observe that in this equation there is no term dependent on time, because of the canceled baseline and the covariates X do not change over time. Once the β vector of

coefficients is known, the hazard ratio is constant κ. From the quotation above, the careful reader will also note that the hazards for the two subjects have to be proportional.

$$\kappa = h_1(t)/h_2(t)$$
$$h_1(t) = \kappa h_2(t)$$

As a thought experiment, let's imagine that covariates can change over time and that there are only two patients p_1 and p_2 in a heart attack study. Consider that in the first moment of the study p_2 was healthier and less prone to suffer a heart attack than p_1. However, after a few weeks p_2 through bad habits adequately quantified in the covariate values $X_2(t)$ of p_2, becomes more likely to suffer a heart attack than p_1.

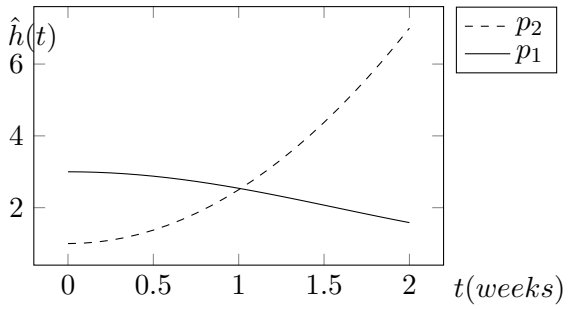


Figure 2.2: Crossing of the hazards of two individuals over time. [KK12]

Note in figure 2.2 that at week one, both patients have the same hazards value, however:

$$h_1(0.5)/h_2(0.5) > 1$$
$$h_1(1.5)/h_2(1.5) < 1$$

Clearly, in this case, the hazard ratio is not constant over time for this pair of individuals, therefore the PH assumption is not met [KK12]. To disqualify the PH assumption, as a rule of thumb one can check if the hazards cross for a pair of individuals. Note, nevertheless, that it is not a sufficient condition to disqualify the PH assumption, meaning that nonproportionality can be found even when the hazards does not cross. One possible workaround, when such crossing is found, is to fit one model for $t < t_c$ and another for $t > t_c$ where t_c is the time of crossing.

2.2.1 Proportional Hazard Likelihood

The β coefficients present in equation 2.13 are produced by maximizing a likelihood function $L(\beta)$. This likelihood is the joint probability of target time-to-event random variable, therefore such formulation often requires the knowledge about the distribution

of the outcome. The likelihood construction for the Cox model is based on the order of events rather than on the joint distribution [KK12].

$$L(\beta) = L_1 \times L_2 ... L_N = \prod_{i=1}^{N} L_i \tag{2.16}$$

N is the number of distinct times which at least one subject failed, for each of these time points a likelihood term is calculated, thus we can assign for each term in $\{L_1, L_2, ...L_n\}$ a corresponding time $\{t_1, t_2, ...t_N\}$. The adjective "partial" is used to denote that only the subjects who failed are assigned respective likelihood terms, i.e. if there was a time t_j when an individual was right-censored but no other subjects failed, we would not have at this time an assigned likelihood term L_j.

The risk set $R(t)$ contains the subjects which have not yet failed up to a given time t, or failed exactly at t. Suppose there are four subjects s_1,s_2, s_3 and s_4 that fail in the times shown figure 2.3.

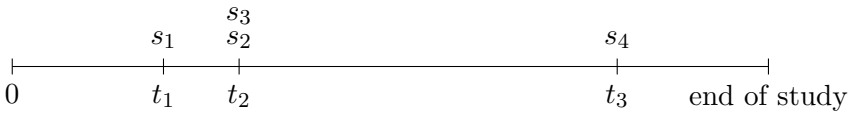


Figure 2.3: Four events example

Following the timeline of figure 2.3, up until t_1 (and including at t_1) all subjects are on the risk set, for $t \leq t_2$ only s_2, s_3 and s_4; and for $t \leq t_3$ only s_4 is on the risk set. Oppositely, the notion of a fail set $F(t)$ indicates the subject which failed at t.

$$F(t) = \begin{cases} \{s_1\}, & t = t_1 \\ \{s_2, s_3\}, & t = t_2 \\ \{s_4\}, & t = t_3 \\ \emptyset, & \text{otherwise} \end{cases}$$

With use of these two notions, each term in equation 2.16 can be expressed. The portion of the likelihood on the ith failure time t_i given the risk and fail set $R(t_i), F(t_i)$ can be defined as, vide [sur97],

$$L_i(\beta) = \frac{h_{m_i}(t)}{\sum\limits_{r \in R(t_i)} h_r(t)} \tag{2.17}$$

where it holds that,

$$F(t_i) = \{m_i\}$$

Despite figure 2.3 two subjects fails at the same time, in the likelihood construction above 2.17 assumes only one subject fails at each time. Theoretically assuming continuous time, it is indeed very unlikely that two subjects fail exactly at the same time. Under the PH assumption, the nonparametric part of the hazard will cancel each other, leaving only the risk score of the ith subject $exp(X_i^T\beta)$, hence the baseline hazard function plays no role in the estimation of the coefficients.

$$L_i(\beta) = \frac{exp(X_i^T\beta)}{\sum\limits_{r\in R(t_i)} exp(X_r^T\beta)} \tag{2.18}$$

Even if errors are made at the time which observations failed, but the order and the above restriction on $F(t)$ were preserved, the likelihood would still be correctly calculated and would give the right coefficients precisely because the partial likelihood construction estimates the coefficients based on the order of events rather than by the distribution of time random variable.

As already mentioned, the suitable Cox coefficients β can be estimated by maximizing likelihood function, or equivalently the coefficients β can be estimated by the log partial-likelihood function.

$$lnL(\beta) = ln\prod_{i=0}^{N} \frac{exp(X_i^T\beta)}{\sum\limits_{r\in R(t_i)} exp(X_r^T\beta)}$$

$$= \sum_{i=0}^{N} X_i^T\beta - ln \sum_{r\in R(t_i)} exp(X_r^T\beta)$$

One way to accomplish such a task is to force the first derivative of the log-likelihood (also known as (Fisher's) score, not to be confused with the risk score) to be equal to zero. There will be p equations, one for each covariate, the right side of the equation above will be derived with respect to the coefficient β. Depending on the specific coefficient β_a, the derivation of the risk score $exp(X_j^T\beta)$ will be multiplied by a different X_{ja} which means simply the ath covariate for the jth subject.

$$\frac{\delta lnL(\beta)}{\delta\beta_a} = \sum_{i=0}^{N} X_{ia} - \frac{\delta \ln \sum_{r\in R(t_i)} exp(X_r^T\beta)}{\delta\beta} = 0$$

$$= \sum_{i=1}^{N} X_{ia} - \sum_{r\in R(t_i)} \left(\frac{X_{ra} exp(X_r^T\beta)}{\sum_{q\in R(t_i)} exp(X_q^T\beta)} \right) = 0 \tag{2.19}$$

The indexes i from 1 to N in equation 2.19 are only for the individual which experienced the event of interest with covariates X_i. The estimator of the variance of β_a

is obtained by computing the information matrix, the negative of the second derivative of the log-likelihood. Hence, by computing the second derivative of equation 2.19 by applying the quotient rule.

$$-\sum_{i=1}^{N} \frac{\left(\sum_{r\in R(t_i)}(X_{ra})^2 exp(X_r^T\beta)\right)\left(\sum_{r\in R(t_i)} exp(X_r^T\beta)\right) - \left(\sum_{r\in R(t_i)} X_{ra} exp(X_r^T\beta)\right)^2}{\left(\sum_{r\in R(t_i)} exp(X_r^T\beta)\right)^2} \tag{2.20}$$

or simply

$$\frac{\delta^2 L(\beta)}{\delta\beta_a^2} = -\left(\sum_{i=1}^{N} \frac{\sum_{r\in R(t_i)} X_{ra}^2 exp(X_r^T\beta)}{\sum_{r\in R(t_i)} exp(X_r^T\beta)} - \left(\sum_{r\in R(t_i)} \frac{X_{ra} exp(X_r^T\beta)}{\sum_{r\in R(t_i)} exp(X_r^T\beta)}\right)^2\right) \tag{2.21}$$

Note that the squared term on the right side of equation 2.21 is the weighted average of the covariate values for subjects in the risk set. Once the second derivative matrix of the estimated coefficients obtained by equation 2.19 are known, the variance of $\hat{\beta}$ is simply the inverse of the information matrix [HLS08].

$$I(\beta) = -\frac{\delta^2 L(\beta)}{\delta\beta^2} \tag{2.22}$$

$$\mathcal{V}(\hat{\beta}) = I(\hat{\beta})^{-1} \tag{2.23}$$

Since only the variance is of interest, only the elements in the diagonal of the information matrix will be needed, this is why equation 2.21 uses the same subscript a. This equation can be further generalized to compute the covariance as seen in [MN04] or [HLS08].

2.2.2 Handling Ties

Because on real-world scenarios, the event time scale is discrete and because continuous times are grouped into intervals [TG00], often the likelihood calculation have to account for tied event times. We already have an example given by figure 2.3 to illustrate this problem, assuming only four subjects in the study, how to account for the contributions of s_2 and s_3? Allow the notation of the risk score contribution to the likelihood function of the ith subject was to be r_i, then there are two ways of computing

the two likelihood contributions:

$$\left(\frac{r_2}{r_2+r_3+r_4+r_5}\right)\left(\frac{r_3}{r_3+r_4+r_5}\right) \tag{2.24}$$

$$\left(\frac{r_3}{r_2+r_3+r_4+r_5}\right)\left(\frac{r_2}{r_2+r_4+r_5}\right) \tag{2.25}$$

To answer this question, different approximations of the partial likelihood function were proposed.

Breslow approximation

The simplest, and perhaps most intuitive case would be to assume that the risk set for each contribution is the same, namely, all subjects at risk including the tied subjects, so $r_2 + r_3 + r_4 + r_5$. The problem with this approach is that we keep both subjects in the risk set of the second contribution, despite knowing that theoretically, it is very unlikely that both of them failed exactly at the same time. Therefore, we are increasing the overall denominator of the likelihood calculation, making it easier to maximize, thus biasing the coefficients towards zero [TG00].

Efron approximation

This most accurate approximation assigns a probability to the likelihood terms in the risk set that has tied events. In our case, it would simply multiply 0.5 to r_2 and r_3 in the denominator of the second contributions because although necessarily all subjects are on the risk set of the first contribution, there's 0.5 chance of either r_2 or r_3 to be in the risk set of the second contribution to the likelihood. This is the method used as default in R [The15] to handle tied events.

2.2.3 Estimating the Survival Function

If one wants to build the survival function for a new subject s_i based on given covariates X_i, it is necessary to estimate the baseline survival function. The estimator of the baseline survival function is often based on the Breslow's estimator of the baseline cumulative hazard rate [KM03].

Given the $\hat{\beta}$ estimates of the coefficients. Let $t_1, t_2, ...t_N$ be the distinct event times and e_i be the number of events at t_i. The estimator of the cumulative baseline hazard rate, $H_0(t) = \int_0^t h_0(x)dx$.

$$\hat{H}_0(t) = \sum_{i:t_i \leq t} \frac{e_i}{\sum_{r \in R(t_i)} exp(X_r^T \hat{\beta})} \tag{2.26}$$

When there are no covariates, the estimator above reduces to the Nelson-Aalen estimator in equation 2.12. An estimator for the baseline survival function can be

obtained via equation 2.9, where the survival and hazard functions are replaced by the baseline survival and baseline hazard functions respectively.

$$\hat{S}_0(t) = exp(-\hat{H}_0(t)) \tag{2.27}$$

The baseline survival estimate uses the covariates of subjects at risk at time t. However, the baseline survival estimates the survival of a subject s_0 with covariate values $X_0 = 0$. The survival function estimate of a subject s_i with covariate values $X_i \neq 0$.

$$\hat{S}(t) = \hat{S}_0(t)^{exp(X_i^T \hat{\beta})} \tag{2.28}$$

2.2.4 Assessing the Proportional Hazards Assumption

The mentioned method of checking for a pair of subjects if their hazards cross (section 2.2), states that if such pair exists, the proportional hazard (PH) assumption is not met. However, the opposite is not true, i.e. if there is no such pair, one cannot claim the PH assumption hold. Therefore, there is no use to check the hazards of all possible pairs of subjects, instead, further investigations must be performed.

Time-dependent interaction term

By adding to the model an interaction term the dependency of time of a covariate V can be assessed. The interaction term, consists of the product of a variable V alleged constant over time with a time-dependent function $g(t)$. One can use different time functions such as polynomial or exponential decay but often very simple fixed functions of time such as linear or logarithmic functions are preferred [BMD+10]. The measure to be observed is the significance of the coefficient δ of the interaction term. Assuming an univariate model

$$h(t) = h_0(t)e^{\beta V + \delta(V g(t))}$$

The test of statistical significance of δ:

$$H_0 : \delta = 0$$
$$H_1 : \delta \neq 0$$

The hazard ratio for a unit increase in the variable V,

$$HR(t) = \frac{h_{V+1}(t)}{h_V(t)} = e^{\beta + \delta g(t)}$$

The key assumption of the PH Cox model, is that the hazard ratio is constant over time [TG00], however, if it is possible to reject $\delta = 0$, then the $HR(t)$ is a function of time [BMD+10], hence the test of statistical significance of δ is converted to a test of the PH assumption.

Schoenfeld residuals

The score equation 2.19 or the first derivative of the partial log-likelihood function can be seen as the sum of Schoenfeld residuals for one covariate, for each time of failure t_i from t_1 to t_N. When there are no tied events, given the coefficients estimate $\hat{\beta}$ the Schoenfeld residual r_{ik} corresponding to ith failure time t_i and subject s_i for the kth covariate, is composed of the observed covariate value minus the expected covariate value [HLS08].

$$r_{ik} = X_{ik} - \sum_{r \in R(t_i)} \frac{X_{rk} exp(X_r^T \beta)}{\sum_{r \in R(t_i)} exp(X_r^T \beta)} \tag{2.29}$$

Additionally, the Schoenfeld residuals for one subject could be represented by the vector of the residuals for all p covariates, $r_i = \{r_{i1}, r_{i2}, ...r_{ip}\}$. As in the scores, the sum of the residuals for a fixed covariate X_k have to equal zero to satisfy the $\hat{\beta}$ estimates.

$$\sum_{i=1}^{N} r_{ik} = 0 \tag{2.30}$$

Calculating Schoenfeld residuals provide evidence that the coefficients are time-dependent $\beta(t)$. Hence PH assumption is not met, once the $HR(t)$ would be a function of time. Grambsch and Therneau have shown in [GT94] that it is possible to estimate $\beta(t)$.

$$\beta_k(t_i) \approx E(r_{ik}^*) + \beta_k$$

where $E(r_{ik}^*)$ is the expected value of the scaled residual at the ith time failure and the kth covariate. Where the scaled residual is given by,

$$r_i^* = (\mathcal{V}(\hat{r_i}))^{-1} \hat{r_i} \tag{2.31}$$

Furthermore, since the variance tends to be constant over time, the inverse of the variance can be approximated by multiplying the estimator covariance matrix by the number of uncensored events N [HLS08].

$$(\mathcal{V}(\hat{r_i}))^{-1} = N\mathcal{V}(\hat{r_i})$$

$$r_i^* = N\mathcal{V}(\hat{r_i})\hat{r_i} \tag{2.32}$$

The expected value of the scaled residual can be approximated by $E(r_{ik}^*) \approx \delta g(t)$, where $g(t)$ is a specified function of time, and one may perform a statistical test to determine if the δ is zero, in a similar fashion from the "Time-dependent interaction term" strategy, for more details please consult Grambsch and Therneau (1994) [GT94].

For simplicity, assuming a nonproportionality test on a univariate model, the Schoenfeld residuals correspond to a N -vector of residuals as in equation 2.29. Meaning that by plotting $E(r_i^*) + \beta$ over time one can access the nonproportionality of the single covariate $\beta(t)$, i.e. if it vary around zero such that it sums up to zero the PH assumption holds.

2.3 Statistical Prediction

Before performing any predictions, the predictive or prognostic model must be constructed. The beginning of this section will be dedicated to the architecture of prediction models.

The static risk prediction considered in this study using the Cox proportional hazards model fixes the covariate values at a prespecified date often called the baseline time point. To avoid confusion with the baseline hazard where the values of the covariates are zero, the name “cutoff” time point will be used for its intuitive meaning. Concretely, it represents the beginning of follow-up. A distinction between the cutoff time point and the time the first covariate value was observed, must be made clear; although both time points can match, they are not necessarily the same, for instance, one can set the cutoff point after a particular state of interest was observed (i.e. sometime after the individual entered the study), for instance, the initial diagnosis of a clinical condition, or the occurrence of a medical procedure such as kidney transplantation [GL17].

The intuition and motivation for naming "cutoff time point", observation of covariates are only up until cutoff, henceforth one may not access covariate values, but rather use past values at or before the cutoff to predict events occurring after the cutoff, referred as an evaluation period. In the case of proportional hazards, these covariates are fixed at cutoff, i.e. for one subject, there is a static vector of p covariate values, where p is the number of covariates. As stated by Greene and Li (2017),

> *event times during the evaluation period after time 0 are related to the baseline covariates under a Cox regression model with two components:*
>
> 1. *a baseline hazard function, which defines the risk for the event at each time during the evaluation period for one arbitrary set values for the baseline covariates, and*
> 2. *a multiplying factor, which proportionally modifies the baseline hazard depending on the actual values of each patient's baseline covariates*

In this setting, the Cox regression coefficients and the baseline hazard function, obtained with the training set, are used to predict the shape of the survival function of a new patient with its cutoff covariates in the multiplying factor. This obtained survival

refers to survival of the new patient beyond the cutoff time, over the evaluation time.

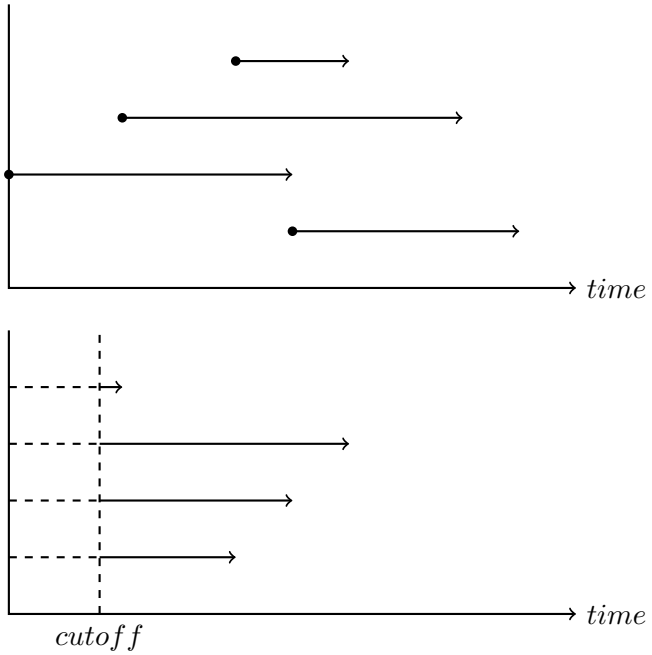


Figure 2.4: Selection of the cutoff over subject's follow-up time

Imagine four subjects follow-up time were observed, in reality, the entry of subjects in the study tend to be staggered and their calendar time rarely match, as can be seen on the top figure 2.4. On the bottom figure, instead of using the calendar time, the subject's follow-up refers to the time from the entry in the study, hence they all match on the time axis.

The cutoff time point represents the boundary of what is "known" to the prediction model, i.e. only covariate values up until cutoff (represented by the dashed line in figure 2.4) may be used during the training of the prediction model. A suitable model should be able to provide reliability measures such as the estimated survival function that are compatible with the follow-up time extension from cutoff (represented by the solid line). As an indicator of how reliable the model is, the survival of the subject on the top of the bottom figure should approximate zero quicker than the other three subjects below [GL17].

The selection of the cutoff time point, therefore, must be made before any observed event (represented as the end of the arrow), otherwise, if a subject experiences an event before the selected cutoff time point it would not be available for prediction, once it is already known when the subject failed.

The prediction problem arises when one wants to predict for a future random variable

T given covariates X and a reference population, used to estimate the coefficients $\hat{\beta}$ and a baseline function $h_0(t)$. There are two types of predictors, point predictors, and probabilistic predictors [Yua08]. Point predictors provides an exact estimation of the expected time of failure $E(T_i)$ of a single subject $subj_i$. Probabilistic predictor gives the probability of survival $P(T_i \geq t|X_i)$ of a subject exceeding a given time t i.e. the survival function; or prediction intervals for $subj_i$, $P(t_l \leq T_i \leq t_r|X_i)$.

Despite the unreliability of point predictors as discussed in section 1.1, the Literature Review, will be presented for the sake of completeness, however, they should not be used to validate a prediction model.

2.3.1 Median Time

The most common point predictor is the first 2-quantile or median survival time, i.e. the time when the survival probability has dropped to half. [Yua08]

2.3.2 Expectation of Life

The expected value of T_i by definition would be the following integral,

$$\mu = \int_0^\infty t f(t) dt \tag{2.33}$$

with integration by parts setting $u = t$ and $dv = f(t)$ then

$$\begin{aligned} \mu &= \Big[tF(t)\Big]_0^\infty - \int_0^\infty F(t)dt \\ &= \int_0^\infty 1dt - \int_0^\infty F(t)dt \\ &= \int_0^\infty 1 - F(t)dt \end{aligned}$$

with equation 2.2 we obtain

$$\mu = \int_0^\infty S(t)dt \tag{2.34}$$

Therefore, the second point prediction metric can be simply calculated by computing the integral of the survival function.

2.4 Extended Cox model

The survival analysis situation involving a covariate that changes over time is also supported by the Cox model, the progression of covariate values is often called the "covariate path" [TG00]. The figure 2.5 bellow displays two subjects, one represented by the dashed line and another by the solid line.

For the sake of simplicity, if one performs a univariate regression model, measuring the influence of total cholesterol (in mg/dL) in heart attack events. Since the values of cholesterol could be measured on several occasions yielding different values, it could be modeled as a covariate dependent on time. As seen in the figure 2.5, $subj_1$, represented by the solid line, suffered a heart attack after 170 days and $subj_2$, represented by the dashed line, after 200.

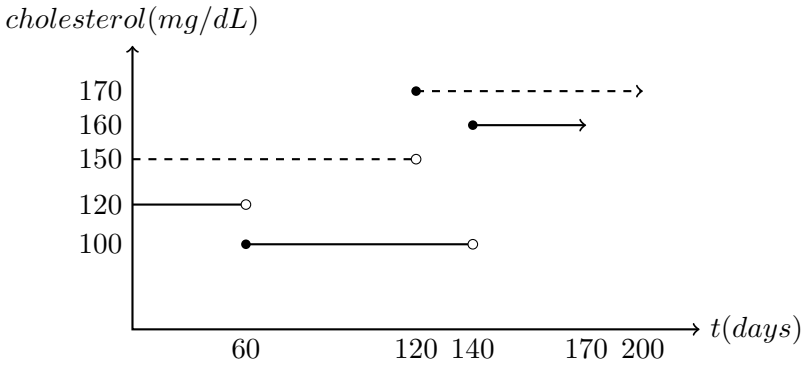


Figure 2.5: Cholesterol covariate paths of two subjects

The time-dependent hazard equation separates coefficients β of p_1-time-independent variables from δ of p_2-time-dependent variables.

$$h_i(t) = h_0(t)exp\Big(\sum_{j=1}^{p_1}\beta_i X_{ij} + \sum_{k=1}^{p_2}\delta_j X_{ik}(t)\Big) \tag{2.35}$$

In this model, although now a time-dependent covariate value change over time, there is only one coefficient associated with each time-dependent covariate. In the case of the heart attack example above, the hazard would have a single coefficient value δ for the cholesterol covariate,

$$h_1(t) = h_0(t)exp(\delta X_1(t)) \tag{2.36}$$

Nevertheless, it is possible to have a time-dependent coefficient $\delta(t)$, Schoenfeld residuals can be used to check if coefficients should be regarded dependent of time. For simplicity, over this section it is assumed that δ is constant.

2.4.1 Assumptions of the Extended model

Firstly, it is important to point that the effect of a time-dependent covariate $X_j(t)$ on the hazard at time t, depends only on $X_j(t)$, any other covariate value $X_j(t')$ for $t' > t$ or $t' < t$ will not contribute to the calculation of $h(t)$. However, the calculation of the survival function may use observations from the past, $t' < t$, since as shown in equation 2.9 the survival can be obtained by integrating the hazards until t; in any case *"you cannot look into the future"* [TCA19].

The time-dependent Cox models work by assigning the risk score $exp(X_j(t)^T\beta)$ for each subject in such a way to maximize the likelihood of obtaining the observed order of events, and it does so by using in the risk score with covariate values at the event time.

It is also assumed that each time-dependent covariate behaves as a function of time, i.e. the time intervals for one subject have only one associated covariate value, with no interval overlaps, as in figure 2.5. On a non-recurrent Cox model (each individual only experience an event once), there is no concern about the correlation between two covariate values for the same subject at different time points.

2.4.2 Hazard Ratio for the Extended model

Once introduced time-dependent variables the PH assumption is no longer met, because the hazard ratio now varies over time [KK12], i.e. if we write equation 2.15 with a single time-dependent covariate $X(t)$.

$$HR(t) = \frac{h_1(t)}{h_2(t)} = exp((X_1(t) - X_2(t))^T\beta)$$

Because in this case the hazard ratio clearly depends on time, it is customary to drop off the "proportional hazards" from the name of this model, although some still loosely call PH. Accounting for multiple covariates, including some that may not be time-dependent, the $HR(t)$ takes the shape,

$$HR(t) = exp\left(\sum_{j=1}^{p_1}\beta_j(X_{1j} - X_{2j}) + \sum_{k=1}^{p_2}\delta_k(X_{1k}(t) - X_{2k}(t))\right) \tag{2.37}$$

One must be careful on an attempt to calculate the $HR(t')$ on a specified t', since would lead to a constant value, e.g. assuming a single time dependent covariate *cholesterol* and the example displayed on figure 2.5 assuming an estimated constant coefficient $\hat{\delta}$,

$$HR_{12}(t = 150) = \frac{h_1(t = 150)}{h_2(t = 150)} = exp(\hat{\delta}(160 - 170))$$

Although the outcome is the constant $e^{-10\hat{\delta}}$, one must not assume the PH assumption holds, once the hazard ratio differs over time for the same two subjects, for instance $HR_{12}(t = 150) \neq HR_{12}(t = 120)$.

2.4.3 Extended Likelihood

The calculation of the likelihood function for the Extended Cox case is quite similar to time-independent covariates (PH) case, however, one must account for the time-dependent covariates when calculating the hazards of subjects in the risk set. The first thing to notice is how the contributions to the likelihood of the same subject at risk at different times may change. Assume two additional subjects in figure 2.5, $subj_3$ and $subj_4$ which failed at 50 and 110 respectively. Since $subj_1$ is at risk at both failure times, when building the likelihood terms $L_3^{t=50}$ and $L_4^{t=110}$ correspondent to two failure times 50 and 110, one need to account in the denominator of equation 2.18 to the risk score of $subj_1$. Let $\mathcal{R}_i^t$ be the risk score of subject i at time t.

	contribution of $subj_1$
at 50	$\mathcal{R}_1^{50} = exp(\hat{\delta}120)$
at 110	$\mathcal{R}_1^{110} = exp(\hat{\delta}100)$

Therefore, unlike in likelihood estimation of the PH model, where each subject $subj_i$ provides the same risk score contribution $\mathcal{R}_i$ to the denominator of equation 2.18 when at risk. An important detail when building the likelihood function in the extended case is, dependent on the time t' associated with the likelihood term $L^{t=t'}$, subject i may at another t'' time, contribute differently $\mathcal{R}_i^{t'} \neq \mathcal{R}_i^{t''}$ when $t' \neq t''$.

2.4.4 Limitations of the Extended Cox model in Prediction

As shown in equation 2.35, the parametric part of the hazard model is now dependent of time, therefore if the future hazard of a subject at t_f after cutoff is sought; the covariate values at $X(t_f)$ must be known. If the covariate values are not predicted, which implies training an additional model solely for predicting covariates values, the hazard of a subject can only be calculated up until cutoff time. The hazard, thus the survival can be computed before cutoff, this computation nonetheless have no use; the subject's survival exceeds the cutoff as established by the prediction model architecture, instead, $S(t) = 1$ for every $t < t_{cutoff}$.

An argument for this limitation is the different paradigm of extended models, the modeled relationship of interest is the association of the patient's instantaneous risk with the patient's most recently measured covariates; instead of the relationship between covariates at cutoff with future risk [GL17].

If performance of extended versus static proportional hazard models would be compared, one should expect to obtain better results from extended model not because of its increased complexity by considering time-dependent covariates but specially because

the extended model has a different task, once "it is less difficult to estimate the present risk than it is to predict future risk" [GL17]. For this reason, extended models won't be further considered in this study.

2.5 Validation

Validation is essential to assess performance of prediction models. Specially in medical contexts, models which did not underwent validation are not adequate to enter clinical practice [MV09]. By "validation" it is normally meant the assessment of the overall model performance; i.e. its discriminative ability and its prediction accuracy. There is a variety of methods reported in the literature to calculate such measures [SVC+09]. In this study, although discrimination will be performed separately, the prediction accuracy will be assessed as a measure of both accuracy and discrimination. Moreover, the overall fit of the model will be described.

2.5.1 Discrimination

Discrimination, sometimes known as "separation", is the extent to which risk estimates from a model characterize different subject's outcomes. Subjects predicted to be at higher risk should exhibit higher event rates than those deemed at lower risk [RA13].

Concordance index [HCP+82] (c-index) is one of the methods that can be used to measure discrimination. A pair of subjects is called concordant if the subject who experiment an event earlier is assigned a higher risk by the prediction model than the one who experiences the event later. Inversely, a pair is also concordant if the subject who experiences an event later is assigned the lower risk. Otherwise, the pair is called discordant.

In the case of Cox proportional hazards, the risk of a subject i is the prognostic index $X_i^T\beta$ (not to be confused with the risk score $exp(X_i^T\beta)$). Hence, the risk factor suffices when comparing risks of individuals. Let t_k be the observed time of failure of subject k, a pair i and j of subjects is concordant if

$$I(X_i^T\beta > X_j^T\beta, t_i < t_j) + I(X_i^T\beta < X_j^T\beta, t_i > t_j) \tag{2.38}$$

where $I(A, B)$ is an indicator function which is 1 if conditions A and B hold and 0 otherwise. The concordance is simply the percentage of concordant pairs over all pairs C_2^n, where n is the number of subjects. According to Austin and Steyerberg [AS17], the c-statistic is equivalent to the area under the receiver operating characteristic (ROC) curve, additionally, Hanley & McNeil [HM82] stated that ROC or equivalently the c-index of 0.5 represents no apparent accuracy while 1.0 represents perfect accuracy.

2.5.2 Overall Performance

The distance between actual and predicted outcomes is a common strategy to quantify the model performance of statistical regression models [SVC+09], such as the Residual Sum of Squares in linear regression. For continuous time outcomes, the distance $I(t) - P(t)$ is the difference between the indicator function $I(t)$, 1 if the subject still at risk at time t and 0 otherwise; and the predicted survival probability $P(t)$ at time t.

The distance between actual and predicted outcomes can also be seen as evaluating the "goodness-of-fit" of a model, with superior models having lower distance values. The main difference between goodness-of-fit and predictive performance, is the context which they are evaluates, the former is usually evaluated in the same data while the latter requires either new data or cross-validation [SVC+09]. Hence, the Brier score when computed on a cross-validating setting is considered the evaluation of the predictive performance, whereas after the final model is defined, the recompuputed Brier score corresponds to the goodness-of-fit.

This measure accounts not only for the discrimination by having smaller values of Brier score when the model is able to distinguish high and low risk of subjects, but also for calibration. Calibration describes how accurately the estimates or predictions of survival from a model reflect the survival in the observed data [RA13]. The survival in the observed data is simply the status function which drops from one to zero when the subject experiences an event. Moreover, the Brier score is a quadratic scoring rule, it is calculated by the squared distance between the subject's observed status and the predicted survival probability. In this study, only the computation of the no censoring case was necessary; in this case, the Brier score can be computed as,

$$BS(t) = \frac{1}{n}\sum_{i=1}^{n}(I(t_i > t) - \hat{S}(t|X_i))^2 \tag{2.39}$$

where t_k be the observed time of failure of subject k and n the number of subjects. The values of the Brier score can be interpreted as the loss which is incurred when the prediction $\hat{S}(t|X_i)$ is issued to a patient whose true status is $I(t_i > t)$ [GCS08]. Additionally, if an overall measure at all times is necessary, one can integrate the Brier score [KR08]. Let t_M be the latest observed time.

$$IBS(t) = \frac{1}{max(t_M)}\int_0^{t_M} BS(x)dx \tag{2.40}$$

2.5.3 Overall fit

Let T be the random variable that describes the failure time event. If T has survival function $S(t)$ by applying probability integral transform (also known as the universality of the uniform), $S(T)$ is a random variable with a uniform distribution $U(0, 1)$. Let

$F_{S(T)}(x)$ represent the cumulative distribution function of $S(T)$.

$$F_{S(T)}(x) = \begin{cases} 0, & x < 0 \\ x, & 0 \leq x < 1 \\ 1, & x \geq 1 \end{cases} \tag{2.41}$$

It is desirable to find the distribution of the cumulative hazard $H(T) = -log(S(T))$ random variable. Hence, one must find the cumulative distribution function of $H(T)$, $F_{H(T)}(y)$. Because $S(T)$ take values between 0 and 1, the random variable $H(T)$ must take values from 0 to ∞, thus $F_{H(T)}(0) = 0$. From 0 to ∞ the following holds:

$$\begin{aligned} F_{H(T)}(y) &= P(-ln(S(T)) \leq y) \\ &= P(S(T) \geq e^{-y}) \\ &= 1 - P(S(T) < e^{-y}) \\ &= 1 - e^{-y} \end{aligned} \tag{2.42}$$

Because $S(T)$ is a continuous random variables the inequalities do not have to be strict, since the probability for any exact value is equal to zero. For this reason, it is possible to use the uniform cumulative distribution function definition on 2.41 on $P(S(T) < e^{-y})$. The resulting cumulative distribution function on equation 2.42 is an exponential distribution with rate parameter 1 [HLS08].

The estimated cummulative hazard $\hat{H}(t)$, thus must agree with the theorectical true cummulative hazard distributional characteristics obtained. Given estimated $\hat{\beta}$ parameters, and considering no event ties, the Cox-Snell [CS68] residual is defined.

$$r_i = \hat{H}(t) = \hat{H}_0(t_i)exp(X_i^T \hat{\beta}) \tag{2.43}$$

where the baseline cumulative hazard can be the Breslow estimation shown in equation 2.26. If the model is correct and $\hat{\beta}$ is close to the true value of β, then r_i's should behave as a sample from a unit exponential distribution [KM03].

If the residuals r_i's were drawn from an unit exponential distribution $Exp(1)$, then $F(r_i) = 1 - e^{-r_i}$ and the survival function from equation 2.2 is $S(r_i) = e^{-r_i}$. Inversely, from equation 2.9, the cumulative hazard can be calculated by the negation of $ln(S(r_i))$.

$$H(r_i) = r_i \tag{2.44}$$

If the model is correct, then the Nelson-Aalen estimator (equation 2.12) of the cumulative hazard applied on the residuals (instead of on time) versus the residuals should be a straight line through the origin with slope 1 [KM03].

CHAPTER 3

Case Study

This section explains the data sources as well as give an overview of the data structure. Furthermore, we will present how the models were built as well as explain the selection of such models using as a background the theory presented in the previous Methodology chapter.

3.1 Data Source and Characteristics

It is of interest to the construction contractor to gather maintenance logs, collected by mechanics for accounting reasons. When a repair needs to be performed, mechanics have to inspect the machine and are often unsuccessful to find the faulty component. The use of the information contained in logs could drastically improve diagnostics, which we aim to explore.

Moreover, the data at hand could be useful to provide prognosis, potentially preventing failure. Although the data was not collected to be used on Predictive Maintenance models, but rather for generating invoices and managing the demand for repairs, they contain desirable but not sufficient information to train such models. Additional to the fixed information e.g. the manufacturer's name, the year of manufacture, the year of purchase, whether it was bought new or used; the mechanics also took notes on the date of repair, working hours, which component received the repair and its costs. Note that the reported date, as well as the other notes, are made at the date of repair not the date of failure which induces biases, especially in this study since the delay from failure to repair was inconsistent. A fundamental difference between the quality of repair is made.

Instandsetzung reactive repair after breakdown preventing the machine from operating, e.g. replacing the motor.

Instandhaltung a routine maintenance, normally a check, verification; or a minor intervention e.g. changing oil or lubrication.

Acknowledge that it is not the description of the repair that define the quality of repair. The same description, e.g. replacement of the gear, could either be applied for precautionary measures although still functional (Instandhaltung); or because the gear failed, preventing the operator to move the machine (Instandsetzung). In the rest of this study, failure may be understood as an Instandsetzung event.

To explore the factors causing the degradation of machines, and optimally perform predictive maintenance, it is evident the need for additional data which potentially correlates with observed breakdown.

This study focused on the analysis of 18 paver machines manufactured by Voegele which telematics data were provided. Sensors installed by the manufacturer recorded the state of the machine over an interval and sent the data to the contractor through a telecommunication device. The new source of data provided daily scans of attributes, such as

- duration of recording interval in seconds (Aufzeichnungsintervall Dauer[s])
- duration of engine running in seconds (Motor an)
- working hours (Betriebstunden)
- driving velocity kilometres per hour (Geschwindigkeit)
- motor rotation speed in 1/minutes (Motordrehzahl)
- load factor M1 (Lastfaktor M1)
- total fuel consumption in litres (Kraftstoffverbrauch gesamt)
- coolant temperature C° (Kuehlmittel temperatur)
- total time of idle engine in hours (Motors leerlauf)
- distance in kilometres (Wegstrecke)
- fuel rate litres per hour (Kraftstoffrate)
- year of construction (Baujahr)

. The covariates were recorded over time intervals, for the relevant ones, the average, maximum and absolute values were stored of over those intervals. Therefore, one could explore the usage of average or maximum covariate values, depending on the relationship of the covariate to the outcome variable at event time.

Although there were only 18 pavers, more than 18 repairs occurred since one machine failed multiple times in several components, additionally, in one failure event one or more repairs were performed, each in a correspondent component. Overall, there were 19 components, each containing a group of corresponding subcomponents representing

a specific part of the machine. Hence, a log reporting an Instandsetzung event on a component does not uniquely identify the faulty component which yielded the breakdown, but rather indicate the component the repair was performed.

To define failure time precisely in this case study, as described in section 2.1, the time origin, scale and meaning of failure must be specified. Firstly the meaning of failure will be explained, followed by the scale and the time origin.

The distribution of Instandsetzung repairs per component present in the logs can be seen in following table 3.1, displaying the frequency count of instandsetzung repairs in each component.

In case studies with abundant data, the ideal method to select the critical component would be by assessing the frequency of failure over downtime associated with failure for relevant components [LWZ+14]. The choice of the component in this study was constrained, however, by the limited amount of data available, since there are only 18 machines. Therefore, the only component selected were those which had at least one Instandsetzung repair on each machine, hence we may analyze, Material Conveyance and Screed, bold columns 13 and 12 respectively, vide table 3.1. General, code 20 in the table, was the component name given when a minor intervention sufficed to repair the machine after the breakdown, but the actual component in the machine varied, therefore it will not be considered.

Although the selection of the component constraint the problem it does not fully specify it, once a failure can happen in multiple parts of such component. To finally specify the meaning of failure, one must select within one component one repair description.

Every machine failed on Material Conveyance and Screed components yet it was cumbersome to select within those components a specific Instandsetzung repair description that occurred over all machines. Given this restriction there are two possibilities ahead, the first is to accept using less than 18 machines although building a model to analyse exclusively repair after failure (Instandsetzung events). The second, to accept Instandhaltung repairs, and assume that such repairs would only be performed for a good reason, i.e. the part replaced displayed some problem, although not serious enough to cause the machine breakdown. In this last case, notice that some bias is introduced, namely, the prediction is not made on the failure event, but on the time of the Instandhaltung repair.

In this work the second solution was considered, allowing predictions of Instandhaltung repairs, the chosen repair description was the replacement of a specific part called “Bunkergummi” on the component Material Conveyance. Therefore any of the three repair descriptions “Bunkergummi erneuert”,“Bunkergummi erneuert/umgebaut” or “Bunkergummi getauscht/erneuert” refers to the same repair activity and was considered an event of interest. The reason these descriptions are aggregated is because they refer to the same repair activity, i.e. to exchange the old “Bunkergummi” for a new one. Note that it is important to be able to aggregate repairs that been given different log descriptions by mechanics, nevertheless, those refer to the same repair on the same machine part; i.e. in this context, these verbs are considered synonyms.

Regarding the scale, the outcome variable targeted for prediction was working hours

	01	**13**	02	81	15	08	07	06	05	03	30	20	09	10	14	11	**12**	04	60	MachCount
id-5	3	**24**	0	0	0	4	1	3	0	5	11	53	9	0	0	1	**38**	5	0	157
id-8	0	**26**	0	0	11	0	0	0	0	8	0	43	10	0	0	0	**66**	7	1	172
id-6	0	**43**	1	0	10	1	1	11	0	9	9	156	9	0	1	3	**110**	14	1	379
id-3	5	**47**	0	3	1	0	1	4	1	12	16	77	16	2	0	0	**56**	7	0	248
id-11	1	**28**	0	1	9	0	0	1	0	4	5	42	9	0	0	1	**48**	4	0	153
id-9	0	**27**	0	0	4	0	0	2	0	4	0	25	8	0	1	0	**39**	3	1	114
id-4	1	**28**	0	0	11	2	1	4	0	4	5	81	0	1	1	0	**44**	8	2	193
id-0	5	**35**	1	0	8	5	1	10	0	15	8	215	16	1	0	0	**126**	30	5	481
id-12	0	**30**	0	0	0	0	0	6	0	0	0	99	3	1	0	0	**107**	5	2	253
id-14	1	**21**	0	0	15	0	1	4	0	7	4	39	2	2	1	0	**42**	5	0	144
id-2	0	**14**	0	0	0	0	0	5	0	2	5	44	8	0	0	0	**28**	3	1	110
id-10	3	**24**	0	0	5	0	1	1	0	5	0	37	6	1	0	0	**66**	3	0	152
id-13	0	**5**	0	0	0	0	0	2	0	3	1	24	2	0	0	0	**39**	0	1	77
id-1	0	**32**	1	1	2	1	2	10	0	23	6	96	13	1	3	0	**106**	12	3	312
id-15	0	**8**	0	0	0	0	1	0	0	5	2	51	9	0	0	0	**29**	3	3	111
id-17	0	**4**	0	0	0	0	0	0	0	0	0	34	0	0	0	0	**8**	0	0	46
id-16	0	**11**	0	0	0	0	0	0	0	4	6	26	2	1	0	0	**12**	2	2	66
id-7	0	**24**	0	0	2	0	1	3	0	4	6	68	4	0	0	0	**42**	6	0	160
CompCount	19	**431**	3	5	78	13	11	66	1	114	84	1210	126	10	7	5	**1006**	117	22	3328

Table 3.1: Overall distribution of repairs per component

at the time of failure instead of simply of time the repair was made. The "working hours" are simply the number of hours the machine has operated from purchase to the current time. Two main issues motivate this choice of scale,

1. There is no information on which day the failure happened, but only the month, and as already mentioned, the logs contain the time when the repair arrived at the construction site instead of when the failure occurred.

2. When using the repair time as the outcome, the hazard would still increase while the machine was idle or already failed.

These problems are solved when using working hours as the time scale. Both considered components (Screed and Material Conveyance) are essential for the operation of the machine, hence after a breakdown, the working hours would not increase, providing the precise condition of the outcome of the machine in the moment of failure. Unlike humans, machines shut off, when predicting calendar time it is cumbersome to model periods which the machine was not functioning. During such periods, the hazard should not grow, since the machine is preserving its state and environmental covariates, such as humidity and temperature, are not considered in this study.

Figure 3.1 shows the preprocessed working hours over time (details on preprocessing will be given in the next section) until the first replacement of the Bunkergummi on the Material Conveyance component. As can be seen in figure 3.1, around January a lack of activity can be seen due to winter when machines do not normally operate. The start of the the study was in the beginning of the month with the earliest date of purchase, on 1st of April. Telematics data, nevertheless, only started recording in the beginning of June 2014. Therefore there was missing information on the working hours of machines 0 and 1. Many problems were spotted in the telematics measurements of working hours, such as

- outliers,
- non-monotonic behaviour and
- missing information.

3.2 Preprocessing Data

Significant effort was necessary to correct not only the problems mentioned in the previous section about the working hours, but also in the covariates. The preprocessing was done in Python using Jupyter Notebook 4.4.0 and mainly pandas, datetime and matplotlib libraries. Firstly, results of the preprocessing will be exhibited then preprocessing of the working hours will be explained and finally of the other covariates.

To understand the behavior of the telematics data at hand and explore similar inconsistencies, plotting each relevant covariate over calendar time was useful (the colors

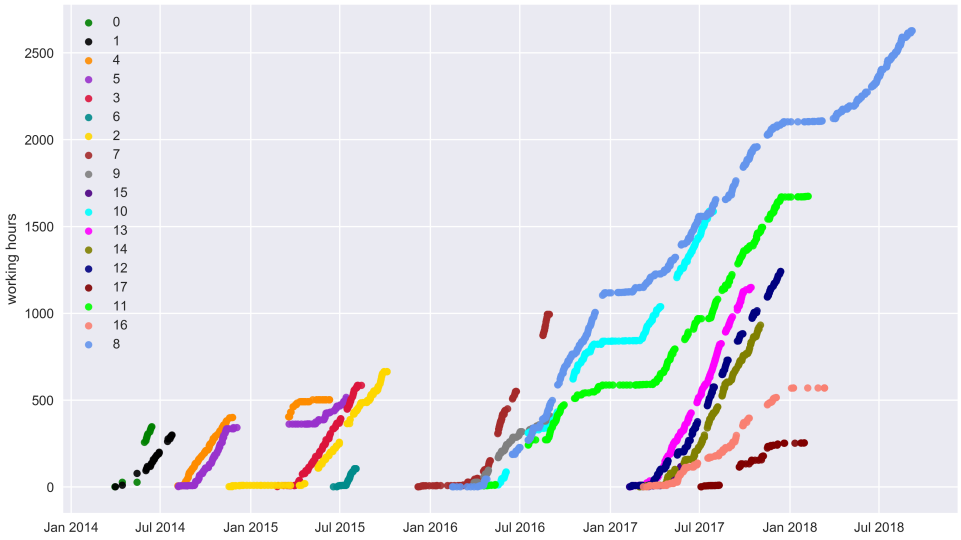


Figure 3.1: Working hours up until the first repair on Bunkergummi on component Material Conveyance

from machines match over the plots). After understanding the underlying behavior, conclusions were drawn and covariates could be preprocessed, figures 3.2 and 3.3 displays the resulting preprocessed covariates over working hours. Notice the values were truncated on the first replacement of "Bunkergummi".

Displaying all covariates for all machines in a lucid fashion was not trivial, machines had to be omitted. Only machines 1,3 and 4 were displayed in 8 of the 11 plots, the covaviates on figures 3.2 and 3.3. The selection was based by plotting all of them at once on the same plot how reckognizing the standard behavior, then selecting the outliers as well as a few machines that epitomizes the behavior. For instance, when depicting the average load factor M1 in 3.3, it is possible to see that machines 3 and 4 follow the pattern while machine 1 represents the outlier, although it is not easy to make this distinction in every plots. Moreover, the selection was constrained to a single set of machines over all plots.

Since the work hours until breakdown is the outcome variable, it should be a monotonic function over time, behaviours such as the one in figure 3.4 should be corrected before building the model.

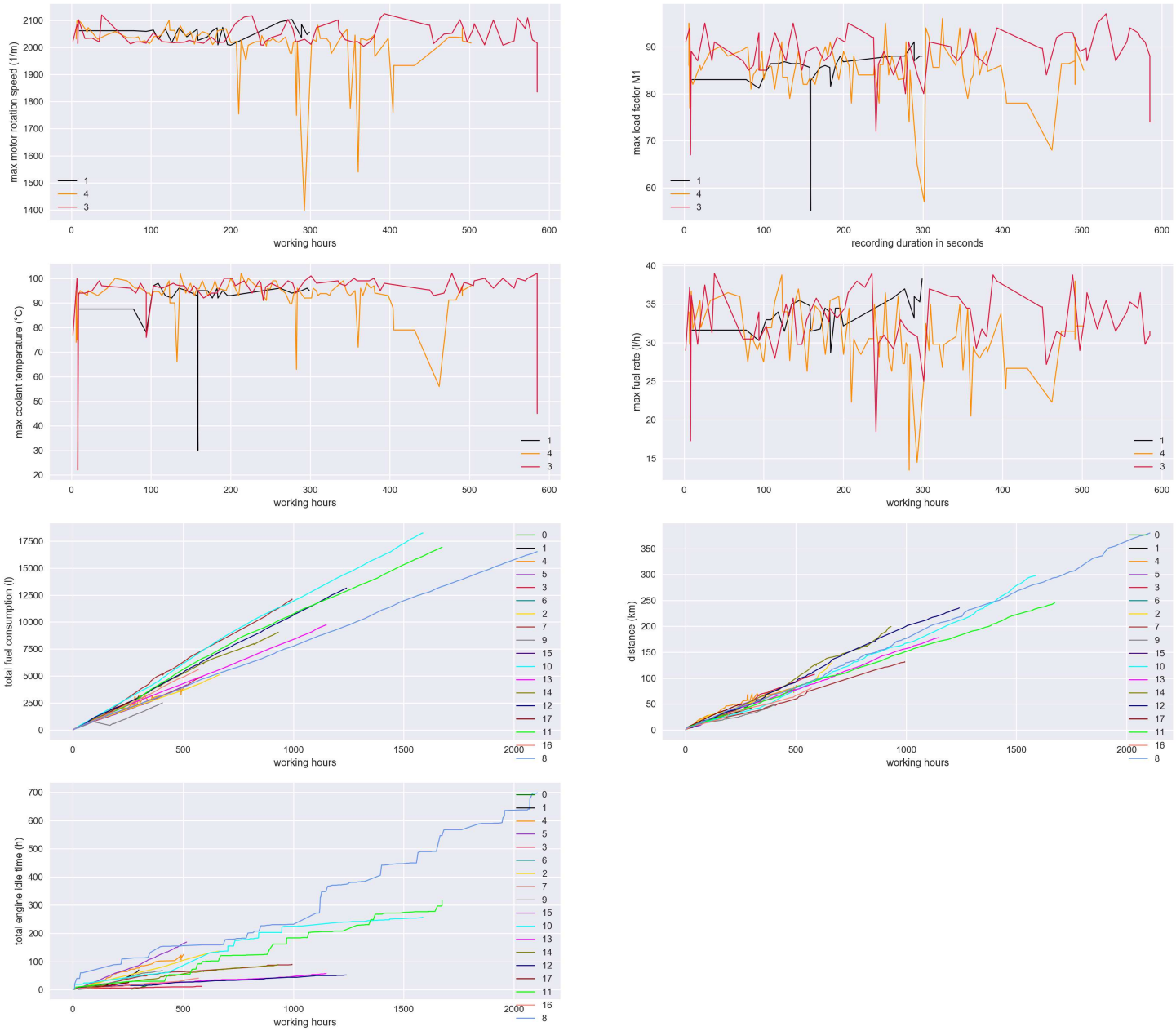


Figure 3.2: Plot of relevant maximum covariates

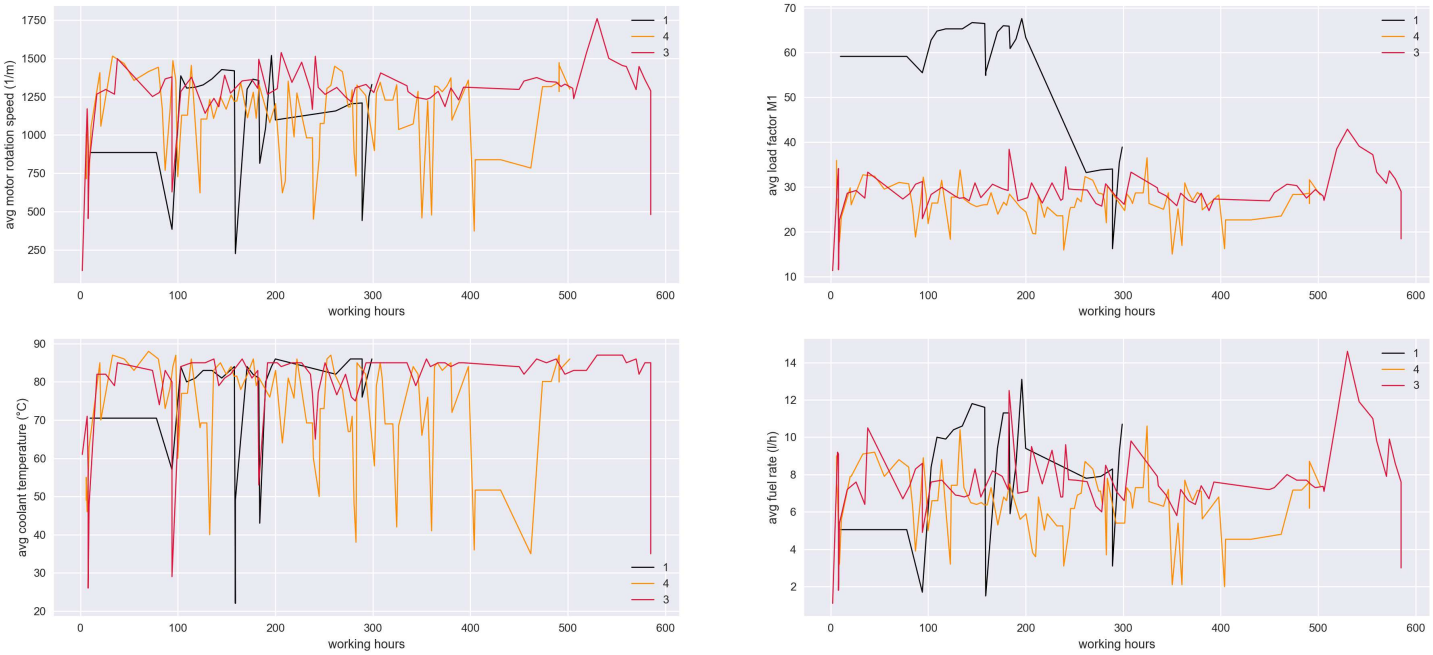


Figure 3.3: Plot of relevant average covariates

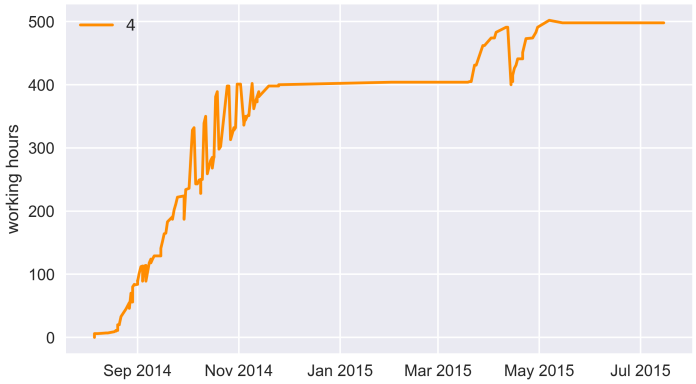


Figure 3.4: Anomaly in working hours of machine four.

Treatment was required, for instance, to remove outliers e.g. a spike displayed on machine one around December 2014 and a sudden jump in machine 4 around December

2015. Spikes or outliers were easily spotted by human eye and manually removed. Anomalies such in figure 3.4 were simply interpolated between two points which the data appeared to be stable, in the case depicted in figure 3.4, values from first of September to the thirteenth of November, using linear interpolation from the pandas library.

Nevertheless, it was not always trivial to determine the behavior and classify as an anomaly or an outlier, the raw data also displayed anomalies such as the one seen in figure 3.5. In the picture there are two parallel trends, to define the trend of the working hours, additional data from the contractor was required. The additional data was superimposed over the former sensor data; the trend which did not coincide was removed. Thereafter, an algorithm to enforce monotonicity was executed, by keeping track of the highest work hours and replacing values lower than the current highest by the highest.

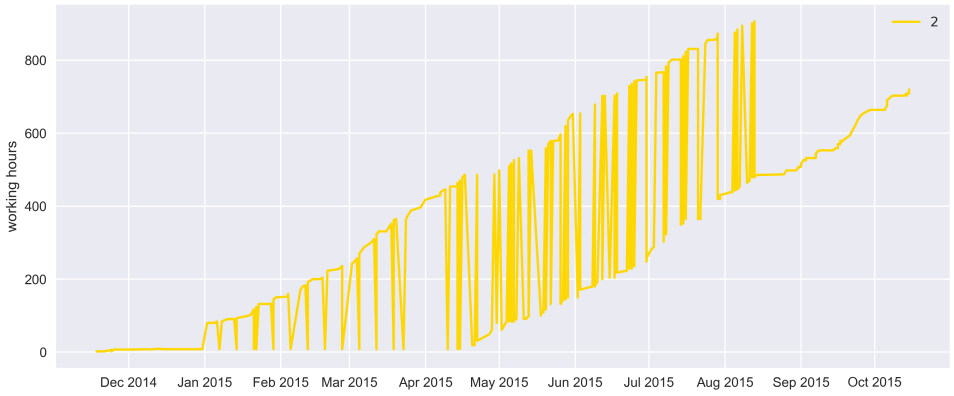


Figure 3.5: Anommaly in working hours of machine two.

Additional problems were also found in the covariate values. Covariates were divided into two categories: cumulative, namely, total fuel consumption, distance and total engine idle time and the rest, noncumulative. At a first moment, the preprocessing of the noncumulative covariates will be presented, then the preprocessing for the cumulative ones.

When plotting the covariate values over the duration of the recording interval, a pattern was observed. For durations lower than 10000 seconds, the covariate values displayed high variance and stabilized as the recording duration increased. One example of such behavior can be seen in figure 3.6, although this pattern consistently appeared overall noncumulative covariates.

Covariates values with such low durations varied too much, therefore the quality of the information recorded by the sensor appeared to be compromised. For this reason, all such covariate values were removed.

The bulk of the preprocessing of covariates was interpolating gaps, i.e. intervals with covariate values zero. Despite not exhibited in figures 3.2 and 3.3, all covariates displayed the same behavior when analyzed over working hours, their values often dropped to zero. Every drop was defined by the two points p_1, p_2 that forms the gap, and its size the amount of working hours the covariate remains zero. Once defined, every gap size bigger than the average gap size was treated. For example, machine 8 and covariate maximum load factor M1, every gap bigger than 4.81 work hours were treated. The selected gaps would then be treated by taking the average of the points in the neighborhood, work hours intervals $[p_1 - \gamma, p_1)$ and $(p_2, p_2 + \gamma]$. The value γ determined the size of the neighborhood, it was a single constant for every covariate, and it was found iteratively by slowly increasing the value such that all selected gaps would be appropriately interpolated, i.e. to find the smallest neighborhood which was big enough so it contained at least one observation excluding p_1 and p_2; γ around 10 was found to be appropriate.

Finally, since the cumulative covariates approximate a linear function over work hours, a linear regression was performed. Note that for zero working hours the noncumulative covariates should all have value zero, therefore the intercept must be zero, the values of the slope, however, were stored and used when building the model. Figures 3.2 and 3.3 depict the raw cumulative covariate values.

3.3 Fitting a Cox PH prediction model

The work hours at the time event varied from 105 to 2627 work hours overall machines. Therefore the cutoff time had to be set before 105 hours, however, no information about the machine's life is used to build the model as the cutoff hours approximate zero. Therefore, the cutoff time was set to 100 work hours. Thus, all information on each machine before 100 working hours is available for the construction of the prediction model.

For the maximum and average noncumulative covariates, the average and the maximum of the covariate values from 0 to 100 work hours were used to build the model. Therefore, note that the four covariates Coolant Temperature, Motor Rotation Speed, Load Factor M1 and Fuel rate, have the maximum and average values of the sensor values, and now the average and maximum values over the work hours from 0 to 100. For the cumulative ones, once the slope was stored, the covariate values for 100 working hours were obtained from the linear function. Therefore all each covariate is represented by a single value as expected by the PH Cox model. Adding to the all of the mentioned covariates, the construction year of each machine was considered.

Two datasets were produced, one with all 20 covariates, and another containing only the maximum values from 0 to 100 working hours from the maximum noncumulative covariates; and average values from 0 to 100 working hours from the average noncumulative covariates, in total 12 covariates. The model was build on a leave-one-out cross-validation setting. Because of the small number of machines, this setting permitted using largest possible sample to train the model. This way, every machine is left out from the training sample once, and its covariates are used once for prediction.

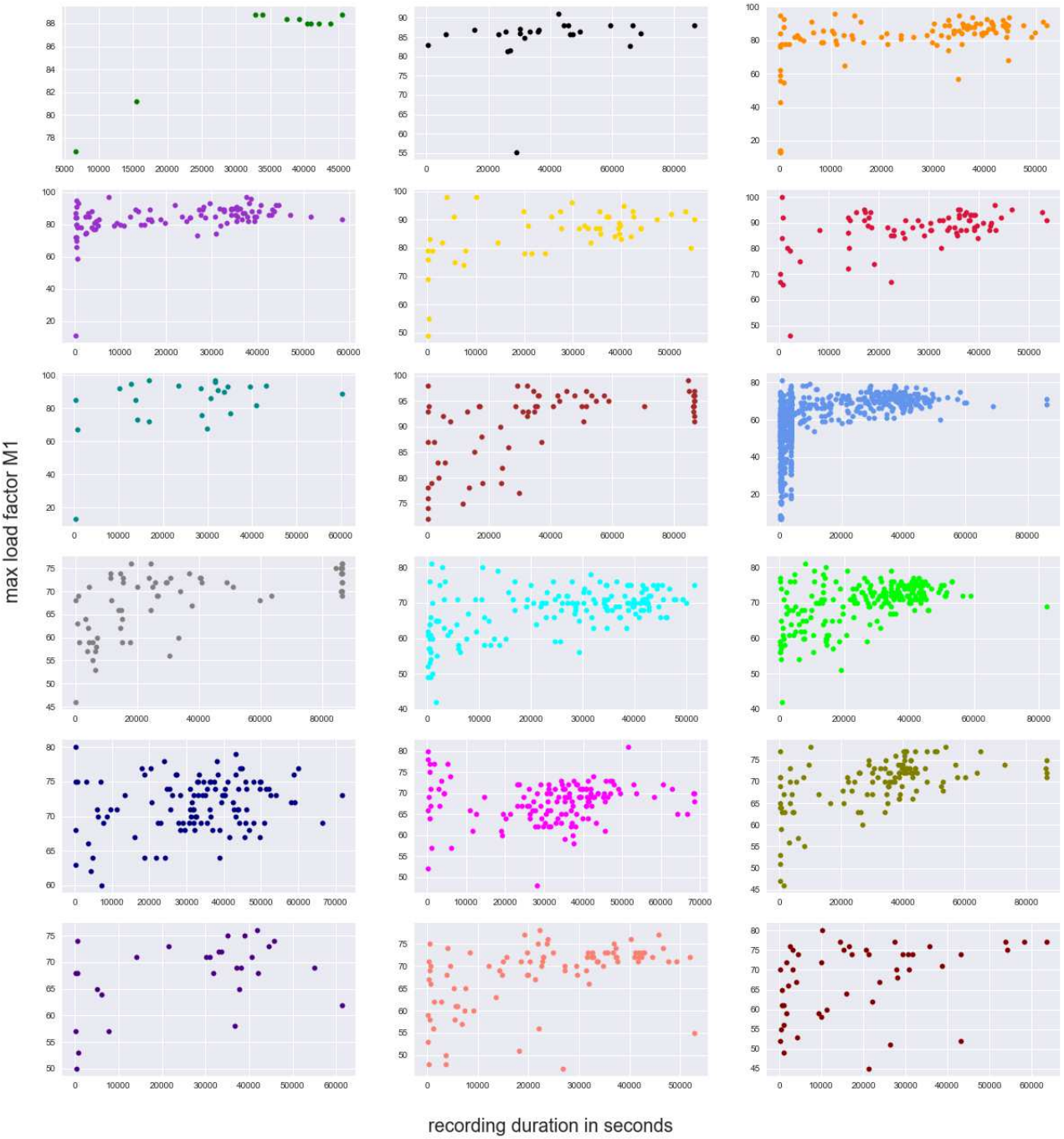


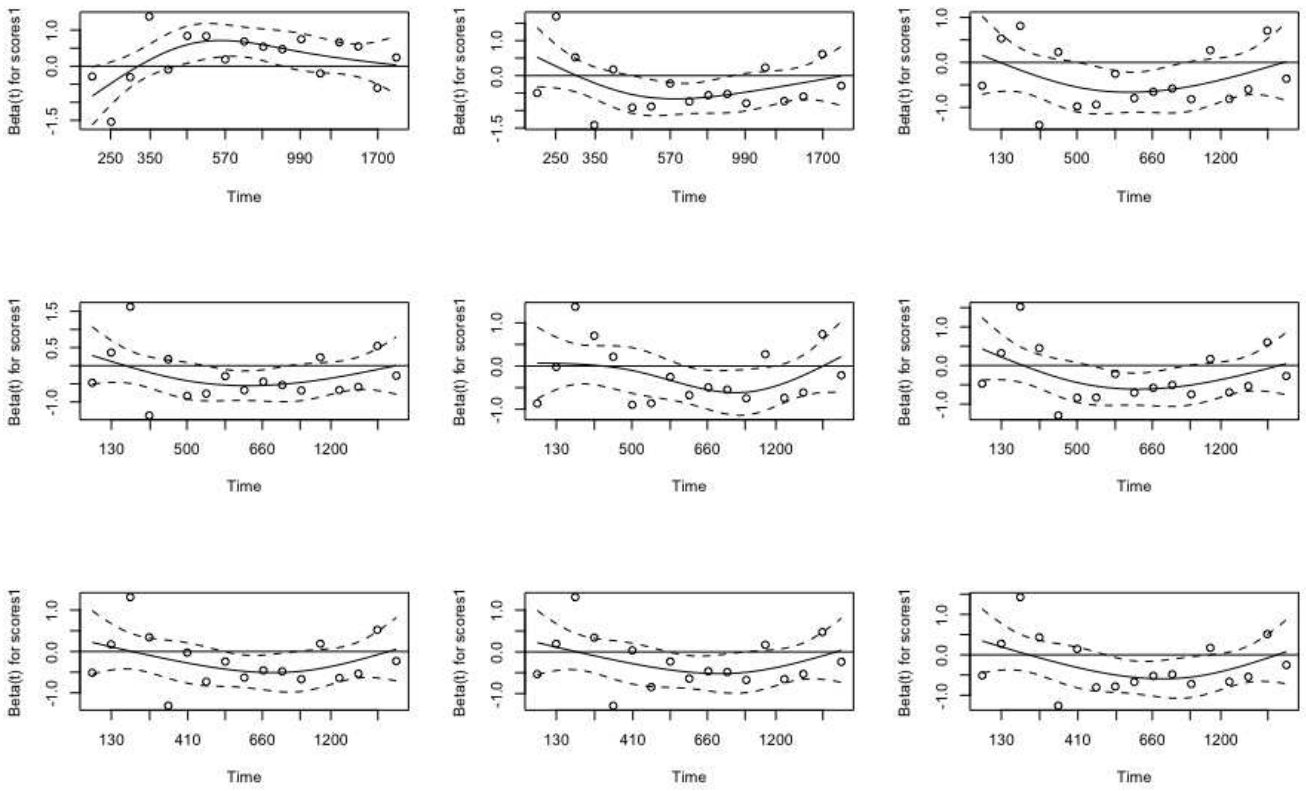
Figure 3.6: Covariate values of maximum load factor M1 over recording duration, for each machine.

Because of the discrepancy of the number of covariates (20) and the number of machines (18), Principal Component Analysis (PCA) was used to transform the original covariates to a set of uncorrelated principal components, as in the work of Wang and Zhang [WZ04]. Since the focus of this thesis is on Survival Analysis, no theory on Principal Component Analysis was presented in the Methodology chapter, however there are enough references in the literature, for example [HTF01]. The purpose of using PCA was to reduce the number of coefficients to estimate, due to the small sample. This work produced one up to five principal components, the optimal number of components will be later discussed in the validation section.

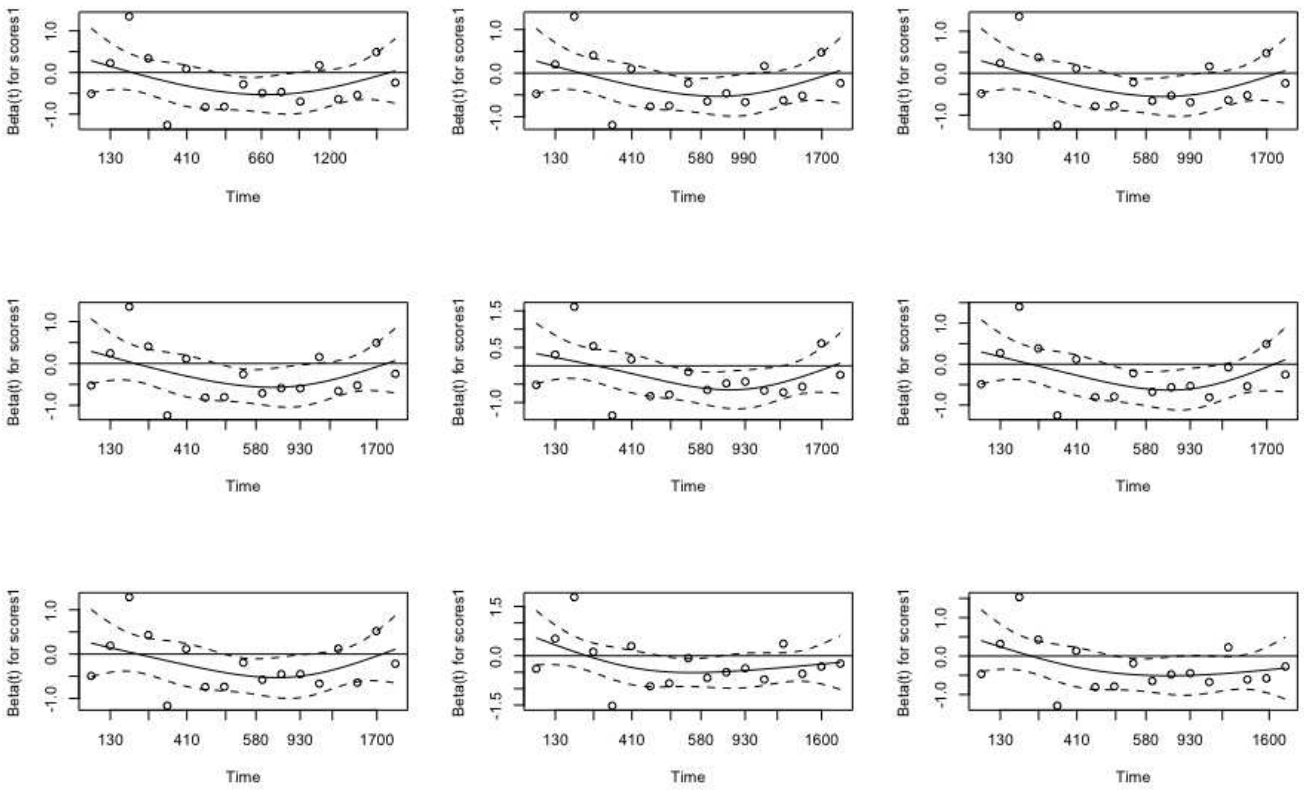
The models and the scaled Schoenfeld residuals were obtained using Therneau's survival package [The15] using R language [R C13] version 1.2. The validation was partially implemented and partially assessed by the survival package.

3.4 The Proportional Hazards Assumption

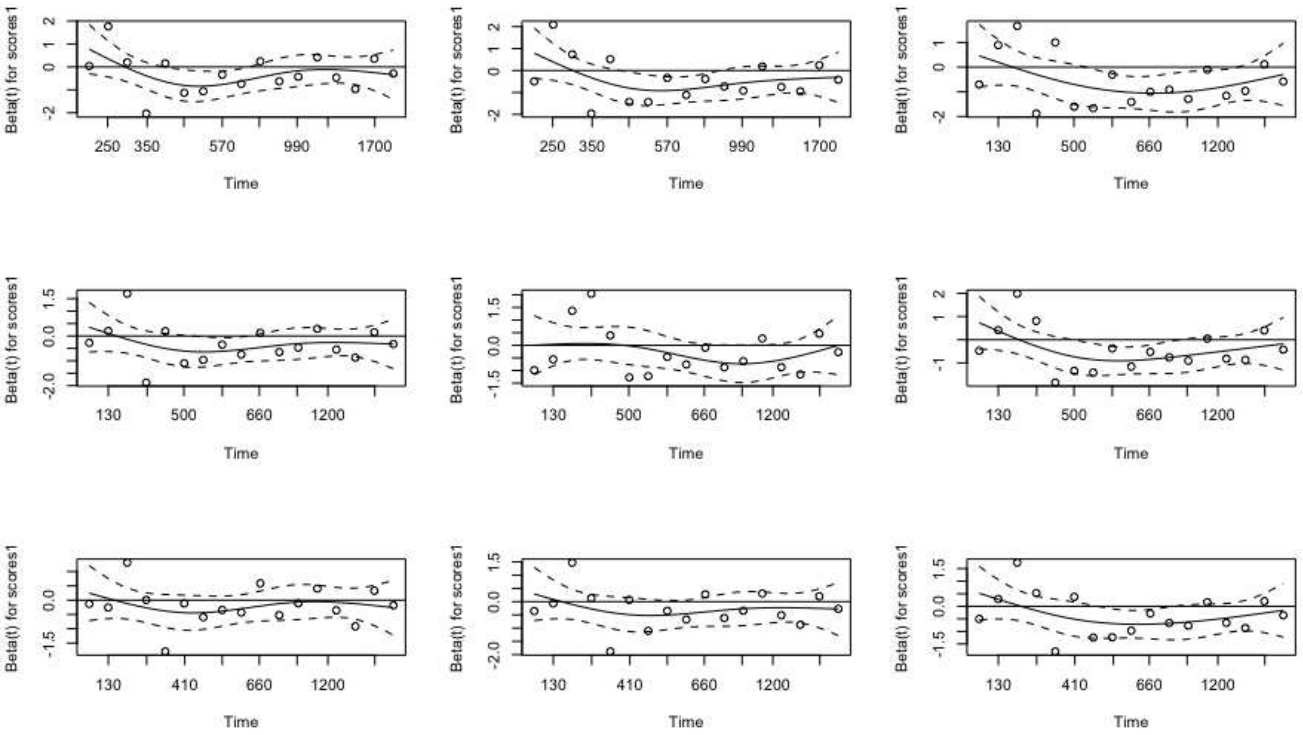
The proportionality of the components was evaluated by the "cox.zph" function [The15]. In the line of the "Schoenfeld residuals" subsection, this tool provides the correlation ψ between the scaled residuals r^* and a function of time $g(t)$ for each covariate. In this case study, used two different functions of time were used, $g(t) = t$ and $g(t) = log(t)$ [Sch92]. A hypothesis test was performed with $H_0 : \psi = 0$ and $H_1 : \psi \neq 0$. However because on the cross-validation setting many models are created, analysing statistical test for each one of them can be discouraging. Plotting the outcome of function " cox.zph" offers a visual test for the proportionality assumption which is more convenient in a cross-validation setting. $E(r_i^*) + \beta$ over time function as an estimate of the $\beta(t)$. When the PH assumption holds theses estimates should add up to zero, therefore the closest the $\beta(t)$ estimates are to the horizontal zero line more confident one can be of the PH assumption. In figures 3.7a, 3.8a, 3.9a, 3.10a are the 36 plots referring to the 36 components estimated during cross validation, using a single principal component built using 12 and 20 covariates.



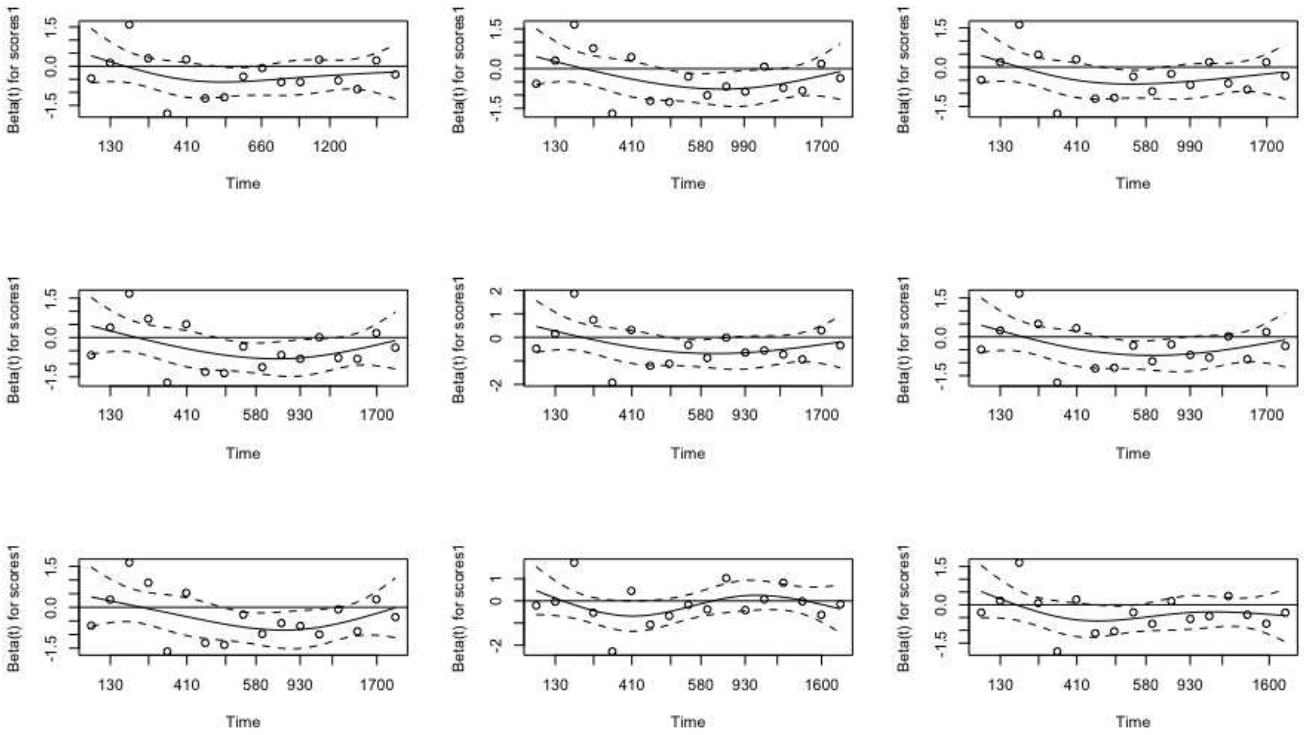
(a) 1-9 $scores1(t)$ estimates for the first principal component when using 20 covariates.



(a) 10-18 $scores1(t)$ estimates for the first principal component when using 20 covariates.



(a) 1-9 $scores1(t)$ estimates for the first principal component when using 12 covariates.



(a) 10-18 $scores1(t)$ estimates for the first principal component when using 12 covariates.

Figures 3.10a and 3.9a vary closer to the zero line when compared to figures 3.7a and 3.8a. Therefore using solely 12 covariates would be the conservative choice to preserve the proportionality assumption.

3.5 Validation

3.5.1 Discrimination

Table 3.2 contains the concordance index for all models built, M refers to models built using all 20 covariates, and M' for those using only 12 covariates. The underscripts indicates the number of principal components used to build the model, so M_n informs that the first n principal components generated from all covariates. Selecting one or two components provide higher discriminative power. Nevertheless, concordance values below 0.7 can be considered a poor or random prediction, vide Hanley & McNeil, [HM82] and Xiaona Jia [Jia18]. The low discriminative performance is due to an imprecise event definition, as mentioned, prediction of not only of failure events but of repair activities are performed, resulting in a biased risk assignment. The higher values of concordance obtained for fewer components can be explained by poor coefficient estimation as the

Model	c-index
M_1	0.620915
M_2	0.627451
M_3	0.5816993
M_4	0.5555556
M_5	0.5359477
$M_1^{'}$	0.6339869
$M_2^{'}$	0.5882353
$M_3^{'}$	0.5490196
$M_4^{'}$	0.5816993
$M_5^{'}$	0.5228758

Table 3.2: Concordance index for all models built

number of components increase, once only 17 observations were available for training.

As stated by Royston and Altman in [RA13], inadequate discrimination is a more important failing than poor calibration, since the latter can be improved by model recalibration (as seen in [vH00] and [MV09]) whereas the former cannot be altered. For this reason, and the observed concordance values in table 3.2, the number of components above two will not be considered in the following validation stages.

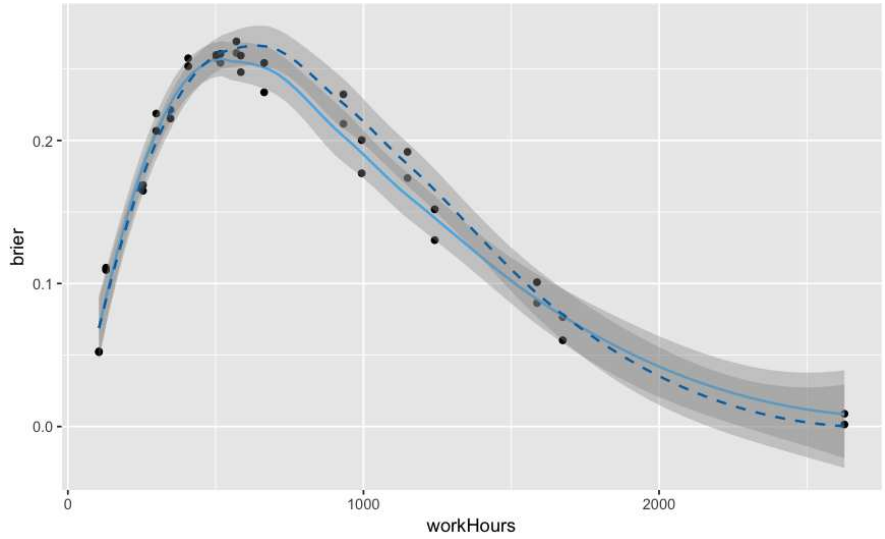
3.5.2 Overall performance

Figure 3.11 displays the Brier score represented as points over time, for the four considered models, namely, using the first and first and second principal components with 20 and 12 covariates.

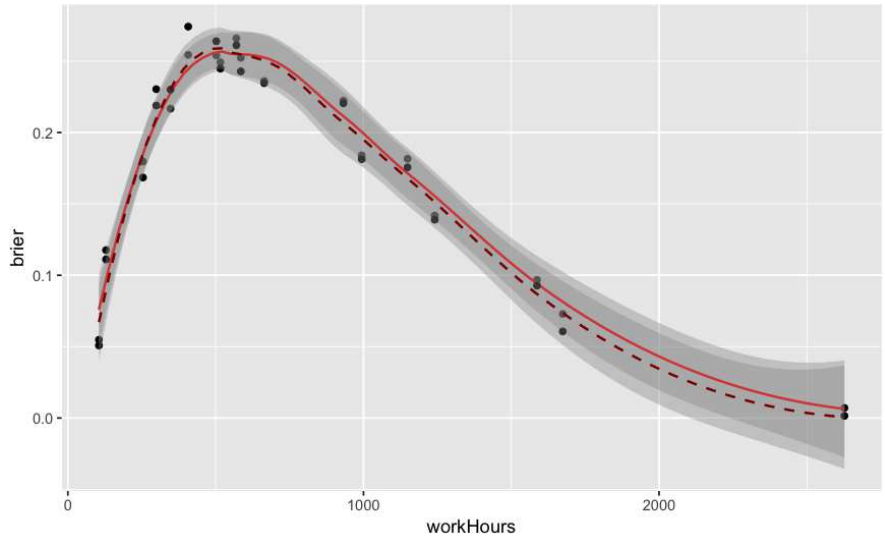
In figure 3.11a, it is not trivial to see which Brier score is in general higher, in such cases the integrated Brier score can be calculated from 0 to $t \to \infty$. Although when calculating concordance-index there is metric in the literature to distinguish overall "good" and "bad"discrimination, there is no such metric for Brier score, allowing one make such judgement comparing the Brier score of different models. Thus, if we use the fact that the time points at which the Brier score was computed for both settings of the covariate choice match; one could simply calculate, at each time point, the difference between the two Brier score curves, then sum these differences. For figure 3.11a the aggregated difference from the solid to the dashed line was -0.09435643, and in figure 3.11b 0.045659. Note that this is an simpler way of to calculate the difference between integrated Brier score when dealing with noncontinuous functions. Therefore, it is preferable to use 12 covariates (depicted as the dashed dark blue line in figure 3.11a) when using solely the first principal component for there is higher discriminative power (0.6339) and a lower brier score loss function.

When using the first and second principal components, in figure 3.11b, it is not so straight forward to decide which covariate setting to use. It can be seen that for working hours above 500, the Brier score depicted by the dashed line is placed consistently bellow

the one depicted by the solid line. Additionally, once the aggregated difference is positive 0.045659, then there is lower brier score in general, which supports the 12 covariate setting. However, there is significant better discrimination when using 20 covariates 0.627451, as opposed to 0.5882353 when using 12 covariates.



(a) Brier score computed using the first principal component. The solid light blue line represent the smoothed Brier score function over time when using 20 covariates, the dashed dark blue line is the smoothed Brier score when using 12 covariates.



(b) Brier score computed using the first and second principal components. The solid red line represents the smoothed Brier score function over time when using 20 covariates, the dashed dark red line is the smoothed Brier score when using 12 covariates.

Figure 3.11: Brier score over time using the first and the first and second principal components.

3.5.3 Overall fit

The Cox-Snell residuals can be used to assess the fit of a model based on the Cox proportional hazards model [KM03]. If the Cox-Snell residuals behave as a sample of the $Exp(1)$ distribution than the residual plot should be a straight line with 45° over the origin. Figure 3.12 contains the four residual plots for the first and first and second principal components, on 20 and 12 covariate setting. The residual plot assists the decision of which covariate setting provides the best fit.

When measuring overall performance through the calculation of the Brier Score, it was not "straight forward to decide which covariate setting to use". However, when looking at figures 3.12a and , it is possible to see that, regardless of the selected components, using 12 covariates instead of 20 provides a better model fit. It remains to decide if the a superior model fit compensate a poorer discrimination, using 12 covariates. Since discrimination is essential for the correct functioning of a prediction model, an inferior fit could be taken.

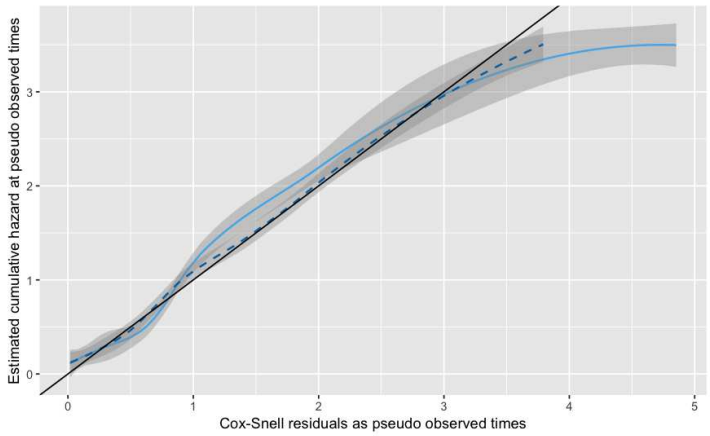
3.6 Conclusion

As mentioned in the Overall performance section 3.5.2, the preferred model uses a single principal component and only 12 covariates, since it produced the highest concordance (0.6339) and a lower brier score loss function when compared to the model with 20 covariates. The first principal component was obtained using "prcomp"from "stats"package [R C13]. The scores for the 18 observations, respective to the first component. The first component consists of proportion of the variance and standard deviation, respectively, 0.2617 and 1.7722.

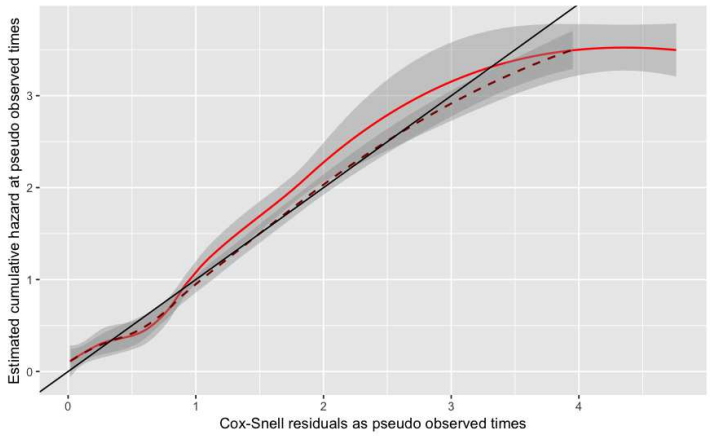
Follow in figure 3.3 the used covariates at cutoff.

Listing 3.1: Principal component scores

```
          1           2           3           4
-4.29897515  0.25181522  0.80136126 -0.88510211
          5           6           7           8
-2.23249731 -1.77954089 -1.44045456 -0.10832180
          9          10          11          12
 1.35710034  0.87029766 -0.17686191  3.07935805
         13          14          15          16
 0.01377039  1.82538322 -0.20065543 -0.08098909
         17          18
 0.24084790  2.76346420
```



(a) Cox-Snell residual plot using the first principal component. The solid light blue line represent the cumulative hazard of the residuals over the residuals when using 20 covariates, the dashed dark blue line when using 12 covariates.



(b) Cox-Snell residual plot using the first principal component. The solid red line represent the cumulative hazard of the residuals over the residuals when using 20 covariates, the dashed dark red line when using 12 covariates.

Figure 3.12: Nelson–Aalen estimator of the cumulative hazard rate of the Cox-Snell residuals over the Cox-Snell residuals, using the first and the first and second principal components.

Covariate	avgavg mean	maxmax mean	avgavg stdev	maxmax stdev
motor rotation speed [1/min]	980.3705	2109.1944	192.111364	20.892680
load factor M1	25.8027	85.1333	13.4734	8.8856
coolant temperature[°C]	63.1066	95.9722	9.3115	5.4136
fuel rate[L/h]	6.5713	37.8027	1.5571	2.4987
	mean	stdev		
distance[km]	17.6674	4.163557		
fuel consumption[L]	982.0037	248.9501		
engine idle[h]	13.4841	10.3206		
construction year	2015.6111	1.2432		

Table 3.3: Covariates at cutoff used for the final model

The final forumula used was $Surv(workHours, event) \sim scores1$ with the single principal component $scores1$, where $workHours$ is the outcome variable and $events$ is a vector indicating if the events were censored or not, in this study all 18 observations failed, so $events$ was a vector of 1s. Follow the output summary of the cox model computed with the "survival" package[The15].

As seen in listing 3.2, for each unit increase in the $scores1$ covariate a corresponding

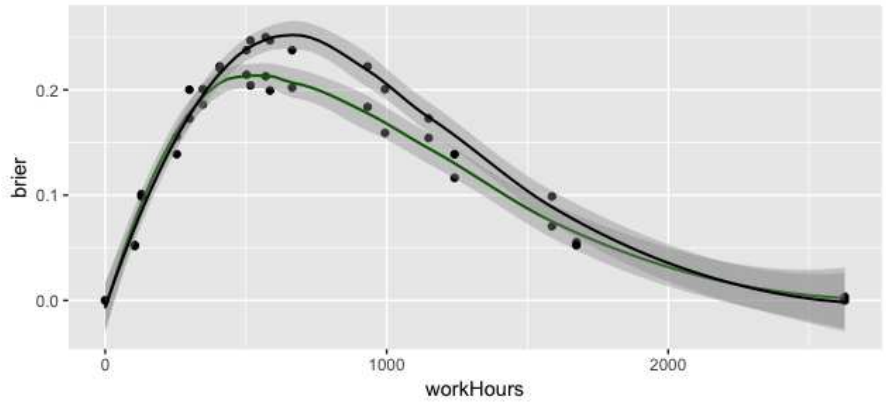
Listing 3.2: R coxph output

```
n= 18, number of events= 18

          coef exp(coef) se(coef)      z Pr(>|z|)
comp1   -0.3521    0.7032   0.1714 -2.054     0.04
---
        exp(coef) exp(-coef) lower .95 upper .95
comp1      0.7032      1.422    0.5026     0.984

Concordance= 0.627  (se = 0.084 )
Likelihood ratio test= 4.21  on 1 df,   p=0.04
Wald test            = 4.22  on 1 df,   p=0.04
Score (logrank) test = 4.28  on 1 df,   p=0.04
```

0.7032 decrease in the machine's risk. The Z-statistics informs that the coeficient is only 2.054 standard errors away to the left of 0, the p-value is also high 0.04 if we are to consider a 5% confidence level. The c-index provided 0.627 is around what was expected, from the cross validation results see table 3.2. In figure , there is the Brier score for the proposed final Cox model, as a reference the Brier score of the Kaplan-Meier prediction model was added to measure the improvement to the loss function by adding covariates. The obtained integrated Brier score of the reference and the Cox model is respectively, 0.127 and 0.114.



(a) Brier score computed using the first principal component. The dark green line represent the smoothed Brier score function over time when using 12 covariates, the black line corresponds to the reference Brier score without covariates.

Although suboptimal, only internal validation was performed. External validation would require an additional external sample of machines from the same model and manufacturer which also underwent replacement of "Bunkergummi". Nevertheless under scarce data, it was not possible to separate such sample from the training and internal validation. Therefore, no claims of generalizability are made for this final model [SV14].

CHAPTER 4

Further study

As a next step, a model-based simply on Instandsetzung repairs could be built on the same framework, to observe if there will be an increase in the performance measures, as suggested in 3. The previous study focus was primarily on the semi-parametric Cox models. In this model, no assumption about the form of the baseline hazard need to be specified. Although it is convenient to abstract from the shape of the distribution, when one has good reasons to make such an assumption and the distribution provides a good fit to the data, they tend to give more precise estimates of the quantities of interest because these estimates are based on fewer parameters [KM03]. The search for reducing the dimensions of the problem was evident by using PCA, perhaps a Weibull model could be built to compare the results with the Cox model.

One of the main issues was the scarce sample of data. Because prediction models, in the line of this current study, consider a single event, this issue could be addressed by using successive events as different events or perhaps recurrent cox models. If the repair description indicates that a full replacement was made, i.e. the old part was fully substituted by a new one, then there is no apparent problem in considering the consecutive lifecycle as a machine. However, implementing such correction may face some problems, for instance of being able to aggregate repair descriptions referring to the same repair activity, in not always trivial and may require the use of natural language processing or expert support. Additionally, the repairs would need to be after failure, i.e. an Instandsetzung repair.

Another problem on the data source, consists of its ambiguous nature, after breakdown multiple repairs are made, all classified as Instandsetzung repairs. It is possible that a mechanic repaired one component which was completely unrelated to the cause of failure, and still present in the data as a repair entry. Pinpointing the repair which fixed the failure will be another challenge. Therefore a better communication between experts and researchers have to be established.

List of Figures

List of Tables

Bibliography

[AS17] Peter C. Austin and Ewout W. Steyerberg. Events per variable (epv) and the relative performance of different strategies for estimating the out-of-sample validity of logistic regression models. 2017.

[BMD+10] Carine A. Bellera, Gaetan MacGrogan, Marc Debled, Christine Tunon de Lara, Veronique Brouste, and Simone Mathoulin-Pelissier. Variables with time-varying effects and the cox model: Some statistical concepts illustrated with a prognostic factor study in breast cancer. *British Medical Journal*, 2010.

[CDCD] Chris Coleman, Satish Damodaran, Mahesh Chandramouli, and Ed Deuel. Making maintenance smarter. predictive maintenance and the digital supply network.

[CL00] Nicholas A. Christakis and Elizabeth B. Lamont. Extent and determinants of error in doctors' prognoses in terminally ill patients: prospective cohort study. *British Medical Journal*, 320:469-73, 2000.

[CO84] D. R Cox and D. Oakes. *Analysis of Survival Data*. CRC Press, first edition, 1984.

[Cox72] D. R. Cox. Regression models and life-tables. *Journal of the Royal Statistical Society*, 1972.

[CS68] D. R. Cox and E. J. Snell. A general definition of residuals. *Journal of the Royal Statistical Society*, 1968.

[Des] Simona Despa. Statnews #78. what is survival analysis? `https://www.cscu.cornell.edu/news/statnews/stnews78.pdf`.

[ELSD16] Lucas Equeter, Christophe Letot, Roger Serra, and Pierre Dehombreux. Estimate of cutting tool lifespan through cox proportional hazards model. *IFAC: International Federation of Automatic Control*, 2016.

[GCS08] Thomas A. Gerds, Tianxi Cai, and Martin Schumacher. The performance of risk prediction models. *Biometrical Journal*, 2008.

[GL17] Tom Greene and Liang Li. From static to dynamic risk prediction: Time is everything. *American Journal of Kidney Diseases*, 69(4):492-494, 2017.

[GT94] Patricia M. Grambsch and Terry M. Therneau. Proportional hazards tests and diagnostics based on weighted residuals. *Biometrika*, 81:515- 526, 1994.

[HCP+82] Frank E. Harrell, Robert M. Califf, David B. Pryor, MD, Kerry L. Lee, and Robert A. Rosati Rosati. Evaluating the yield of medical tests. *Journal of the American Medical Association*, 1982.

[HJS01] Robin Henderson, Margaret Jones, and Janez Stare. Accuracy of point predictions in survival analysis. *Statistics in Medicine*, 20:3083-3096, 2001.

[HK05] R. Henderson and N. Keiding. Individual survival time prediction using statistical models. *Journal of Medical Ethics*, 31:703-706, 2005.

[HLS08] David W. Hosmer, Stanley Lemeshow, and May Susanne. *Applied Survival Analysis. Regression Modeling of Time-to-Event Data*. John Wiley & Sons, second edition, 2008.

[HM82] James A. Hanley and Barabara J. McNeil. The meaning and use of the area under a receiver operating characteristic (roc) curve. *Radiology*, 1982.

[HMRM] Jonathan Holdowsky, Monika Mahto, Michael E. Raynor, and Cotteleer Mark. Inside the internet of things (iot). a primer on the technologies building the iot.

[HTF01] Trevor Hastie, Robert Tibshirani, and Jerome Friedman. *The Elements of Statistical Learning*. Springer Series in Statistics. Springer New York Inc., New York, NY, USA, 2001.

[IKS+91] Kenji Ikeda, Hiromitsu Kumada, Satoshi Saitoh, Yasuji Arase, and Chayama Kazuaki. Effect of repeated transcatheter arterial embolization on the survival time in patients with hepatocellular carcinoma. an analysis by the cox proportional hazard model. 1991.

[Jia18] Xiaona Jia. A cox-based risk prediction model for early detection of cardiovascular disiese. Master's thesis, AUCKLAND UNIVERSITY OF TECHNOLOGY, New Zealand, 2018.

[KBS19] Hårvard Kvamme, Ørnulf Borgan Borgan, and Ida Scheel. Time-to-event prediction with neural networks and cox regression. *Journal of Machine Learning Research*, 2019.

[KK12] David G. Kleinbaum and Mitchel Klein. *Survival Analysis, A Self-Learning Text*. Springer, third edition, 2012.

[KM03] J.P. Klein and M.L. Moeschberger. *Survival analysis: Techniques for censored and truncated data*. Springer, second edition, 2003.

[KR08] Louis-Philippe Kronek and Anupama Reddy. Logical analysis of survival data: prognostic survival models by detecting high-degree interactions in right-censored data. *Bioinformatics*, 2008.

[KW97] Dhananjay Kumar and Ulf Westberg. Maintenance scheduling under age replacement policy using proportional hazards model and ttt-plotting. *European Journal of Operational Research*, 1997.

[LEA97] Kwan-Moon Leung, Robert M. Elashoff, and Abdelmonem A. Afifi. Censoring issues in survival analysis. *Annu. Rev. Public Health*, 18:83–104, 1997.

[LG91] Charles E. Love and R. Guo. Using proportional hazard modelling in plant maintenance. *QUALITY AND RELIABILITY ENGINEERING INTERNATIONAL*, 1991.

[LWZ+14] Jay Lee, Fangji Wu, Wenyu Zhao, Masoud Ghaffari, Linxia Liao, and David Siegel. Prognostics and health management design for rotary machinery systems — reviews, methodology and applications. *Mechanical Systems and Signal Processing*, 2014.

[MJ92] Viliam Makis and Andrew K.S. Jardine. Optimal replacement in the proportional hazards model. *INFOR: Information Systems and Operational Research*, 1992.

[MKRM99] Richards Marcus, Diana Kuh, Hardy Rebecca, and Wadsworth Michael. Lifetime cognitive function and timing of the natural menopause. *American Academy of Neurology*, 1999‘.

[MMBS14] Mittal, Madigan, Burd, and Suchard. High-dimensional, massive sample-size cox proportional hazards regression for survival analysis. *Biostatistics*, 2014.

[MN04] Jay I. Myung and Daniel J. Navarro. Information matrix. *Encyclopedia of Behavioral Statistics. Wiley.*, 2004.

[MV09] Karel G. M. Moons and Yvonne Vergouwe. Prognosis and prognostic research: application and impact of prognostic models in clinical practice. *British Medical Journal*, 2009.

[R C13] R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2013. ISBN 3-900051-07-0.

[RA13] Patrick Royston and Douglas G. Altman. External validation of a cox prognostic model: principles and methods. *BioMed Central Medical Research Methodology*, 2013.

[Rod07] Germán Rodríguez. Generalized linear models lecture notes, 2007.

[Sch92] Michael Schemper. Cox analysis of survival data with non-proportional hazard functions. *Royal Statistical Society*, 1992.

[SN15] Hisham M. Said and Tony Nicoletti. Telematics data-driven prognostics system for construction heavy equipment health monitoring and assessment. Construction Specialty Conference, 2015.

[SNPH15] Hisham M. Said, Tony Nicoletti, and Peter Perez-Hernandez. Utilizing telematics data to support effective equipment fleet-management decisions: Utilization rate and hazard functions. *Journal of Computing in Civil Engineering*, 2015.

[sur97] Survival analysis in public health research. *Annu. Rev. Public Health.*, 1997.

[SV14] Ewout W. Steyerberg and Yvonne Vergouwe. Towards better clinical prediction models: seven steps for development and an abcd for validation. *European Heart Journal*, 2014.

[SVC+09] Steyerberg, Vickers, Cook, Gerds, Gonen, Obuchowski, and Kattan Pencina. Assessing the performance of prediction models. a framework for traditional and novel measures. *Lippincott Williams & Wilkins*, 2009.

[SWHZ] Xiao-Sheng Si, Wenbin Wang, Chang-Hua Hu, and Dong-Hua Zhou. Remaining useful life estimation – a review on the statistical data driven approaches. *European Journal of Operational Research*.

[TCA19] Terry Therneau, Cynthia Crowson, and Elizabeth Atkinson. *Using Time Dependent Covariates and Time Dependent Coefficients in the Cox Model*, March 2019.

[TG00] Terry M. Therneau and Patricia Grambsch. *Modeling Survival Data. Extending the Cox Model*. Springer, first edition, 2000.

[The15] Terry M Therneau. *A Package for Survival Analysis in S*, 2015. version 2.38.

[top] Despite today's turbulent markets, it is possible to reach top quartile performance. `https://www.emerson.com/en-us/about-us/featured-stories/top-quartile-performance`, note = Emerson, featured stories.

[TTT+10] Yasuko Takeda, Yutaka Takeda, Shigehiro Tomimoto, Tomomitsu Tani, Hitomi Narita, and Genjiro Kimura. Bilirubin as a prognostic marker in patients with pulmonary arterial hypertension. *British Medical Journal*, 2010.

[vH00] H. C. van Houwelingen. Validation, calibration, revision and combination of prognostic survival models. *Stat Med*, 2000.

[WHSS11] M. Wang, Y. He, L. Shi, and C. Shi. Multivariate analysis by cox proportional hazard model on prognosis of patient with epithelial ovarian cancer. *European Journal of Gynaecological Oncology*, 2011.

[Wol] Gary Wollenhaupt. Iot slashes downtime with predictive maintenance. `https://www.ptc.com/en/product-lifecycle-report/iot-slashes-downtime-with-predictive-maintenance`.

[WZ04] Wenbin Wang and Wenjuan Zhang. A model to predict the residual life of aircraft engines based upon oil analysis data. *Wiley Periodicals, Inc.*, 2004.

[Yua08] Yan Yuan. *Prediction Performance of Survival Models*. PhD thesis, University of Waterloo, 2008.