Linköping studies in science and technology
Licentiate Thesis. No. 1779

Data-driven lead-acid battery lifetime prognostics

Sergii Voronov

Department of Electrical Engineering
Linköping University, SE-581 33 Linköping, Sweden

Linköping 2017

Linköping studies in science and technology
Licentiate Thesis. No. 1779

This is a Swedish Licentiate's Thesis.
Swedish postgraduate education leads to a Doctor's degree and/or a Licentiate's degree.
A Doctor's degree comprises 240 ECTS credits (4 years of full-time studies).
A Licentiate's degree comprises 120 ECTS credits,
of which at least 60 ECTS credits constitute a Licentiate's thesis.

Sergii Voronov
sergii.voronov@liu.se
www.vehicular.isy.liu.se
Division of Vehicular Systems
Department of Electrical Engineering
Linköping University
SE-581 33 Linköping, Sweden

Copyright © 2017 Sergii Voronov.
All rights reserved.

Voronov, Sergii
Data-driven lead-acid battery lifetime prognostics
ISBN 978-91-7685-504-1
ISSN 0280-7971

Typeset with LaTeX 2ε
Printed by LiU-Tryck, Linköping, Sweden 2017

To my parents

Abstract

To efficiently transport goods by heavy-duty trucks, it is important that vehicles have a high degree of availability and in particular avoid becoming standing by the road unable to continue the transport mission. An unplanned stop by the road does not only cost due to the delay in delivery, but can also lead to a damaged cargo. High availability can be achieved by changing components frequently, but such an approach is expensive both due to the frequent visits to a workshop and also due to the component cost. Therefore, failure prognostics and flexible maintenance has significant potential in the automotive field for both manufacturers, commercial fleet owners, and private customers.

In heavy-duty trucks, one cause of unplanned stops are failures in the electrical power system, and in particular the lead-acid starter battery. The main purpose of the battery is to power the starter motor to get the diesel engine running, but it is also used to, for example, power auxiliary units such as cabin heating and kitchen equipment. Detailed physical models of battery degradation is inherently difficult and requires, in addition to battery health sensing which is not available in the given study, detailed knowledge of battery chemistry and how degradation depends on the vehicle and battery usage profiles.

The main aim of the given work is to predict the lifetime of lead-acid batteries using data-driven approaches. Main contributions in the thesis are: a) the choice of the Random Survival Forest method as the model for predicting a conditional reliability function which is used as the estimator of the battery lifetime, b) variable selection for better predictability of the model and c) variance estimation for the Random Survival Forest method.

When developing a data-driven prognostic model and the number of available variables is large, variable selection is an important task, since including non-informative variables in the model have a negative impact on prognosis performance. Two features of the dataset has been identified, 1) there are few informative variables, and 2) highly correlated variables in the dataset. The main contribution is a novel method for identifying important variables, taking these two properties into account, using Random Survival Forests to estimate prognostics models. The result of the proposed method is compared to existing variable selection methods, and applied to a real-world automotive dataset.

Confidence bands are introduced to the RSF model giving an opportunity for an engineer to observe the confidence of the model prediction. Some aspects of the confidence bands are considered: a) their asymptotic behavior and b) usefulness in the model selection. A problem of including time related variables is addressed in the thesis with arguments why it is a good choice not to add them into the model. Metrics for performance evaluation are suggested which show that the model can be used to find and optimize cost of the battery replacement.

Acknowledgments

First of all, I would like to thank Scania CV and Vehicular Systems group at Linköping University for giving me an opportunity to participate in the project. All work and results would be impossible without a chance that I have received.

I would like to express my gratitude to my supervisor Dr. Erik Frisk and co-supervisor Dr. Mattias Krysander for all the help and advises they gave me, and also for their patience during these years. I appreciate the effort and time they spent to make me a better researcher in the areas of prognostics and machine learning. I would also like to thank my supervisor at Scania CV Dr. Jonas Biteus for the great disscussions and insights into the problem under study and indispensable help with data extraction and analysis. I am grateful to all colleagues from Vehicular Systems group for a nice and pleasant working environment.

The very special acknowledgement goes to my family and, particularly, to my parents. Thank you for always being with me and believing in me, even at times when the hope is lost. Only you can cheer me up in situations when I do not listen to anyone.

This all acknowledgement section would be impossible without mentioning my friends. Thank you all a lot for the great memories and fun time we have had together. No matter where you are right now, in Ukraine or Sweden, in Portugal or China, or in USA, you are in my heart and I always remember how lucky I am for having you in my life.

Linköping, May 2017
Sergii Voronov

Contents

Chapter 1

Introduction

Intelligent, condition based, maintenance is an effective and cost-effective tool to improve availability and uptime in many industrial applications. High availability can be achieved by changing components frequently, but such an approach is expensive both due to the frequent visits to a workshop and also due to the component cost. Therefore, failure prognostics and flexible maintenance has significant potential.

Areas of significant current interest are autonomous systems and autonomous vehicles where the main aim are systems that can perform required tasks without any supervision from humans. Self-driving vehicles should deliver goods to the customers and help at work at the construction sites. Autonomous operation of dynamic systems with high uptime requirements and no direct user feedback is a challenging task and flexible, condition based, maintenance is then even more important.

Also in more established applications, like heavy-duty trucks that is the main motivating business for this thesis work, it is important that vehicles have a high degree of availability and in particular avoid standing by the road unable to continue the transport mission. An unplanned stop by the road or at the construction site does not only cost due to the delay in delivery, but can also lead to damaged cargo or influence the work plan of other self-driving vehicles. Therefore, maintenance planning becomes important in the automotive industry, in particular where car or truck manufactures do not only produce and deliver cars and trucks, but also provide maintenance services that will allow fleet owners or regular customers to avoid unexpected failures. In heavy-duty trucks, one cause of unplanned stops are failures in the electrical power system, and in particular, the lead-acid starter battery. Flexible maintenance and fault prognosis of lead-acid batteries is the topic of this thesis. It is an industrially relevant component and prognostics is technically challenging which motivates this research.

1.1 Lead-acid batteries

It can be a surprising fact, but first batteries could exist as early as thousands years ago. One such battery was found in 1936 near Baghdad. However, some scientists believe it was not used as the source of energy, but the object has all elements that make it possible to consider the discovered object as a battery. The battery was in the form of a clay jar and had iron rod as a positive terminal, copper cylinder as a negative terminal and probably a vinegar solution as an electrolyte.

Considering modern batteries, there are a few different types of batteries available, for example lead-acid, nickel-based, lithium-ion, sodium-sulfur etc. This thesis concerns lead-acid batteries that are used in heavy-duty trucks. The main purpose of the battery is to power the starter motor to get the diesel engine running, but it is also used to, for example, power auxiliary units such as cabin heating and kitchen equipment. Figure 1.1 shows one example of a lead-acid starter battery used in Scania heavy-duty trucks.

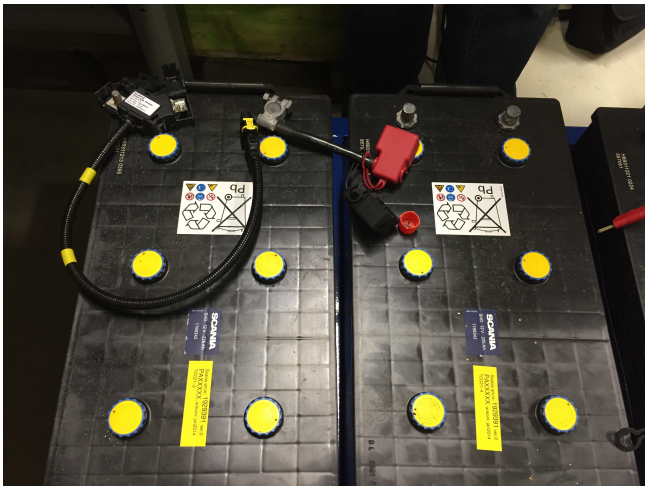
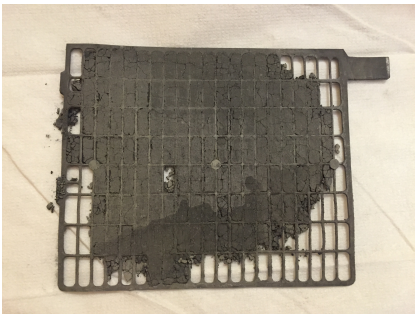


Figure 1.1: Lead-acid starter battery used at Scania CV.

This thesis concerns tracking battery degradation and main reasons for the lead-acid battery degradation are sulfation, corrosion, and internal short. Sulfation happens when the battery is not charged to its full charge value. This is common for the trucks operated within cities as they start the battery quite often and do not charge it fully because of the frequent stops and short travelling distance within the cities. Corrosion appears due to high charge voltage which can happen if a vehicle is operated in a broad temperature range and charging procedure is not designed properly. Internal short is a type of degradation that happens due to the fact that with time part of the lead from the plates accumulates at the bottom of the battery container which can lead to

(a) healthy positive plate



(b) degraded positive plate

Figure 1.2: Healthy, left figure, and degraded, right figure, positive plates in lead acid batteries.

a conducting layer formation that connects two plates. Figure 1.2 demonstrates a healthy and a degraded plate from a lead-acid battery.

1.2 Lifetime prognostics

This section gives a brief introduction to the research methods of lifetime prognostics area highlighting main differences between them and making a bridge to the further discussion in this thesis.

There are, coarsely, two categories, (Roemer et al., 2005), of approaches to lifetime prognostics, namely, model based and data-driven methods. Cornerstones of the model based methods are physical laws and equations that describe degradation of the components. However, accurate predictions of the model based methods rely on the accurate degradation models. It is sometimes, and this is certainly true for lead-acid batteries, hard to develop an accurate degradation model for a particular system, and then data-driven methods can be an alternative if reliability data is available. Data-driven models work with data which is logged during the operation of the systems under study.

A basic notion in lifetime prognostics is Remaining Useful Life (RUL), which is either the remaining time until component failure or to the point where it can no longer fulfill its function. It is common in both approaches to aim for a reliable estimate of the remaining useful life. In general, RUL is estimated using sensors that give health related information of the component, meaning, there is a possibility to track the state of health related parameters during the lifetime of the component. Examples of model-based prognostics are given in (Daigle and Goebel, 2011; Hanachi et al., 2015; Saha and Goebel, 2009). Authors in (Daigle and Goebel, 2011) developed a detailed physics-based model of a pneumatic valve in a cryogenic refueling system and predict the RUL of the component based on a discrete sequence of observations and a particle filter

as a predictive technique. Prognostics and health management of gas turbine engines is addressed in (Hanachi et al., 2015) where a comprehensive nonlinear thermodynamic model is developed. In (Saha and Goebel, 2009) authors model li-ion battery capacity depletion in a particle filtering framework similar to (Daigle and Goebel, 2011). Here, an empirical model that describes battery depletion during a discharge cycles is suggested, then, it is used in a particle filter to predict the RUL. It is worth to notice that the mentioned works all have the possibility to either measure key parameters of the models during the operational time or at least during the period of tuning the model. However, and this is a key observation for the battery problem studied in this thesis, this is not the case for the data under study.

Data-driven models use machine learning methods to either estimate RUL, the health of the component, or prognostic related information. These data-driven methods can be categorized into parametric and non-parametric methods. A parametric approach assumes a parametric form of the distribution of the RUL and the parameters of interest are estimated based on observations. An approach based on Hidden Markov Models is proposed by the authors in (Medjaher et al., 2012) as machine learning method to predict the degradation of the bearings. There, it is assumed that the observation probability densities are a mixture of Guassians and that there are signals from the sensors that can capture degradation of the bearings. Thus, data from the sensors can be used to estimate parameters of the distributions and as the result the RUL. A parametric method COSMO is proposed in (Fan et al., 2015) to detect and predict faults in air pressure compressors of a fleet of city busses. The signals from the busses are collected continuously and compared to measurements from the failed busses aggregated in the distribution. In turn, non-parametric data-driven methods use machine learning methods that do not have any basic assumption regarding the underlying distribution of the RUL (Ishwaran et al., 2008; Prytz et al., 2015). In (Ishwaran et al., 2008) the model is the ensemble of tree-based classifiers, Random Survival Forest, which are built with the help of a function that splits the sample into two in a such way that the difference between the two samples is as large as possible. For the battery case, this corresponds to separating the set of vehicles such that vehicles in the same sample has similar battery degradation properties. Another non-linear ensemble method, called Random Forest, is used in a data-driven approach proposed by Prytz et al. (2015) which combines pattern recognition with RUL estimation. The type of data set, which is available for the study in (Prytz et al., 2015), is similar to the data set used in this thesis. Another example of a non-parametric data-driven model is an artificial neural network (Cheng and Titerington, 1994). Artificial neural networks do not need any assumptions about the underlying degradation distribution and can learn degradation profiles from data. Nowadays, hybrid methods, a fusion of model-based and data-driven, are proposed. For example, in (Zhao et al., 2015) an integrated prognostic method is proposed that uses Paris' law as a degradation model of a gear and Bayesian inference as the tool to address the changes in the

physical model depending on operating conditions.

1.3 Vehicle fleet data

For a data-driven approach, the available data is of key importance. The data used throughout the thesis is introduced in the current section and its distinctive characteristics are described. This thesis is part of a project together with Scania CV, our industrial partner, and access to the data has been generously given by them.

The data set used here is different from the data that is mentioned in Section 1.2. First, the data under study is static, i.e., there is only one set of measurements for each vehicle. This means that it is not possible to track battery lifetime related measurements during a vehicle's lifetime. This fact makes it difficult to apply a model based method, because they mostly rely on time series sensor measurements of health related parameters of the battery. Another distinctive feature of the data set is that the true underlying degradation profiles are not known. This means that, in addition to the lack of time series for a vehicle, it is not possible to compare a prediction of a model with a true degradation curve.

The data set contains stationary and variable information. Stationary data is given in the form of discrete variables and represent specification of a vehicle, for example engine model, battery position, if there is a kitchen equipment and so on. As an example, a battery position variable takes one of three possible values: right, left mounting point and rear frame position. Variable data is represented in the form of histograms which show how the vehicle has been operated during the period up to the visit to the workshop when the data is collected. For example, there is a battery voltage histogram which has 10 bins. Each bin shows what fraction of time out of the total operational time a vehicle has been operated in some voltage range. Every bin of the histogram is treated as a separate variable and therefore the voltage histogram contributes with 10 variables to the data set.

It should be noted that measurements directly related to battery health are not available. For the case of the voltage histogram, the logged voltage is the battery voltage right before the vehicle is switched on. The measured value of the voltage is not the open circuit voltage, because the battery requires hours of relaxation time to get to the open circuit voltage.

Two data sets which represent two versions of similar information are provided by Scania CV. The data sets have information from vehicles operating in 5 European markets. For example, the older data set contains 33,603 vehicles each with 284 variables, while the newer data set has data from 56,163 vehicles with 536 variables. Battery failures are observed for only a small fraction of the vehicles. A vehicle in the database that has a functioning battery is called censored, since the future failure time of the battery is unknown. The censoring rate for the two data sets is as high as 80 and 90 percent. Therefore, the vast

majority of the vehicles do not experience battery problems throughout time of the study and the proposed method should handle censored data in a systematic way. Some characteristics of the data sets are summarized below:

- 56,163/33,603 vehicles from 5 EU markets
- 536/284 variables stored in each vehicle snapshot
- One single snapshot per vehicle
- Heterogeneous data, i.e., a mixture of categorical and numerical data
- Histogram variables
- Censoring rate about 80/90 percent
- Significant missing data rate about 40 percent

The data sets that are being used merge information from three different databases: a) one contains static information/specification of the vehicles, b) another has variable information/histogram variables and c) the third database contains information about workshop visits. Information whether the vehicles had problem with a battery or not is extracted from the workshop information after some data preprocessing. After this the three databases are merged into a new one based on a unique vehicle identifier used in all databases.

1.4 Scope and aim

The main aim of the work is to construct a data-driven prognostic framework applicable to lifetime prognostics not only for the lead-acid batteries in heavy-duty trucks, but also to the other components of the vehicle in the future when the data set is similar to the one considered in the current work. Taking into account the characteristics of the data set described in Section 1.3, a conditional probability function is selected as the target for the prognostic framework instead of the RUL. The conditional probability function being estimated in the given work is the probability for a random variable T, the time of a battery failure, be larger than $t + t_0$ time units given that the battery has survived t_0 time units. The time t_0 represents the time when making the prognosis, e.g., when visiting a workshop with a functioning battery, but one would like to estimate the expected lifetime of the battery or the next suitable time for a battery replacement. The conditional probability function is formally written as

$$\mathcal{B}^{\mathcal{V}}(t; t_0) = P(T > t + t_0 \mid T \geq t_0,\ \mathcal{V}) \tag{1.1}$$

where $\mathcal{V}$ are the measurements from the vehicle retrieved at the workshop at time point t_0. It is worth emphasizing that neither true degradation profiles nor time series of measurements are available for a vehicle, therefore, instead of

determining the RUL, the objective stated here is to estimate the probability of failure or survival at a given time point.

The following problems are addressed in this licentiate thesis: a) the prognostic framework construction process together with an accuracy estimation of the model and b) selection of variables important for prognostics in the case of highly correlated variables. These two directions in the work are done in the Random Survival Forest framework developed in (Ishwaran et al., 2008). There is also an alternative approach explored, in the form of a technical report, that uses neural network, a multilayer perceptron, and a method to predict the conditional reliability function.

1.5 Contributions

The main contributions are summarized for each included paper.

Data-driven battery lifetime prediction and confidence estimation for heavy-duty trucks

Submitted to a journal

A method for estimating the conditional probability function for predicting the lifetime of lead-acid batteries in heavy-duty trucks based on fleet data is demonstrated. A method for estimation of the confidence bands for the RSF method is developed as an extension of existing techniques for estimating variance of Random Forest predictions. The following aspects of the confidence bands are considered: a) their asymptotic behavior and b) usefulness in model selection. A problem of time related variables is identified in the paper by analyzing performance of the models with and without time related variables.

Heavy-duty truck battery failure prognostics using random survival forests

Published in Proceedings of the IFAC Symposium on *Advances in Automotive Control, Norrköping, Sweden, 2016*

A new method for identifying important variables using random survival forests is proposed which is applied for battery failure prognosis. A key property of the approach is that it aims at selecting important variables in a large set of variables where many of them are uninformative and/or correlated. Advantage of our approach is demonstrated by the comparison of variable selection to the existing approaches.

Variable selection for heavy-duty vehicle battery failure prognostics using random survival forests

Published in Proceedings of *European Conference of the PHM Society, Bilbao, Spain, 2016*

For the data set introduced in Section 1.3, two features has been identified, 1) there are few informative variables, and 2) there are highly correlated variables in the dataset. The main contribution is a novel method for identifying important variables, taking these two properties into account, using Random Survival Forests to estimate prognostics model. Compared to "Heavy-duty truck battery failure prognostics using random survival forests", this paper introduces additional mechanics to the variable selection method and its properties in the case of large number of correlated variables important for prediction and significant amount of noise are analyzed. Prognostic models with all and reduced set of variables are generated and differences between the model predictions are discussed, and favorable properties of the proposed approach are highlighted.

Battery failure prognostics using multilayer perceptron

Technical report

In this work a multilayer perceptron model is introduced for the given data from the fleet of the vehicles. The multilayer perceptron model can be used for estimating the conditional probability function used for maintenance planning. The model is designed in a such way that it can work with categorical and numerical variables, i.e., the type of data available in the study. The main contributions are the design of the input layer of the network as well as the choice of the parameters in the hidden layers. Another contribution is the introduction of the classification-survival framework. At first, a class of a vehicle is determined, i.e., failed or censored vehicle, then based on the answer from the classification step the individual lifetime of the battery is predicted. The influence of different network architectures on the prediction is also identified.

Chapter 2

Theoretical basis

This chapter will introduce the theoretical basis and methods that are used in the included papers. Survival data and the machinery behind survival analysis is presented first and an introduction to Random Forest (RF) and Random Survival Forests (RSF) follows next including a derivation of the Infinitesimal Jackknife variance estimate for bagged predictors finishes the chapter.

2.1 Survival analysis

Introduction of the survival analysis below is based on the book by Cox and Oakes (1984). Survival analysis concerns the prediction of future events called failures for a group or groups of individuals. The failure time is defined as either the time when end of life for the particular individual occurs or when the state of individual is such that execution of its intended task is no longer possible with the required quality. Example applications of survival analysis could be prediction of component failure times in a gas turbine, the survival times of patients in medicine or the prediction of economic crisis occurrence in economics.

A distinctive feature of survival analysis is the process called censoring. Usually, only a fraction of the individuals fail during trials or time of observation which means that the remaining part of the population do not experience failures, i.e., the failure times are censored. Typical data used for survival analysis is presented in Figure 2.1 where the time of censoring is depicted with circles and time of failures with squares. Figure 2.1 shows that 3 out of 7 individuals are censored and the method for prediction has to take this information into account. In general, if there is a random variable T_i of failure time for the i^{th} individual and a censoring time which is represented by a random variable c_i, then the observed variable $X_i = \min(T_i, c_i)$ together with a response variable $R_i = 0$ for censored and $R_i = 1$ for failed individuals are the targets for the survival analysis.

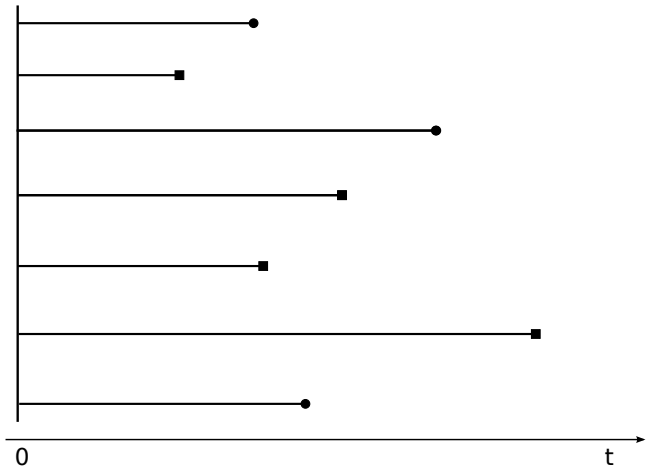


Figure 2.1: Lifetime for failed, represented by squares, and censored, represented by circles, individuals as data for survival analysis.

Survival analysis can be performed either by building models that use only survival time for prediction, i.e., the time when an individual experience a failure or is being censored, or model prediction that can rely on so called variables/covariates/features that explaining the health degradation of individual. For instance, a variable that shows how long a battery has been operated under low voltages or high temperatures can be used in a predictive method.

For a non-negative random variable T which represent time of failure the survival, or reliability, function is defined as

$$R(t) = P(T \geq t). \tag{2.1}$$

The function gives the probability that the failure time T occurs after t time units. Usually, it is handy to work with the probability density function for the random variable T which is related to the reliability function $R(t)$ as

$$f(t) = -R'(t). \tag{2.2}$$

The functions in (2.1) and (2.2) are two different ways to define the distribution of the failure time T. Another special notion in survival analysis is the hazard function which defines probability of instantaneous failure as

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

The hazard function plays an important role in survival analysis. The relationship between the reliability function and the hazard function can be seen by denoting

the cumulative distribution function for the random variable T with $F(t)$ and expanding (2.3) as

$$h(t) = \lim_{\delta \to 0+} \frac{P(t \leq T < t + \delta | T \geq t)}{\delta} = \lim_{\delta \to 0+} \frac{1}{P(T \geq t)} \frac{P(t \leq T < t + \delta)}{\delta} =$$
$$= \frac{1}{R(t)} \lim_{\delta \to 0+} \frac{F(t + \delta) - F(t)}{\delta} = \frac{f(t)}{R(t)} =$$
$$= -\frac{\frac{d}{dt}(1 - F(t))}{R(t)} = -\frac{\frac{d}{dt}R(t)}{R(t)} = -\frac{d}{dt} \log R(t)$$

Then the relation between the hazard and reliability functions is

$$R(t) = \exp\left(-\int_0^t h(u)du\right) = \exp\left(-H(t)\right) \tag{2.4}$$

where $H(t)$ is called the integrated or cumulative hazard rate. Two classes of methods are available when it comes to modelling the survival time T, parametric and non-parametric. The parametric methods assume a parametric distribution as a model for the random variable T, for instance exponential or log-normal. In turn, non-parametric methods do not make any such assumption, making use only of the observations. When one is confident in the underlying distribution of the survival times, then it is better to choose a parametric method since they will give a more accurate prediction compared to a non-parametric one. However, if information about the survival time distribution is not known, non-parametric methods can be used. In the given study non-parametric methods are chosen due to the fact that actual degradation profiles of the batteries are not known.

Concepts introduced above such as the reliability function and hazard rate are defined for the case when the random variable T has a continuous distribution. In general, a discrete distribution is used in non-parametric methods. Therefore, let time points $t_1 < t_2 < \ldots < t_n$ be the time points where either censoring or failure occurs. Then, a non-parametric Kaplan-Meier estimator, (Kaplan and Meier, 1958), of the reliability function is given by

$$\hat{R}(t) = \prod_{t_j < t} (1 - \hat{h}_j) \tag{2.5}$$

where $\hat{h}_j$ is maximum likelihood estimator of hazard rate h_j defined at time point t_j. The estimate $\hat{h}_j$ is represented as

$$\hat{h}_j = \frac{d_j}{r_j} \tag{2.6}$$

where d_j is the number of failed cases at time point t_j and $r_j = \sum_{i=j+1}^{n}(d_i + c_i)$ is the number of available cases at time point t_j with c_i being the number of censored cases at time t_i.

Greenwood's formula, (Cox and Oakes, 1984), is another tool which is often used in the survival analysis. It estimates the variance of the reliability estimate $\hat{R}(t)$ under the assumption that (2.5) is efficient, i.e., reaches its Cramer-Rao bound and using a linearization approach, as follows

$$\text{var}\left[\hat{R}(t)\right] = \left(\hat{R}(t)\right)^2 \sum_{t_j<t} \frac{d_j}{r_j(r_j - d_j)}. \tag{2.7}$$

The confidence bands with $(1-\alpha)\%$ confidence, under a Gaussian assumption, for the reliability estimator $\hat{R}(t)$ are computed as

$$\hat{R}(t) \pm z_\alpha \left(\text{var}\left[\hat{R}(t)\right]\right)^{\frac{1}{2}}$$

at each time point t where $z_\alpha = -\Phi^{-1}(\alpha/2)$ is a quantile of the normal distribution with mean zero and unit variance being computed using cumulative density function $\Phi(x)$.

The Kaplan-Meier estimator and Greenwood's formula are useful tools to estimate the reliability and estimator variance if the classes of the individuals are known, i.e. classes of individuals with different degradation profiles. However, it is not the case for the data under study in the current work, therefore, one can not apply the aforementioned estimates directly.

Another concept from the survival analysis being used in the papers of the current work is the Nelson-Aalen estimator which is a non-parametric estimator of the cumulative hazard rate $H(t)$. The estimate is written in terms of d_j and r_j as

$$\hat{H}(t) = \sum_{t_j<t} \frac{d_j}{r_j} \tag{2.8}$$

and according to (2.4) it holds that $\hat{H}(t) = -\log\left(\hat{R}(t)\right)$. Introduced concepts from the survival analysis will be used in the consecutive sections and form part of the theoretical basis of the given work.

2.2 Bagged predictors

In the selection of a suitable data-driven method for the studied battery prognostic problem the following needs to be taken into account. The underlying baseline hazard functions are not known in the data set under study, in addition, it is not clear how to estimate parameters in the case of a parametric model such as in Cox regression, see (Cox, 1972). Therefore a non-parametric approach is chosen. Random Survival Forest (RSF), (Ishwaran et al., 2008), is a non-parametric method that gives the ability to handle different types of data, direct applicability of the method to survival analysis, and automatic missing data imputation. The output from the RSF model is an estimate of the reliability

function which can be directly used in the analysis of this work. The basic idea of the RSF model is to group individuals with similar degradation profiles and estimate the reliability function for that particular group of individuals.

Before Random Survival Forest is summarized a brief introduction to basic classification and regression trees and Random Forest methods are given. Classification and regression trees are machine learning techniques that maps/predicts a feature or variable space X into a space of outcomes Y by means of binary trees (Breiman et al., 1984) where the features and outcome for a particular case are considered as a pair (x_i, y_i). Target values y_i from the outcome space could be continuous valued in case of regression and discrete in case of a classification problem. A decision tree is a non-linear estimator

$$\hat{\theta}(x_i) = \hat{y}_i \tag{2.9}$$

where $\hat{\theta}(x)$ is built by partitioning the feature space X, which can contain many features/variables, into disjoint regions R_m with some fitting model for each region. For a regression problem a fitting model is a real value that fits data in a region R_m best, for instance the mean, while for the classification the fitting value is, for example the majority class among all classes in the given region. An example of the aforementioned process is illustrated in Figure 2.2 where the feature space X has two variables v_1 and v_2 and regions R_1, R_2 and R_3 are formed in such way that green, blue and red classes are maximally separated, i.e. a region R_i contains as few individuals from the minority classes as possible.

The aforementioned partitioning process happens at every node of the tree, see Figure 2.3 where a structure of the ordinary classification and regression tree is presented. For a basic decision tree the best splitting variable and splitting value is determined in a greedy manner, namely, all variables and every possible splits are accessed based on a cost function. The split with the lowest value of the cost function is then selected. A tree node where a process of splitting stops is called a terminal node, nodes with s_i variables in Figure 2.3. Splitting stops if either a selected minimal number of individuals in the node is reached or the tree has grown to the predetermined value of maximum depth. Decision trees can be applied to data sets with different types of variables and another advantage is interpretability as rules can be built from a single decision tree. A decision tree is a weak classifier, (Hastie et al., 2009), and generally performs well on the training data, however, they may generalize poorly on unseen data, i.e., have big variance of the predictor.

Therefore, ensemble of trees, a Random Forest (RF) model, was introduced by Breiman (2001). There are different implementations of ensemble of trees such as (Dietterich, 2000) and (Ho, 1998), however, the basic Breiman model is described here since the RSF model is an extension of RF. There are two techniques that are distinctive features of the RF method, namely, bootstrap aggregation, also known as bagging, and a step that reduces correlation between trees in the forest. When the number of data samples is small, bootstrap is a powerful method for estimating statistics of an estimator. By sampling from

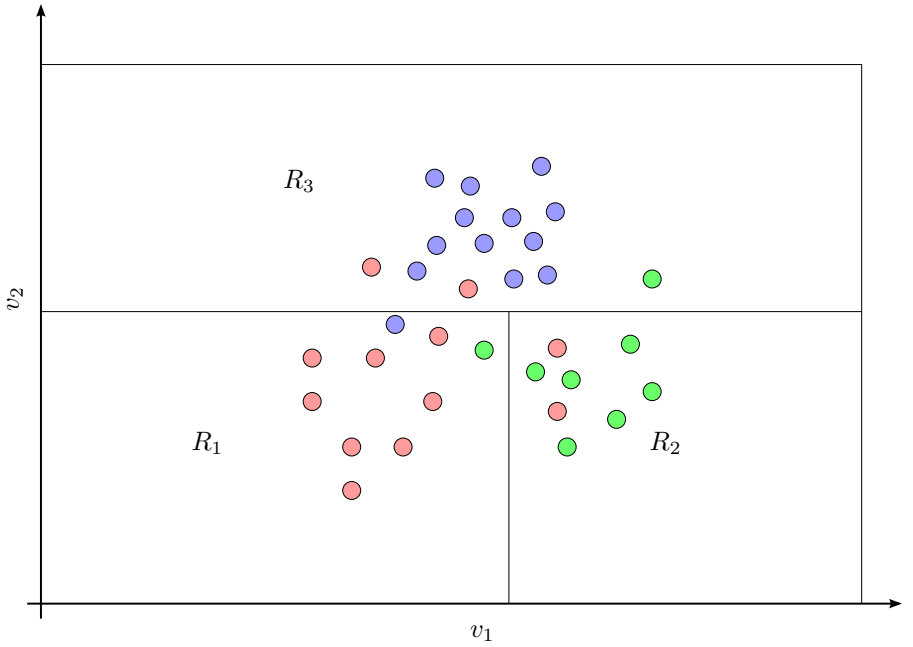


Figure 2.2: Example of the partitioning of the space X with two variables (v_1, v_2) into disjoint regions R_i.

the given data samples with replacement one can construct a large set of new samples that can be used to estimate target statistics. Bootstrap aggregation is an ensemble method that combines predictions from different machine learning models. An example of bootstrapping and bootstrap aggregation is given next.

Example 1 (Bootstrap samples and bootstrap aggregation)**.** *This example will show what a bootstrap sample is and how the results can be used to improve the properties of an estimator. Consider a sample* $(y_1, \ldots, y_5)$ *obtained by sampling from the normal distribution with expected value 2 and variance 4. One possible set of samples is given below*

$$(y_1, \ldots, y_5) = (4.6918, 3.9905, 3.0924, -1.8254, 5.8425).$$

The bootstrap method creates new samples, bootstrap samples, from the original $(y_1, \ldots, y_5)$ *by sampling from it with replacement. For instance, one realization*

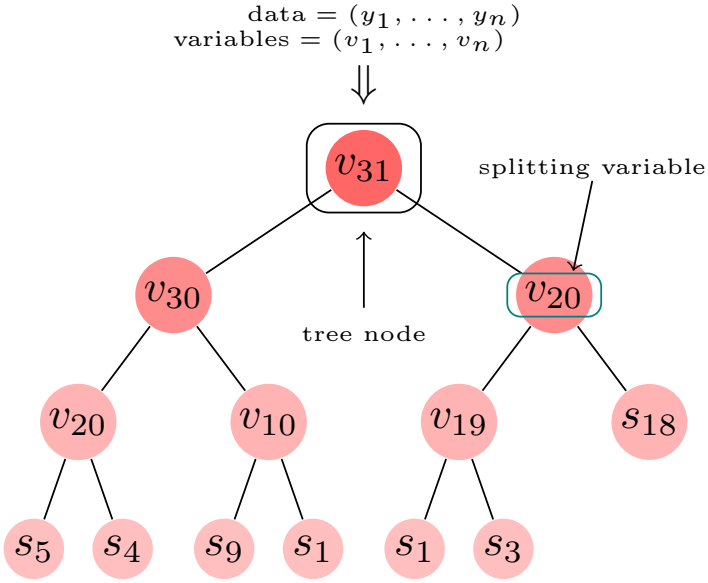


Figure 2.3: Structure of the ordinary binary classification and regression tree.

of 3 bootstrap samples is presented below as

$$(y_1^1, \ldots, y_5^1) = (5.8425, 4.6918, 4.6918, -1.8254, -1.8254)$$
$$(y_1^2, \ldots, y_5^2) = (3.0924, 3.0924, 5.8425, 3.9905, 4.6918)$$
$$(y_1^3, \ldots, y_5^3) = (4.6918, -1.8254, 3.9905, 3.0924, 3.0924).$$

With a non-linear estimator $\hat{\theta} = \theta(y)$, the results of the bootstrap estimates $\hat{\theta}^i = \theta(y^i)$ can be aggregated as

$$\frac{1}{B}\sum_{i=1}^{B}\hat{\theta}^i$$

where B is the number of bootstrap samples. This bootstrap aggregation can be very beneficial in some situations, for example in Random Forest models.

A basic use of using bootstrap samples to estimate variance of an estimator is illustrated in the next example.

Example 2 (Bootstrapping for variance estimation)**.** *This example will show an example how bootstrap samples can be used to estimate statistics of a specific estimator, in this case the variance of a standard estimator for the expected*

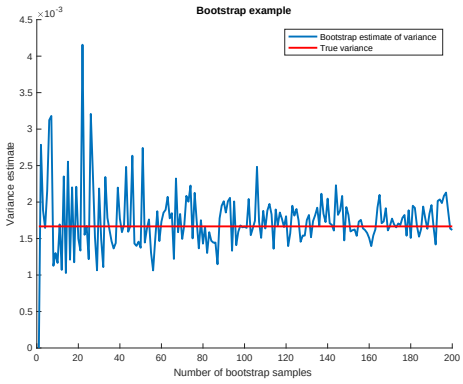


Figure 2.4: Estimate of the estimator variance compared to the true variance as a function of number of bootstrap samples.

value. A main observation of the example is how the method is applicable to any *statistics and without* any *assumptions on underlying distributions.*

Therefore, consider an independent set of samples $(y_1, \ldots, y_N)$ *from an unknown distribution. An expected value estimator is then*

$$\hat{\mu} = \frac{1}{N} \sum_{i=1}^{N} y_i.$$

Assume now that we want an estimate of the covariance of the estimate $\hat{\mu}$*. In this case, since the estimator is simple and linear, the true variance of the estimator is*

$$var\ \hat{\mu} = \frac{1}{N} \sigma_y^2 \tag{2.10}$$

where σ_y^2 *is the variance of the data.*

The bootstrap approach, applicable to any non-linear estimator, is then to generate a set of bootstrap samples y^i $i = 1, \ldots, B$ *and then apply the estimator on each bootstrap sample computing* $\hat{\mu}^i$*. The variance estimate for the estimator is then*

$$\frac{1}{B} \sum_{i=1}^{B} (\hat{\mu}^i - \hat{\mu}^{(\cdot)})^2, \quad \hat{\mu}^{(\cdot)} = \frac{1}{B} \sum_{i=1}^{B} \hat{\mu}^i.$$

Consider the case where the sample y_i *are drawn from a uniform distribution between* 0 *and* 1 *and the number of samples* $N = 50$*. Figure 2.4 shows how the bootstrap variance estimate compares to the theoretical value from* (2.10) *for different number of bootstrap samples* B*. It is clear that the bootstrap method can estimate the variance but with no assumption on linearity or knowledge on underlying distribution.*

In the case of a forest of trees, a number of sets of bootstrap samples are created and then a classification or regression tree model is fitted for each of

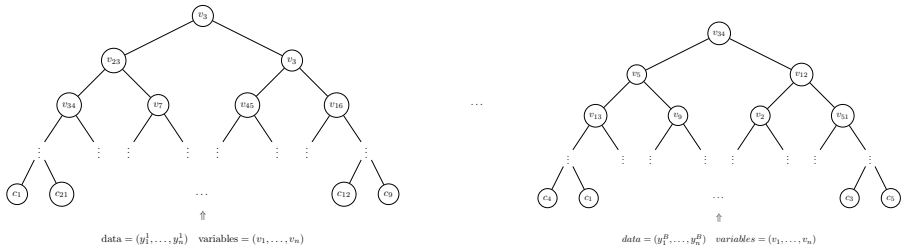


Figure 2.5: Structure of the RF model.

the bootstrap samples. As mentioned, a single tree model is sensitive to unseen data, but by combing outputs from a set of trees, grown on different bootstrap samples, the resulting output has reduced variance of a predictor compared to the single tree model. In regression, the output from a bootstrap aggregation model is the mean of outputs of all trees

$$\hat{\theta}_{\text{BAGG}}(x) = \frac{1}{B} \sum_{i=1}^{B} \hat{\theta}_i(x) \tag{2.11}$$

where $\hat{\theta}_i(x)$ is a tree model fitted to the i^{th} bootstrap sample, and B is the number of trees/bootstrap samples. It was suggested by Breiman (2001) that introducing randomness into the procedure of choosing variables for splitting reduces correlation between trees and increase performance of the aggregated model. Therefore, instead of choosing all m available variables for split at each node, only a fraction p of them is considered. This step also increases speed of the algorithm as it requires less variables to check at each split. Structure of the random forest is presented in Figure 2.5.

A Random Survival Forest (RSF) model is an RF model modified for the purpose of survival analysis (Ishwaran et al., 2008). Structurally, an RSF model is similar to an RF except for the following changes. The cost function used for splitting is so called log-rank test (Ciampi et al., 1986). It is a hypothesis test which compares survival distributions of samples that are formed by dividing data available at the splitting node into two samples which will be the part of the two child nodes. The best split corresponds to a variable with a value under which two samples have as distinctive degradation profiles as possible. The log-rank test is non-parametric and designed for censored data, a type of data encountered in survival analysis. At each terminal node, a node at which splitting no longer is performed, the Nelson-Aalen estimate (2.8) of the cumulative hazard rate is computed (Cox and Oakes, 1984). The estimated cumulative hazard rate $\hat{H}(t)$ of the whole forest is computed by averaging over tree hazard rates and, finally, the estimate $\hat{R}(t)$ of the reliability function is

computed as

$$\hat{R}(t) = e^{-\hat{H}(t)} \tag{2.12}$$

The estimate $\hat{R}(t)$ of the reliability function is the forest output.

2.3 Confidence estimate of a bagged predictor

This section describes the process behind finding IJ estimates for some bagged predictor and, then the explicit expressions for the variance estimates and bias are derived in the case of RF and RSF models. The results are available in the existing literature, but the complete derivation can not be found in any publication. Assume a bagged predictor (2.11) which is complex, nonlinear, and deriving an explicit expression for the estimation covariance is infeasible. Then, one option is to use a bootstrap technique. Since the estimator already uses bootstrap, a bootstrap strategy for estimating the variance would require to compute bootstrap of bootstraps which is not computationally feasible (Efron, 2014). Another approach is to use the original bootstrap samples and structure of the bagged model to estimate the variance of the predictor. One such procedure is the Infinitesimal Jackknife (IJ) variance estimate suggested by (Efron et al., 2014). The theoretical fundamentals are described, based on the works by Efron et al. (2014) and then extended to RSF models.

To summarize the results from (Efron et al., 2014), consider the i^{th} bootstrap sample $\boldsymbol{Y_i^*} = (y_{i1}^*, y_{i2}^*, \ldots, y_{in}^*)$ which is sampled from the initial data set $\boldsymbol{Y} = (y_1, y_2, \ldots, y_n)$ where y_{ij}^* represents the number of times a particular data point y_j, a set of variables for the vehicle from the data set mentioned in Section 1.3, is included in bootstrap sample $\boldsymbol{Y_i^*}$. Introduce a resampling vector as

$$\boldsymbol{P} = (p_1, p_2, \ldots, p_n) \tag{2.13}$$

where p_i denotes the probability of selecting y_i in a bootstrap sample. This vector belongs to a set such that

$$\mathcal{L}_n = \left\{ \boldsymbol{P} : P_i \geq 0, \sum_{i=1}^{n} P_i = 1 \right\} \tag{2.14}$$

The resampling vector represents the weight each data point y_i in the initial sample $\boldsymbol{Y} = (y_1, y_2, \ldots, y_n)$ has in the i^{th} bootstrap sample. For example, the resampling vector $\boldsymbol{P}^0 = (\frac{1}{n}, \ldots, \frac{1}{n})$ is associated with an initial sample $\boldsymbol{Y}$ where each element of the sample has equal weight.

The distribution for the resampling vector $\boldsymbol{P}$ under the bootstrap procedure is a scaled multinominal distribution

$$\boldsymbol{P} \sim \frac{\text{Mult}(n, \boldsymbol{P}^0)}{n}$$

with mean and covariance matrices

$$\left(\boldsymbol{P}^0, \frac{\boldsymbol{I}}{n^2} - \frac{\boldsymbol{P}^{0\prime}\boldsymbol{P}^0}{n}\right).$$

By definition, the variance of $\hat{\theta}_{\mathrm{BAGG}}$ is

$$\mathrm{var}\left[\hat{\theta}_{\mathrm{BAGG}}\right] = E\left[\hat{\theta}_{\mathrm{BAGG}} - E\left[\hat{\theta}_{\mathrm{BAGG}}\right]\right]^2.$$

An expansion of the nonlinear estimator $\hat{\theta}_{\mathrm{BAGG}}$ using directional derivatives around resampling vector $\boldsymbol{P}^0$ keeping only a linear term gives

$$\hat{\theta}_{\mathrm{BAGG}} = \hat{\theta}_{\mathrm{BAGG}}(\boldsymbol{P}) = \hat{\theta}_{\mathrm{BAGG}}(\boldsymbol{P}^0) + (\boldsymbol{P}-\boldsymbol{P}^0)\cdot\boldsymbol{U} + \mathcal{O}((\boldsymbol{P}-\boldsymbol{P}^0)\cdot(\boldsymbol{P}-\boldsymbol{P}^0)'). \tag{2.15}$$

The column vector $\boldsymbol{U}$ consists of the directional derivatives U_i defined as

$$U_i = \lim_{\epsilon\to 0} \frac{\hat{\theta}_{\mathrm{BAGG}}(\boldsymbol{P}^0 + \epsilon(\boldsymbol{\delta}_i - \boldsymbol{P}^0)) - \hat{\theta}_{\mathrm{BAGG}}(\boldsymbol{P}^0)}{\epsilon}, \quad i = 1, \ldots, n \tag{2.16}$$

with $\boldsymbol{\delta}_i$ being the ith coordinate vector. Taking the expectation of (2.15), ignoring higher order terms, gives

$$\begin{aligned} E\left[\hat{\theta}_{\mathrm{BAGG}}\right] &\approx \hat{\theta}_{\mathrm{BAGG}}(\boldsymbol{P}^0) + E\left[\boldsymbol{P} - \boldsymbol{P}^0\right]\cdot\boldsymbol{U} = \\ &= \left\{E\left[\boldsymbol{P} - \boldsymbol{P}^0\right] = E\left[\boldsymbol{P}\right] - \boldsymbol{P}^0 = \boldsymbol{P}^0 - \boldsymbol{P}^0 = 0\right\} = \hat{\theta}_{\mathrm{BAGG}}(\boldsymbol{P}^0). \end{aligned}$$

Thus, the variance of $\hat{\theta}_{\mathrm{BAGG}}$ becomes

$$\begin{aligned} var\left[\hat{\theta}_{\mathrm{BAGG}}\right] &\approx E\left[\hat{\theta}_{\mathrm{BAGG}}(\boldsymbol{P}^0) + (\boldsymbol{P} - \boldsymbol{P}^0)\cdot\boldsymbol{U} - \hat{\theta}_{\mathrm{BAGG}}(\boldsymbol{P}^0)\right]^2 = \\ &= E\left[(\boldsymbol{P} - \boldsymbol{P}^0)\cdot\boldsymbol{U}\right]^2 = E\left[\left(p_1 - \frac{1}{n}, \ldots, p_n - \frac{1}{n}\right)\cdot\begin{pmatrix} U_1 \\ \vdots \\ U_n \end{pmatrix}\right]^2 = \\ &= E\left[\sum_{i=1}^n \left(p_i - \frac{1}{n}\right)U_i\right]^2 = E\left[\sum_{i=1}^n \left(p_i - \frac{1}{n}\right)^2 U_i^2\right] + \\ &+ 2E\left[\sum_{i\neq j} \left(p_i - \frac{1}{n}\right)U_i\left(p_j - \frac{1}{n}\right)U_j\right] = \sum_{i=1}^n E\left[\left(p_i - \frac{1}{n}\right)^2\right]U_i^2 + \\ &+ 2\sum_{i\neq j} E\left[\left(p_i - \frac{1}{n}\right)\left(p_j - \frac{1}{n}\right)\right]U_iU_j = \sum_{i=1}^n \frac{1}{n^2}\left(1 - \frac{1}{n}\right)U_i^2 + \\ &+ 2\sum_{i\neq j}\left(-\frac{1}{n^3}\right)U_iU_j = \frac{1}{n^2}\sum_{i=1}^n U_i^2 - \frac{1}{n^3}\left(\sum_{i=1}^n U_i\right)^2 \end{aligned} \tag{2.17}$$

Now, let us show that the sum of all directional derivatives U_i is 0. First, the gradient vector D is defined as

$$D = \begin{pmatrix} D_1 \\ \vdots \\ D_n \end{pmatrix} \qquad D_i = \frac{\partial}{\partial p_i} \hat{\theta}_{\mathrm{BAGG}}(\boldsymbol{P}) \bigg|_{\boldsymbol{P}=\boldsymbol{P}^0}.$$

Thus, the directional derivative U_i can be expressed as

$$U_i = (\boldsymbol{\delta}_i - \boldsymbol{P}^0) \cdot D.$$

From this follows that

$$\begin{aligned} U_i &= \left(\underbrace{-\frac{1}{n}, \ldots, -\frac{1}{n}}_{\text{i - 1}}, 1 - \frac{1}{n}, -\frac{1}{n}, \ldots, -\frac{1}{n} \right) \cdot \begin{pmatrix} D_1 \\ \vdots \\ D_n \end{pmatrix} = \\ &= \sum_{j \neq i} \left(-\frac{1}{n} \right) \cdot \frac{\partial}{\partial p_j} \hat{\theta}_{\mathrm{BAGG}}(\boldsymbol{P}) \bigg|_{\boldsymbol{P}=\boldsymbol{P}^0} + \left(1 - \frac{1}{n} \right) \cdot \frac{\partial}{\partial p_i} \hat{\theta}_{\mathrm{BAGG}}(\boldsymbol{P}) \bigg|_{\boldsymbol{P}=\boldsymbol{P}^0} \end{aligned} \tag{2.18}$$

It is evident from (2.18) that the sum of U_i is 0. Taking this result into account and using (2.17), the infinitesimal jackknife (IJ) variance estimate $\hat{V}_{\mathrm{IJ}}$ of the true variance $\mathrm{var}\left[\hat{\theta}_{\mathrm{BAGG}}\right]$ of the bagged predictor is

$$\hat{V}_{\mathrm{IJ}} = \frac{1}{n^2} \sum_{i=1}^{n} U_i^2. \tag{2.19}$$

To compute the variance estimator, we then need the directional derivatives. For a bagged estimator $\hat{\theta}_{\mathrm{BAGG}}$, it turns out that there exists an explicit expression for the asymptotic, with respect to number of bootstrap samples B, expression of the directional derivatives. Now follows a derivation of the directional derivatives for an RF model.

Consider again a vector $\boldsymbol{Y}^* = (y_1^*, y_2^*, \ldots, y_n^*)$ which represents the number of times a particular sample appears in the bag/bootstrap sample of a tree. The distribution of $\boldsymbol{Y}^*$ is a multinomial

$$\boldsymbol{Y}^* \sim \mathrm{Mult}(n, \boldsymbol{P}^0).$$

Consider a forest with $B = n^n$ trees which corresponds to all possible samples. Then, the Random Forest estimator $\hat{\theta}_{\mathrm{RF}}(\boldsymbol{P})$ could be written as

$$\hat{\theta}_{\mathrm{RF}}(\boldsymbol{P}) = \sum_{i=1}^{B} P(\boldsymbol{Y}_i^*) t_i^* \tag{2.20}$$

where $P(\boldsymbol{Y}^*)$ is the probability of bootstrap sample under multinomial distribution, t_i^* is the output from the ith tree in the forest. We explicitly state that the Random Forest estimator depends on the resampling vector $\boldsymbol{P}$ associated with the bootstrap sample $\boldsymbol{Y}^*$.

Let $P_0(\boldsymbol{Y}_i^*)$ and $P(\boldsymbol{Y}_i^*)$ denote the probability of a bootstrap sample under the multinomial distributions with $\boldsymbol{P}_0$ and a general $\boldsymbol{P}$ respectively. Then, using the fact that under $\boldsymbol{P}_0$ all bootstrap samples are equally probable

$$P(\boldsymbol{Y}_i^*) = \frac{P(\boldsymbol{Y}_i^*)}{P_0(\boldsymbol{Y}_i^*)} P_0(\boldsymbol{Y}_i^*) = \frac{\frac{n!}{y_{i1}^*!\ldots y_{in}^*!} p_1^{y_{i1}^*} \cdot \ldots \cdot p_n^{y_{in}^*}}{\frac{n!}{y_{i1}^*!\ldots y_{in}^*!} p_0^{y_{i1}^*} \cdot \ldots \cdot p_0^{y_{in}^*}} \cdot \frac{1}{B} = \frac{1}{B} \prod_{k=1}^{n} (np_k)^{y_{ik}^*} \tag{2.21}$$

Combining results in (2.20) with (2.21) gives

$$\hat{\theta}_{\mathrm{RF}}(\boldsymbol{P}) = \sum_{i=1}^{B} \frac{1}{B} \prod_{k=1}^{n} (np_k)^{y_{ik}^*} t_i^*$$

Let a vector $\boldsymbol{P}_j$ be $\boldsymbol{P}_j = \boldsymbol{P}^0 + \epsilon(\boldsymbol{\delta}_j - \boldsymbol{P}^0)$, then

$$\hat{\theta}_{\mathrm{RF}}(\boldsymbol{P}_j) = \sum_{i=1}^{B} \frac{1}{B} \underbrace{\prod_{k=1}^{n} (n(p_0 + \epsilon(\delta_{jk} - p_0)))^{y_{ik}^*}}_{\mathrm{A}(\epsilon)} t_i^*$$

$$A(\epsilon) = (1 + \epsilon(n-1))^{y_{ij}^*} (1-\epsilon)^{\sum_{k \neq j} y_{ik}^*} = (1 + \epsilon(n-1))^{y_{ij}^*} (1-\epsilon)^{n - y_{ij}^*}. \tag{2.22}$$

A Taylor expansion of non-linear function $A(\epsilon)$ around 0 leads to

$$\begin{aligned} A(\epsilon) = 1 + \Big(& y_{ij}^*(n-1)\,(1+\epsilon(n-1))^{y_{ij}^*-1}\,(1-\epsilon)^{n-y_{ij}^*} - \\ & -(n-y_{ij}^*)\,(1+\epsilon(n-1))^{y_{ij}^*}\,(1-\epsilon)^{n-y_{ij}^*-1} \Big)\Big|_{\epsilon=0} \epsilon + \mathcal{O}(\epsilon^2) = \\ & = 1 + n(y_{ij}^* - 1)\epsilon + \mathcal{O}(\epsilon^2) \end{aligned}$$

Substituting the obtained result in the forest estimate from (2.22) gives us

$$\hat{\theta}_{\mathrm{RF}}(\boldsymbol{P}_j) = \frac{1}{B} \sum_{i=1}^{B} \left(1 + n(y_{ij}^* - 1)\epsilon + \mathcal{O}(\epsilon^2)\right) t_i^*. \tag{2.23}$$

Noticing that under multinomial distribution with resampling vector $\boldsymbol{P}^0$ forest estimator becomes

$$\hat{\theta}_{\mathrm{RF}}(\boldsymbol{P}^0) = \sum_{i=1}^{B} \frac{1}{B} \prod_{k=1}^{n} (np_0)^{y_{ik}^*} t_i^* = \frac{1}{B} \sum_{i=1}^{B} t_i^*, \tag{2.24}$$

gives the result for computing the directional derivatives U_j, (2.16), as

$$U_j = \lim_{\epsilon \to 0} \frac{\hat{\theta}_{\mathrm{RF}}(\boldsymbol{P}^0 + \epsilon(\boldsymbol{\delta}_j - \boldsymbol{P}^0)) - \hat{\theta}_{\mathrm{RF}}(\boldsymbol{P}^0)}{\epsilon} =$$
$$= \lim_{\epsilon \to 0} \left(\frac{\frac{1}{B}\sum_{i=1}^{B} t_i^* + \frac{1}{B}\sum_{i=1}^{B} n(y_{ij}^* - 1)\epsilon}{\epsilon} + \frac{\frac{\mathcal{O}(\epsilon^2)}{B}\sum_{i=1}^{B} t_i^* - \frac{1}{B}\sum_{i=1}^{B} t_i^*}{\epsilon} \right) =$$
$$= \frac{n}{B}\sum_{i=1}^{B} \left(y_{ij}^* - 1\right) t_i^*. \quad (2.25)$$

Let us show that the result for U_j in (2.25) can be expressed as

$$U_j = n \cdot \widehat{\mathrm{Cov}}_j \qquad \text{where } \widehat{\mathrm{Cov}}_j = \mathrm{cov}(y_{ij}^*, t_i^*) = \frac{1}{B}\sum_{i=1}^{B} \left(y_{ij}^* - 1\right)(t_i^* - \bar{t})$$

and $\bar{t} = \frac{1}{B}\sum_{i=1}^{B} t_i^*$. Denote column vector $\boldsymbol{Y}_{\cdot j}^*$ as

$$\boldsymbol{Y}_{\cdot j}^* = \begin{pmatrix} y_{1j}^* \\ y_{2j}^* \\ \vdots \\ y_{Bj}^* \end{pmatrix}.$$

The vector $\boldsymbol{Y}_{\cdot j}^*$ shows how many times j^{th} individual appears in all bags. According to the distribution of vector $\boldsymbol{Y}^*$ it holds that

$$E\left[\boldsymbol{Y}_{\cdot j}^* - 1\right] = 0.$$

Taking into account that $B = n^n$ and every possible bootstrap combination is considered, it follows that

$$E\left[\boldsymbol{Y}_{\cdot j}^* - 1\right] = \frac{1}{B}\sum_{i=1}^{B}(y_{ij}^* - 1) = 0$$

Therefore, the directional derivative U_j can be rewritten as

$$U_j = \frac{n}{B}\sum_{i=1}^{B} \left(y_{ij}^* - 1\right) t_i^* = \frac{n}{B}\sum_{i=1}^{B} \left(\left(y_{ij}^* - 1\right) t_i^* - \left(y_{ij}^* - 1\right)\bar{t}\right) =$$
$$= \frac{n}{B}\sum_{i=1}^{B} \left(y_{ij}^* - 1\right)(t_i^* - \bar{t}) = n \cdot \widehat{\mathrm{Cov}}_j$$

Taking into account the last result we can write the estimate of variance of $\hat{\theta}_{\mathrm{RF}}$ as

$$\hat{V}_{\mathrm{IJ}} = \sum_{i=1}^{n} \widehat{\mathrm{Cov}}_i^{\,2} \quad (2.26)$$

where

$$\widehat{\mathrm{Cov}}_i = \frac{1}{B}\sum_{b=1}^{B}(y^*_{bi}-1)(t^*_b-\bar{t}). \tag{2.27}$$

Usually estimators are biased and it would be great to take it into account. It turns out that for the RF model it is possible to find an estimate for the bias in an explicit form. To find a bias expression consider $\hat{V}_{\mathrm{IJ}}$ which is the perfect estimator when $B \to \infty$,

$$\hat{V}^{\infty}_{\mathrm{IJ}} = \sum_{j=1}^{n}(\mathrm{Cov}_j)^2$$

where Cov_j means perfect covariance estimate of (2.27) when $B \to \infty$. Thus, the bias of the estimator can be written as follows

$$\begin{aligned}\mathrm{Bias} = E\left[\hat{V}_{\mathrm{IJ}}\right] - \hat{V}^{\infty}_{\mathrm{IJ}} &= \sum_{i=1}^{n}\left(E\left[\widehat{\mathrm{Cov}}_i^{\,2}\right] - (\mathrm{Cov}_i)^2\right) = \\ &= \sum_{i=1}^{n}\left(E\left[\widehat{\mathrm{Cov}}_i^{\,2}\right] - \left(E\left[\widehat{\mathrm{Cov}}_i\right]\right)^2\right) = \sum_{i=1}^{n}\mathrm{var}\left[\widehat{\mathrm{Cov}}_i\right]\end{aligned}$$

Assuming independence of y^*_{ij} and t^*_i, we have the following for the large enough samples, i.e., when $n \to \infty$,

$$\begin{aligned}\mathrm{Bias} &= \sum_{j=1}^{n}\mathrm{var}\left[\widehat{\mathrm{Cov}}_j\right] = n\cdot\mathrm{var}\left[cov_1\right] = \\ &= n\cdot\mathrm{cov}\left[\frac{1}{B}\sum_{i=1}^{B}\left(y^*_{i1}-1\right)\left(t^*_i-\bar{t}\right); \frac{1}{B}\sum_{j=1}^{B}\left(y^*_{j1}-1\right)\left(t^*_j-\bar{t}\right)\right] = \\ &= \frac{n}{B^2}\sum_{i=1}^{B}\sum_{j=1}^{B}\big(E\left[(y^*_{i1}-1)(y^*_{j1}-1)(t^*_i-\bar{t})(t^*_j-\bar{t})\right] - \\ &\qquad -E\left[(y^*_{i1}-1)(t^*_i-\bar{t})\right]E\left[(y^*_{j1}-1)(t^*_j-\bar{t})\right]\big). \end{aligned} \tag{2.28}$$

Assuming that original sample $\boldsymbol{Y}$ is large enough and that t^*_i and y^*_{ij} are independent, (2.28) simplifies as

$$\begin{aligned}\mathrm{Bias} = \frac{n}{B^2}\sum_{i=1}^{B}\sum_{j=1}^{B}\big(&E\left[(y^*_{i1}-1)(y^*_{j1}-1)\right]E\left[(t^*_i-\bar{t})(t^*_j-\bar{t})\right] - \\ &-E\left[(y^*_{i1}-1)\right]E\left[t^*_i-\bar{t}\right]E\left[(y^*_{j1}-1)\right]E\left[t^*_j-\bar{t}\right]\big). \end{aligned} \tag{2.29}$$

The random variable y_{ij}^* has the following properties

$$E\left[(y_{i1}^* - 1)(y_{j1}^* - 1)\right] = \operatorname{cov}\left[y_{i1}^*; y_{j1}^*\right] = \frac{1}{n} \to 0,$$
$$n \to \infty, b \neq j$$
$$E\left[y_{i1}^* - 1\right] = E\left[y_{j1}^* - 1\right] = 0$$
$$E\left[(y_{i1}^* - 1)^2\right] = \operatorname{var}\left[y_{i1}^*\right] = 1 - \frac{1}{n} \to 1,$$
$$n \to \infty, b = j$$

Therefore, if it is assumed that size of sample $n \to \infty$ and n tends to infinity faster than B, the bias becomes

$$\text{Bias} = \frac{n}{B^2} \sum_{i=1}^{B} \left(E\left[(y_{i1}^* - 1)^2\right] E\left[(t_i^* - \bar{t})^2\right]\right) = \frac{n}{B^2} \sum_{j=1}^{B} (t_j^* - \bar{t})^2 \tag{2.30}$$

An improved unbiased estimator of $\operatorname{var}\left[\hat{\theta}_{\text{RF}}\right]$ using the results in (2.26) and (2.30) can then be written as

$$\hat{V}_{\text{IJ-U}} = \hat{V}_{\text{IJ}} - \frac{n}{B^2} \sum_{b=1}^{B} (t_b^* - \bar{t})^2. \tag{2.31}$$

2.3.1 IJ variance estimate of the ratio of random variables

This subsection demonstrates the extension of the IJ variance estimate to the RSF model and lifetime function as in (1.1). The lifetime function could be expressed through the ratio of the reliability function estimates as

$$\hat{\mathcal{B}}^{\mathcal{V}}(t, t_0) = \frac{\hat{R}^{\mathcal{V}}(t + t_0)}{\hat{R}^{\mathcal{V}}(t)} \tag{2.32}$$

where $\hat{R}^{\mathcal{V}}(t) = P(T \geq t|\mathcal{V})$ is the output from the RSF model. There are two main differences between IJ variance estimate of the RF model compared to the variance estimate of the lifetime function $\mathcal{B}^{\mathcal{V}}(t, t_0)$. First, the output of the RF model is either a class or regression value, but in the RSF case the output function is time dependent, and secondly, the lifetime function is a ratio of the reliability estimates.

For the first difference mentioned above, the reliability function is computed on the predefined grid of time points, the variance estimate $\hat{V}_{\text{IJ}}^{\text{RSF}}(t)$ of the true forest variance $\operatorname{var}\left[\hat{\theta}_{\text{RSF}}\right]$ becomes

$$\hat{V}_{\text{IJ}}^{\text{RSF}}(t) = \sum_{i=1}^{n} \widehat{\text{Cov}}_i^2(t) \tag{2.33}$$

where

$$\widehat{\mathrm{Cov}}_i(t) = \frac{1}{B}\sum_{b=1}^{B}(y_{bi}^* - 1)(\hat{R}_b^{\mathcal{V}}(t) - \hat{R}^{\mathcal{V}}(t)) \tag{2.34}$$

Here, the reliability $\hat{R}_b^{\mathcal{V}}(t)$ is the output reliability from the bth tree for a particular vehicle with data $\mathcal{V}$ and $\hat{R}^{\mathcal{V}}(t)$ is the output from the forest. These values correspond to t_b^* and $\bar{t}$ in (2.26) respectively. An unbiased IJ variance estimate $\hat{V}_{\text{IJ-U}}^{\text{RSF}}$ in analogy with Efron's estimate is then

$$\hat{V}_{\text{IJ-U}}^{\text{RSF}}(t) = \hat{V}_{\text{IJ}}^{\text{RSF}}(t) - \frac{n}{B^2}\sum_{b=1}^{B}(\hat{R}_b^{\mathcal{V}}(t) - \hat{R}^{\mathcal{V}}(t))^2 \tag{2.35}$$

For the second property, the variance estimate for the lifetime function $\hat{\mathcal{B}}^{\mathcal{V}}(t, t_0)$, which is a ratio of the outputs of the random survival forest, is estimated and summarized in the next theorem.

Theorem 1. *Let $\mathcal{B}^{\mathcal{V}}(t, t_0)$ be the battery lifetime function. Then*

$$\hat{\mathcal{B}}^{\mathcal{V}}(t, t_0) = \frac{\hat{R}^{\mathcal{V}}(t + t_0)}{\hat{R}^{\mathcal{V}}(t_0)}$$

is the RSF estimate of $\mathcal{B}^{\mathcal{V}}(t, t_0)$ and a first order IJ variance estimate is given by

$$var\left[\hat{\mathcal{B}}^{\mathcal{V}}(t, t_0)\right] \approx \left(\frac{\mu_X}{\mu_Y}\right)^2 \cdot \left(\frac{var[X]}{\mu_X^2} + \frac{var[Y]}{\mu_Y^2} - 2\frac{cov[X, Y]}{\mu_X \mu_Y}\right) \tag{2.36}$$

where the random variable X is the reliability function $\hat{R}^{\mathcal{V}}(t + t_0)$ at time point $t + t_0$ and the random variable Y is the reliability function $\hat{R}^{\mathcal{V}}(t_0)$ at time point t_0 and

$$\begin{aligned}
\mu_X &\approx \hat{R}^{\mathcal{V}}(t + t_0) \\
\mu_Y &\approx \hat{R}^{\mathcal{V}}(t_0) \\
var[X] &= \hat{V}_{IJ\text{-}U}^{RSF}(t + t_0) \\
var[Y] &= \hat{V}_{IJ\text{-}U}^{RSF}(t_0) \\
cov[X, Y] &= cov_{Bias}[X, Y] - Bias
\end{aligned}$$

Result for the estimation of $cov[X, Y]$ is given in Lemma below.

Proof. As mentioned in (2.32) lifetime function can be expressed as ratio of the reliability functions $\hat{R}^{\mathcal{V}}(t)$. Assume that $\hat{R}^{\mathcal{V}}(t + t_0)$ is a random variable X and $\hat{R}^{\mathcal{V}}(t_0)$ is a random variable Y. Then, the variance of the lifetime function can be estimated using a Taylor series expansion as (2.36) where instead of μ_X and μ_Y the outputs from the forest $\hat{R}^{\mathcal{V}}(t + t_0)$ and $\hat{R}^{\mathcal{V}}(t_0)$ are used at

time $t + t_0$ and t_0 respectively. The variances $\text{var}[X]$ and $\text{var}[Y]$ correspond to IJ variance estimates $\hat{V}_{\text{IJ-U}}^{\text{RSF}}(t)$ computed at time $t + t_0$ and t_0 respectively. Covariance $\text{cov}[X, Y] = \widehat{\text{cov}}\left[\hat{R}^{\mathcal{V}}(t + t_0), \hat{R}^{\mathcal{V}}(t_0)\right]$ is a covariance between two random variables which are represented by the values of two points from the reliability curve $\hat{R}^{\mathcal{V}}(t)$ at time $t + t_0$ and t_0. □

The only missing part and the main contribution to the theorem is the derivation of $\text{cov}[X, Y] = \widehat{\text{cov}}\left[\hat{R}^{\mathcal{V}}(t + t_0), \hat{R}^{\mathcal{V}}(t_0)\right]$ which is given in the next lemma.

Lemma 1. *Let $\hat{R}^{\mathcal{V}}(t)$ be an RSF model with B trees grown on the original sample $\boldsymbol{Y} = (y_1, y_2, \ldots, y_n)$ with size n. Assume that the tree output $\hat{R}_b^{\mathcal{V}}(t)$ is independent from one data point y_{ij}^* from the ith bag, then an asymptotic expression of the infinitesimal jackknife estimate of $\widehat{cov}\left[\hat{R}^{\mathcal{V}}(t + t_0), \hat{R}^{\mathcal{V}}(t_0)\right]$ and its bias correction are*

$$cov[X, Y] = cov_{Bias}[X, Y] - Bias \tag{2.37}$$

where

$$cov_{Bias}[X, Y] = \widehat{cov}\left[\hat{R}^{\mathcal{V}}(t + t_0), \hat{R}^{\mathcal{V}}(t_0)\right] = \sum_{i=1}^{n} \widehat{Cov}_i(t_0)\widehat{Cov}_i(t + t_0) \tag{2.38}$$

$$Bias = \frac{n}{B^2}\sum_{i=1}^{B}(\hat{R}_i^{\mathcal{V}}(t_0) - \hat{R}^{\mathcal{V}}(t_0))(\hat{R}_i^{\mathcal{V}}(t + t_0) - \hat{R}^{\mathcal{V}}(t + t_0)) \tag{2.39}$$

as the sample size $n \to \infty$, the number of trees $B \to \infty$, and n tends to infinity faster than B.

Proof. Following the steps similar to derivation of variance estimate and its bias of the RF estimator formula (2.38) and (2.39) can be achieved. Full derivation is given in "Data-driven battery lifetime prediction and confidence estimation for heavy-duty trucks" paper, included as Paper C in this thesis. □

References

L. Breiman. Random forests. *Machine learning*, 45(1):5–32, 2001.

L. Breiman, J. Friedman, R. Olshen, and Stone C. *Classification and regression trees*. Taylor and Francis, 1984.

J. Cheng and D.M. Titerington. Neural networks: A review from a statistical perspective. *Statistical Science*, 9(1):2–54, 1994.

A. Ciampi, J. Thiffault, J.-P. Nakache, and B. Asselain. Stratification by stepwise regression, correspondence analysis and recursive partition: A comparison of three methods of analysis for survival data with covariates. *Computation Statistics and Data Analysis*, 4:185–205, 1986.

D. R. Cox and D. Oakes. *Analysis of survival data*, volume 21. CRC Press, 1984.

D.R. Cox. Regression model and life-table. *Journal of the Royal Statistical Society*, 34(2):187–220, 1972.

M. Daigle and K. Goebel. A model-based prognostics approach applied to pneumatic valves. *International Journal of Prognostics and Health Management*, 2(2):1–16, 2011.

T.G. Dietterich. An experimental comparison of three methods for constructing ensembles of decision trees: bagging, boosting, and randomization. *Machine Learning*, 40(2):139–157, 2000.

B. Efron. Estimation and accuracy after model selection. *Journal of the American Statistical Assosiation*, 109:991–1007, 2014.

B. Efron, T. Hastie, and S. Wager. Confidence intervals for random forests: The jackknife and the infinitesimal jackknife. *Journal of Machine Learning Research*, 15:1625–1651, 2014.

Y. Fan, S. Nowaczyk, and T. Rögnvaldsson. Evaluation of self-organized approach for predicting compressor faults in a city bus fleet. *Procedia Computer Science*, 53:447–456, 2015.

H. Hanachi, J. Liu, A. Banerjee, Y. Chen, and A. Koul. A physics-based modeling approach for performance monitoring in gas turbine engines. *IEEE Transactions on Reliability*, 64(1), 2015.

T. Hastie, R. Tibshirani, and J. Friedman. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction.* Springer, 2009.

T.K. Ho. The random subspace method for constructing decision forests. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 20(8):832–844, 1998.

H. Ishwaran, U. Kogalur, E. Blackstone, and M. Lauer. Random survival forests. *The Annals of Applied Statistics*, pages 841–860, 2008.

E. L. Kaplan and P. Meier. Nonparametric estimation from incomplete observations. *Journal of the American Statistical Association*, 53(282):457–481, 1958.

K. Medjaher, D. A. Tobon-Mejia, and N. Zerhouni. Remaining useful life estimation of critical components with application to bearings. *IEEE Transactions on Reliability*, 61(2), 2012.

R. Prytz, S. Nowaczyk, T. Rögnvaldsson, and S. Byttner. Predicting the need for vehicle compressor repairs using maintenance records and logged vehicle data. *Engineering applications of artificial intelligence*, 41:139–150, 2015. ISSN 0952-1976.

M. Roemer, C. Byington, G. Kacprzynski, and G. Vachtsevanos. An overview of selected prognostic technologies with reference to an integrated phm architecture. In *Proceedings of the First International Forum on Integrated System Health Engineering and Management in Aerospace*, Napa, CA, USA, 2005.

B Saha and K. Goebel. Modeling li-ion battery capacity depletion in a particle filtering framework. In *Proceedings of the Annual Conference of the Prognostics and Health Managment Society*, San Diego, CA, USA, 2009.

F. Zhao, Z. Tian, E. Bechhoefer, and Y. Zeng. An integrated prognostics method under time-varying operating conditionds. *IEEE Transactions on Reliability*, 64(2), 2015.

Publications

Paper A

Heavy-duty truck battery failure prognostics using random survival forests*

A

*Published in Proceedings of the *IFAC Symposium on Advances in Automotive Control, Norrköping, Sweden, 2016.*

Heavy-duty truck battery failure prognostics using random survival forests

Sergii Voronov, Daniel Jung, and Erik Frisk

Vehicular Systems, Department of Electrical Engineering, Linköping University, SE-581 83 Linköping, Sweden.

Abstract

Predicting lead-acid battery failure is important for heavy-duty trucks to avoid unplanned stops by the road. There are large amount of data from trucks in operation, however, data is not closely related to battery health which makes battery prognostic challenging. A new method for identifying important variables for battery failure prognosis using random survival forests is proposed. Important variables are identified and the results of the proposed method are compared to existing variable selection methods. This approach is applied to generate a prognosis model for lead-acid battery failure in trucks and the results are analyzed.

1 Introduction

Heavy-duty trucks are important for transporting goods, working at mines, or construction sites and it is vital that vehicles have a high degree of availability. In particular, this means to avoiding unplanned stops by the road which does not only cost due to the delay in delivery, but can also lead to damaged cargo.

One cause of unplanned stops is a failure in the electrical power system, and in particular the lead-acid starter battery. The main purpose of the battery is to power the starter motor to get the diesel engine running, but it is also used to, for example, power auxiliary units such as heating and kitchen equipment.

Prognostics and health management is an important part to prevent unexpected failures by more flexible maintenance planning. The purpose is to replace the battery before it fails but avoid changing it too often. Coarsely, there are two main approaches in prognostics, data-driven and model-based techniques but also hybrid approaches that combines the two are possible. Model-based prognostics utilizes a model of the monitored system and the fault to monitor to predict the degradation rate and Remaining Useful Life (RUL), see for example (Daigle and Goebel, 2011). Statistical data-driven methods generate a prediction model based on training data to predict RUL, see for example (Si et al., 2011), and is the approach followed here.

The main contribution in this work is a data-driven method to identify important variables from a set of variables, where many are not relevant for lead-acid battery failure prognosis, and use them to build prognostic models. The goal is to find important variables to design a battery failure prognostics model for automotive applications based on random survival forests (Ishwaran et al., 2008). This type of analysis is also important to better understand which factors that are correlated with battery failure rate and also what is causing it.

The outline is as follows. The problem is motivated in Section 2 and some background on random survival forests and variable importance are given in Section 3. Evaluation of existing methods for variable importance in random survival forests is presented in Section 4 showing the need for methodological developments in variables selection. The proposed variable selection method is described in Section 5. Then, the method is analyzed in detail in Section 6 and used to generate a random survival forest prognostic model in Section 7. Finally, some conclusions are presented in Section 8.

2 Problem motivation

The prognostic problem studied here is to estimate a battery lifetime prediction function based on recorded vehicle data. The lifetime prediction function is defined as

$$\mathcal{B}^{\nu}(t; t_0) = P(T > t + t_0 \mid T \geq t_0,\ \nu)$$

where T is the random variable failure time of the battery and ν the vehicle data at $t = t_0$. The function $\mathcal{B}^{\nu}(t; t_0)$ is a function of t and gives the probability that

the battery will function at least t time units after t_0. The data ν is recorded operational data for a specific vehicle which is further described in Section 2.1.

The reliability function (Cox and Oakes, 1984) is defined as

$$R(t) = P(T \geq t) \tag{1}$$

which is the probability that the battery of the specific vehicle will survive at least t time units. Then, the battery lifetime prediction function can be rewritten using the reliability function as

$$\mathcal{B}^{\nu}(t; t_0) = P(T > t + t_0 \mid T \geq t_0,\ \nu) = \frac{R^{\nu}(t + t_0)}{R^{\nu}(t_0)}. \tag{2}$$

Random Survival Forests (RSF) is a data-driven method that can be used for computing maximum-likelihood estimates of the reliability function, as illustrated by Fig. 1. The main objective in this work is to use Random Survival Forests to identify, from data, which vehicle data that is relevant for building RSF models to predict battery failures.

Figure 1: A random survival forest computes the maximum likelihood estimate $\hat{R}^{\nu}(t)$ of the reliability function given a vehicle represented by the data ν. With the estimate $\hat{R}^{\nu}(t)$, the battery lifetime prediction function $\mathcal{B}^{\nu}(t; t_0)$ in (2) can be computed.

2.1 Operational data

In this work a vehicle fleet database is provided, where one snapshot of data is available from each vehicle including information regarding how the truck has been used and the configuration of the specific truck. There is also information if the battery has failed or not. The database contains a lot of information from the truck, not always related to battery degradation, meaning that it is not known what available information is relevant for this specific task. Therefore, it is relevant to identify which variables are relevant for predicting battery lifetime. Previous works considering this vehicle data set are presented in (Frisk et al., 2014) and (Frisk and Krysander, 2015).

The choice of using RSF is motivated by the properties of the available database. Its main characteristics can be summarized as follows:

- 33603 vehicles from 5 EU markets
- 284 variables stored for each vehicle snapshot

- A single snapshot per vehicle
- Heterogeneous data, i.e., it is a mixture of categorical and numerical data
- Availability of histogram variables
- Censoring rate more than 90 percent
- Significant missing rate

The database contains different types of variables, including both categorical and numerical data. The censoring rate refers to that less than 10 percent of the vehicles in the database have had battery failures. This means that for most vehicles it is not known how long the battery will last. Also, there is a significant amount of missing data for the different vehicles, a property of database handled by RSF. One reason for the missing rate is due to the fact that data was recorded for different type of vehicles for which some variables are not applicable.

Another main characteristic of the database is that there are no time series available for a vehicle. It means that there is only one snapshot ν of the variables in the database for each vehicle. Information describing how the vehicle has been used is stored as histogram data where different variables represent how often specific sensor data is measured within different intervals. For example, there is a histogram describing how much time the vehicle has been subjected to different ambient temperatures.

When applying RSF to the data in the database, the objective is to find classes of vehicles with similar battery degradation properties. The reliability computed for a given class is an approximation of the true vehicle reliability which can be used to prognose battery failure. Due to the non-specific purpose of the database, it is probable that only small number of variables from set ν influence prediction of the battery failure rate. Thus, identifying the important variables in order to remove irrelevant ones, may improve the performance of a battery prognosis model. This problem is considered and explained in the successive subsection.

2.2 Variable selection using Random Survival Forests

The problem of identifying a set of important variables from a large set of variables is a relevant topic in machine learning, usually referred to as variable, or feature, selection, see (Guyon and Elisseeff, 2003). There are several reasons why variable selection is important when working with data-driven models. First, it is possible to improve the prediction performance by reducing the number of variables, for example, the quality of the predictor may become bad if the number of noisy variables (those that have no effect on battery failures) is large.

In the following illustrative example, two RSF are trained using synthetic data to show how the number of noisy variables can have a negative impact on prognostics performance.

Synthetic data is created with the following properties. Let h_0 be a constant nominal hazard rate h_0 for battery failure. The hazard rate

$$h(t) = \lim_{dt \to 0} \frac{P(t \leq T < t + dt \mid t \leq T)}{dt} \tag{3}$$

represents the probability of a battery failure at a particular time t, see (Cox and Oakes, 1984) for more details. In this example, the hazard rate does not change with time and the nominal hazard rate corresponds to an expected 10 years of battery life. It is assumed that there is one variable v_1 that explains how vehicle usage profile influences failure rate and changes h_0 to three hazard rates

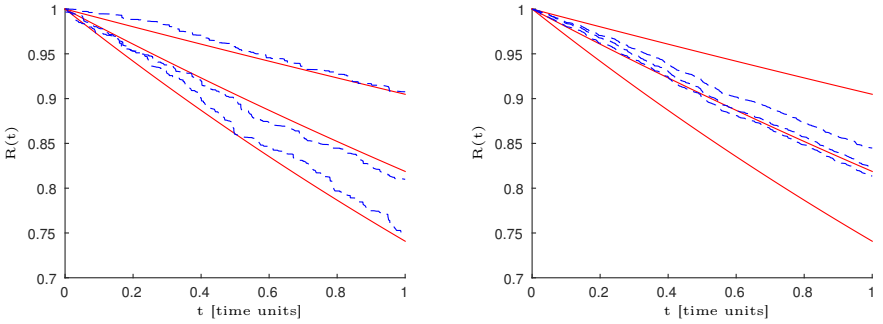
$$h = \begin{cases} 1 \cdot h_0, & \text{if } v_1 = 1 \\ 2 \cdot h_0, & \text{if } v_1 = 2 \\ 3 \cdot h_0, & \text{if } v_1 = 3. \end{cases} \tag{4}$$

The scaling factors show how particular usage of the vehicle, described by v_1, changes the failure rate. Thus, there are three classes of batteries with different degradation profiles. Data for 3000 vehicles is generated with a censoring rate about 80 percent. The censoring rate is selected high to resemble the real vehicle database since censoring rate significantly affects the prediction performance of the RSF model. Two models with different numbers of noisy variables are considered to observe how it changes the RSF prediction.

In the first dataset, two noisy variables are added in addition to v_1, and in the second dataset, 100 noisy variable are added to v_1. After generating two RSF models, one for each dataset, one vehicle from each degradation profile is sampled from validation data and fed to the forest to generate predictions. It is shown in Fig. 2 (a) that predictions from the RSF for the case of 2 noisy variables (dashed blue curves) are following the theoretical reliability functions (red solid curves) significantly better than the predictions from the RSF for the case with 100 noisy variables, see Fig. 2 (b). Note that comparing the results shows a larger number of noisy variables results in worse prediction. The estimated reliability functions follow the theoretical values better with fewer noisy variables. This is something that can be expected.

One measure to evaluate prediction performance of RSF is error rate which should be low and is discussed further in Section 3. The error rate for the case with two noisy variables is 0.4088, for the case with 100 noisy variables is 0.4188. An important observation is that both cases give comparable error rates. However, Fig. 2 shows that there is a significant difference between the two predictors indicating the limitations of using error rate as a performance measure. The given situation happens due to the fact that for the case with a large number of noisy variables, it is hard for the model-building algorithm to find the relevant variables.

This example is illustrative, showing the effects of keeping a lot of noisy variables when generating the RSF model. The true reliability curves are in general unknown but the evaluation using the simulated data shows the advantage



(a) Model 1. Important variable v_1 and 2 noisy.

(b) Model 2. Important variable v_1 and 100 noisy.

Figure 2: Predictive performance of RSF for different amount of noisy variables evaluated on synthetic data. Blue dashed curves correspond to RSF predictions, red solid curves - to theoretical reliabilities.

Table 1: Example of forest generation time given different number of variables.

Number of variables	Time (s)
3	41.6
51	104.3
101	165.9
201	275.19

of reducing the number of noisy variables to improve prediction performance. It motivates the relevance of finding the important variables in a set of data and at the same time remove noisy variables.

A second motivation for variable selection is better interpretability of the results. It is often useful to understand which factors are important for battery failure to utilize this knowledge, for instance, for engineers to improve the design of the vehicle to mitigate degradation of the battery, or to design better models for understanding battery degradation. The interpretability of the model is easier when the model is based on fewer variables.

A third motivation is to reduce model generation time. By reducing the number of variables used for generating the RSF, computational time can be saved. Table 1 shows time spent to grow random survival forest models for different number of variables on a standalone PC. There is a linear dependence of time on number of variables.

The motivations discussed in this subsection show that variable selection is a relevant problem when generating prognosis models.

3 Random survival forests

Random survival forest is here used to make predictions of battery degradation in terms of the lifetime function $B^\nu(t, t_0)$ in (2). This section will give a brief overview of the basic principles and describe what are the basic tools for variables selection related to the given method. RSF was first introduced by (Ishwaran et al., 2008). It is a survival analysis (Cox and Oakes, 1984) extension of a machine learning method called Random Forest (RF) (Breiman, 2001) which is a decision tree based classifier mostly used for regression and classification problems. In this work, the RSF models are generated in `R` using the `RandomForestSRC` package (Ishwaran and Kogalur, 2007).

The difference between an ordinary decision tree classifier and a random forest is that there is randomness of two kinds injected into the process of growing the forest. The first source is the usage of a bootstrap procedure. Each tree is grown using its own bag of cases which are sampled from the training set. Second, for each node in a tree, splitting variables are selected from a randomly sampled subset. RSF extends the RF approach to right-censored survival data, i.e., objects in the study without experiencing a failure. The output from each tree $\mathcal{T}$ is the Nelson-Aalen estimate of cumulative hazard function (Cox and Oakes, 1984).

Let $t_1^\mathcal{T} < t_2^\mathcal{T} < \ldots < t_N^\mathcal{T}$ be N distinct event times when failures of objects under study occur. Then, the Nelson-Aalen estimate for tree $\mathcal{T}$ and vehicle (data) ν is

$$\hat{H}_\mathcal{T}(t|\nu) = \sum_{t_j^\mathcal{T} \le t} \frac{f_{j,n_i}}{s_{j,n_i}} \tag{5}$$

where f_{j,n_i} and s_{j,n_i} are number of failures and survived objects in terminal node n_i of a tree $\mathcal{T}$ at event time $t_j^\mathcal{T}$ respectively. Terminal node n_i is determined by dropping vehicle ν down through the forest. The cumulative hazard estimate $\hat{H}(t|\nu)$ for the whole forest is received by averaging over all $\hat{H}_\mathcal{T}(t|\nu)$. Finally, reliability function $R^\nu(t)$ from (2) obtained from the fact

$$R^\nu(t) = e^{-\hat{H}(t|\nu)} \tag{6}$$

and then $\mathcal{B}^\nu(t; t_0)$ can be computed from (2).

3.1 Prediction error

A performance measure of the RSF is the prediction error (Ishwaran and Kogalur, 2007). It estimates the probability that for two randomly selected out of bag objects, i.e., not used in growing the forest, RSF incorrectly ranks the battery lifetime. It should be noted that prediction error does not fully capture performance of the model. The example in Section 2 shows that the two RSF models generate predictions with similar prediction error. However, it is shown in Fig. 2 that the quality of the predictions is significantly different.

3.2 Measures of variable importance and RSF

There is a tool incorporated in RSF called variable importance VIMP. It measures for a given variable the increase in prediction error when the variable is randomized when used as a splitting variable in the forest. A larger increase indicates that the variable is important for correct classification while a low increase (or even a decrease) in prediction error indicates that the variable is not important.

VIMP is a candidate tool for variable selection by selecting a subset of variables with sufficiently high VIMP values. The variable selection can be done by manually selecting a threshold to separate important from noisy variables. However, previous analyses, (Ishwaran et al., 2011), have shown that VIMP can have problems when there are many correlated variables, a situation that is expected in our case. If several important variables are correlated they will share importance and the computed VIMP will be low even if the variables are important. Thus, there is a risk that important variables will be lost and result in degraded prediction performance. It should be noted that it is not necessary that VIMP fails in our case, but uncertainty motivates an investigation of an alternative approach in selecting important variables.

As an alternative to VIMP, a candidate measure called minimal depth for variable selection in RSF has been proposed (Ishwaran et al., 2010, 2011). Minimal depth for variable v is the distance from the root to the closest node where it appears. The motivation for this measure is that important variables should have a higher probability to be selected as splitting variables at low levels, close to the root, when generating trees. Thus, the average minimal depth for important variables in the forest should be lower compared to noisy variables. A distribution for minimal depth D_v of noisy variables can be derived as (Ishwaran et al., 2010, 2011)

$$P(D_v = d \mid v \text{ is noisy variable}) = \\ = \left(1 - \frac{1}{p}\right)^{L_d} \left[1 - \left(1 - \frac{1}{p}\right)^{l_d}\right], \; 0 \leq d \leq D(T) - 1 \quad (7)$$

where $D(\mathcal{T})$ is a depth of a tree, l_d is number of nodes at depth d, $L_d = l_0 + l_1 + \ldots + l_{d-1}$ and p is number of variables chosen to split node. Then, a threshold to separate important variables from noisy variables can be selected as the mean value for variable distribution (7). If the minimal depth measure of a variable mean value is less than the threshold, it is treated as important, otherwise as noise. The minimal depth measure is evaluated in (Ishwaran et al., 2010) and (Ishwaran et al., 2011) where it is shown to be successful for finding important variables in problems with few important variables and large number of noisy ones, even when the data set is relatively small.

4 VIMP and minimal depth evaluation

VIMP and minimal depth are used to analyze the variables in the vehicle database. For the analysis, three random variables were generated and included into the database to evaluate if the two approaches are able to identify them as non-important. VIMP is evaluated and the value for different variables is shown in Fig. 3. Large positive values correspond to important variables, while values close to zero or negative to non-important variables. To compare the different variable selection methods, five specific variables in the database are highlighted. Four of them are variables that can intuitively contain information about battery degradation. The first one shows if there are battery powered kitchen facilities in a truck, idnicating that the barrery is used not only for starting the combustion engine. Low batttery voltages and low temperatures are important for battery health, and the second variable is therefore a histogram bin with low temperatures of battery voltage histogram. Further, starter motor time and road slope are two bins from respective histograms where first one correlates with battery load and the second one with vehicle usage. The last variable, noise, is one of the added noisy variables and used for testing purposes. It could be seen in Fig. 3 that battery voltage and kitchen equipment are identified as important and noise variable as non-important. It is a positive sign. However, there is no confidence that road slope and starter motor time are not important, because of the problem of the correlated variables VIMP has.

The minimal depth approach is applied to the vehicle database using the recommended configuration described in (Ishwaran et al., 2011). The result of the minimal depth approach is shown in Fig. 4. The x-axis is the mean minimal depth and the y-axis show the mean value of the second minimal depth. Second minimal depth is the distance to the root from a node, in another branch of the tree, where the variable appears the second time. Important variables are thus expected to appear in the lower left corner and non-important in the upper right. The computed threshold, based on (7), is shown as the red vertical line. Most variables are located below the threshold, including the known noisy variable, meaning that the variable selection is not able to distinguish the important variables from noisy variables.

The minimal depth of one of the noisy variables shown as a blue cross in Fig. 4 and is located lower than the computed threshold. The other two noisy variables have similar positions. The minimal depth approach was not able to remove the noisy variables and did not work satisfactory. The previous results presented in (Ishwaran et al., 2011) were based on medical databases. There could be several reasons for different performances where different types of data in the databases could be one reason. Another is the censoring rate, which for the vehicle database is more than 90 percent and much higher than considered in the previous paper. The analysis shows that there seem to be limitations with the existing proposed importance measures and that they are not suitable in this case.

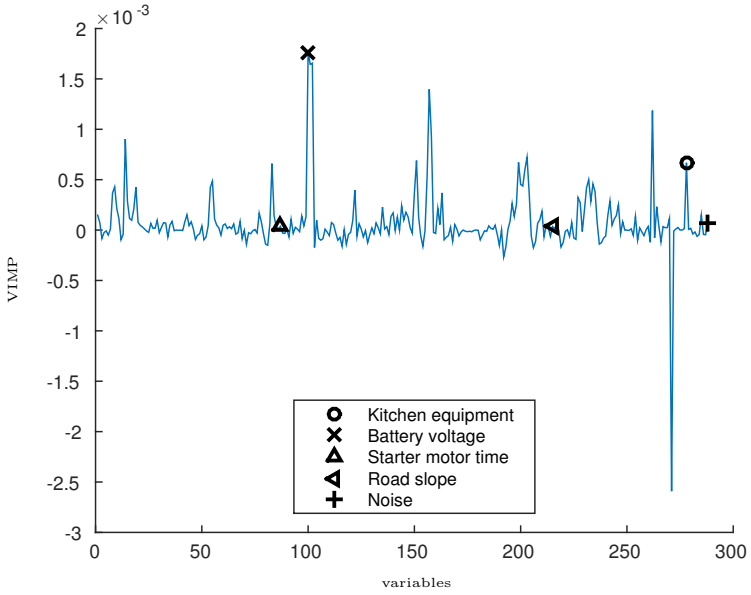


Figure 3: VIMP values for all variables from vehicle database where x axis corresponds to variables from vehicle database, y axis shows VIMP value for a particular variable. Large positive values of VIMP corresponds to important variables.

5 Measure for variable selection

Due to the limitations using the VIMP and minimal depth measures, as discussed in the previous section, a new measure of variable importance is proposed. The principle of the proposed measure is similar to minimal depth but considers not the mean of the first appearance of a variable in a tree, but that the probability that a splitting variable is used varies with different levels of the tree. An important variable should be used more often as a splitting variable at lower tree levels, close to the root, and less at higher tree levels. If noisy variables are selected as splitting variables the probability should be low for low tree levels and not change as much between different tree levels, maybe increase slightly for higher levels. Fig. 5 illustrates the qualitative shapes of the probability distributions with respect to the tree levels for important and noisy variables. Thus, main idea of the new variable importance measure is to evaluate for a given splitting variable the probability that it is used at different levels of the trees in the RSF.

Let $d = 1, 2, \ldots, \max(D(\mathcal{T}))$, where $D(\mathcal{T})$ is a tree depth, be all possible tree levels in a RSF and $v \in \nu$ is a splitting variable. Consider d as a random variable, and define $P(v, d)$ which describes the joint probability that v is selected as a

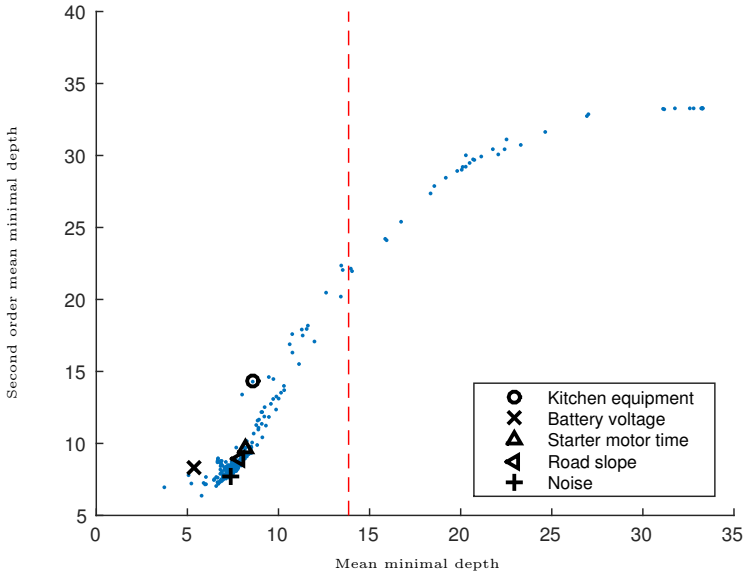


Figure 4: Minimal depth approach applied to vehicle database where x axis corresponds to mean value of the first appearance of a variable in a tree, y axis corresponds to mean value of the second appearance of a variable in a tree. Dashed red line is a threshold that separates important and non-important variables. Important ones should lay to the left from the threshold.

splitting variable in a node at a tree level d. Then,

$$P(d|v) = \frac{P(v|d)P(d)}{P(v)} \tag{8}$$

where, $P(v|d)$ denotes the conditional probability that v is selected as a splitting variable in a node given tree level d. The probability $P(d)$ is an a priori probability to select a specific level in the tree, independent of splitting variable, and $P(v)$ is the marginal probability of selecting v as a splitting variable for the whole tree. It is assumed that there is no a priori knowledge of $P(d)$, thus, the probability is set equal for all levels, i.e., $P(d) = \frac{1}{\max(D(T))}$, $\forall d$. The conditional probability $P(d|v)$ can be interpreted as the a posteriori probability of selecting a tree level given that v is used as a splitting variable. The a posteriori distribution (8) is here considered a relevant measure of the importance of the splitting variable v in the RSF. The measure avoids the problem, for example, VIMP has where the importance will be shared between the correlated variables. This is because (8) consider the probability of selecting different tree levels given that a splitting variable is selected and does not depend on the probability of selecting v which is reduced if variables are correlated.

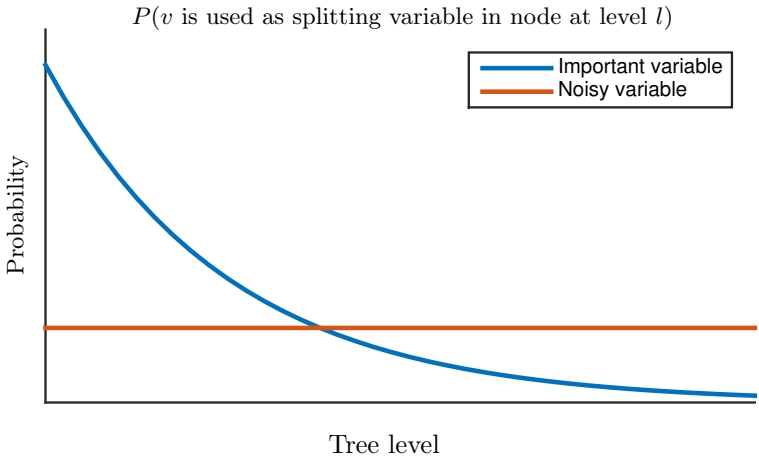


Figure 5: Illustrative example of the probability that a given splitting variable is used in a node at different tree levels.

The conditional probability (8) will be used as a variable importance measure. However, the true probability is not known because it depends on many different factors, for example, the parameters when generating the RSF. However, it can be estimated from the RSF by computing the mean ratio for all trees that v is used as a splitting variable in a node for each level d of the tree. This, can be done by first computing

$$\phi_v(d) = \frac{\sum_{\mathcal{T}} \frac{l_{d,v}}{l_d}}{\#\text{ of trees in RSF}}$$

where $l_{d,v}$ is number of nodes at level d where v is splitting variable. Equation (9) is then used to compute the estimate

$$P_v(d) = \frac{\phi_v(d)}{\sum_k \phi_v(k)}. \tag{9}$$

which will be used when analyzing the RSF.

Generating an RSF for identifying important variables differs from generating an RSF for battery life prediction. To identify important variables it is useful to generate the RSF such that the chance of having significant variations between variables is increased. Thus, each tree in the forest is allowed to grow deep to have as many levels and branches as possible. Therefore, the minimal terminal node size was chosen to be two. This parameter choice is not suitable for battery life prediction where instead a minimal terminal node size of 200 was used. In the later case, the focus is in quality of prediction and taking into account the fact that there are no time series for each vehicle, it should be associated with a class of vehicles with similar usage profile. Therefore, small values of

minimal node size could be a bad choice. However for variable selection, trees are required to be as deep as possible, because quality of (9) depends on it. Based on experience, but also to compare the results with the minimal depth approach, 1000 trees was selected to be generated in the RSF.

After growing an RSF with minimal terminal node size 2 and calculating probability mass functions (pmf) according to (9), five probability mass functions for different splitting variables are shown in Fig. 6. The variables stating whether a vehicle has kitchen equipment or not and how long the battery has had a voltage within a given interval are intuitively important since they indicate how the battery is used. This is visible in the figure since $P_v(d)$ is large for small d and decreasing with increasing d, while the noise variable is more flat starting from level 5. The estimated $P_v(d)$ with respect to the road slope where the vehicle has been run has the same shape as the noise indicating that it is not important regarding battery degradation. The estimate $P_v(d)$ of the starting motor time does not have the same shape as kitchen equipment but there is still more likely that it is used as a splitting variable close to the root in a tree indicating that it still has some importance. This is reasonable since a degraded battery can be correlated with that the starting motor is used more. These observations indicate that the shape of $P_v(d)$ can be used to measure variable importance.

6 Identifying important variables for battery failure prognostics

The proposed variable importance measure estimate (9) is here used to analyze data from the vehicle database. There is a number of different histogram variables describing how the vehicle is used where each bin represents how much a variable has been measured within a specific interval. As a first step, the analysis will evaluate if it is possible to identify important operating regions and vehicle configurations which are correlated with battery degradation. The results are discussed based on expert knowledge for variables related to battery voltage, fuel consumption, starter motor usage, ambient temperature, and configuration variables. Then, in a second step an automatic procedure is outlined and applied to the full set of variables.

Based on the observations in Fig. 6, the shape of $P_v(d)$ for high d is more noisy due to varying sizes of the different trees. Thus, for better visualization (9) is plotted for each histogram variable but only for tree levels $d \leq 20$. Fig. 7 upper plot shows $P_v(d)$ for different histogram bins of battery voltage variable when the battery is used, where bin 1 represents low battery voltage and bin 9 high battery voltage. The three lowest histogram bins have higher values at lower tree levels indicating that the time the battery in the truck is having low voltage is important for battery health prediction. It is also visible that the bin 9 significantly higher at lower tree levels, compared to the bins 4-8, meaning that

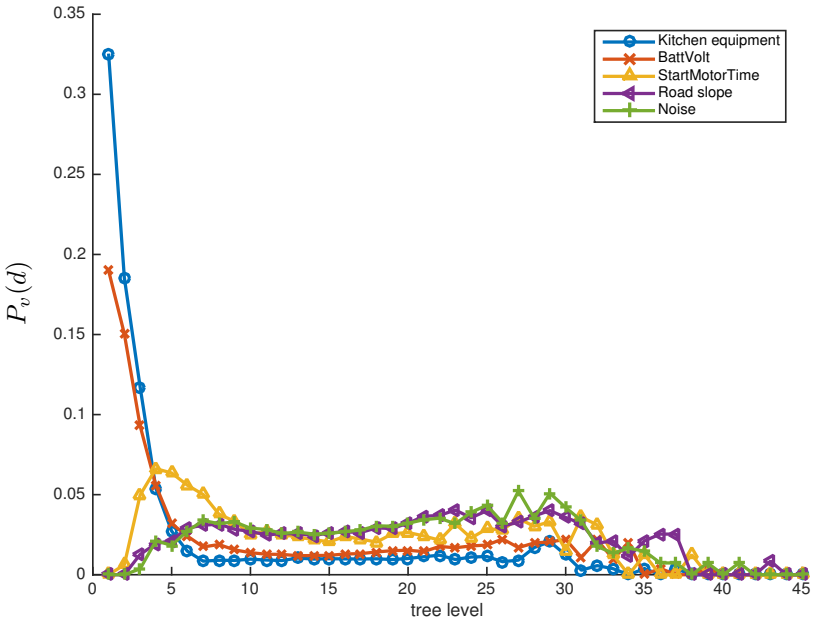


Figure 6: Probability mass functions $P_v(d)$ for 5 variables from vehicle database calculated according to (9).

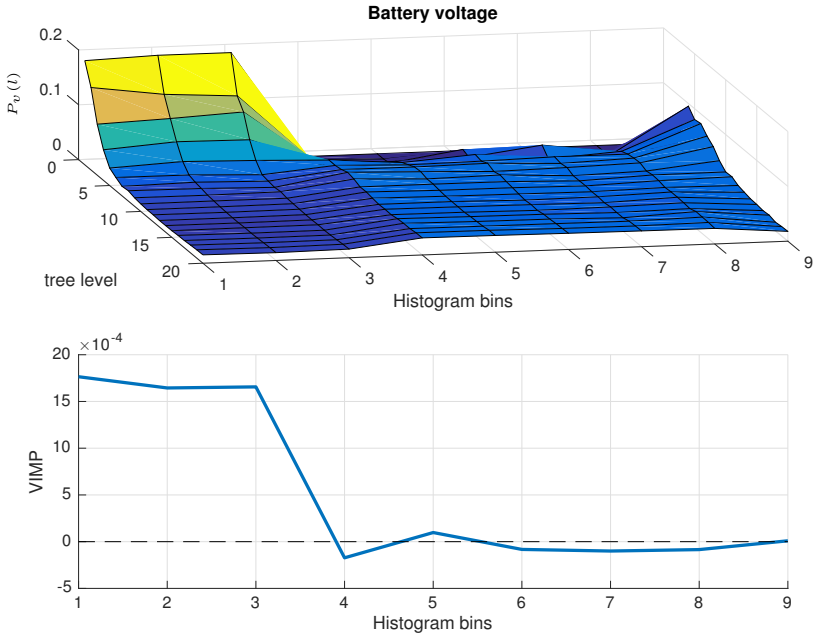


Figure 7: Variable importance analysis of battery voltage histogram variables.

high voltages are also relevant for battery health prognostics. When comparing the result to VIMP, Fig. 7 lower plot, it is visible that both methods identifies low voltages as important, high positive values of VIMP mean important variables, but the high voltage is not identified by VIMP.

Another variable is fuel consumption speed which is shown in Fig. 8. It is visible from upper plot that mainly lower fuel consumption speeds are correlated to battery failure which could be related to city driving with lots of starts and stops increasing the usage of the battery. VIMP, lower plot, varies more and it is more difficult to identify any bin as more important.

The analysis of the time that the starter motor is used is shown in Fig. 9. Compared to noisy variables it is indicated that the starter motor time has some relevance, see Fig. 6 upper plot, for the whole interval, but not as much as, for example, low battery voltage. Also, note that there is a small trend of increasing importance with increasing starter motor time indicating that battery failure is correlated with when the starter motor is used more often which is reasonable since the battery is used more. The computed VIMP measure, lower plot, does not have any clear indication of any bin being important, except possibly bin 6.

It is known that cold temperatures are not good for battery health which is also visible in Fig. 10. It is mainly the lowest temperature bin that is relevant

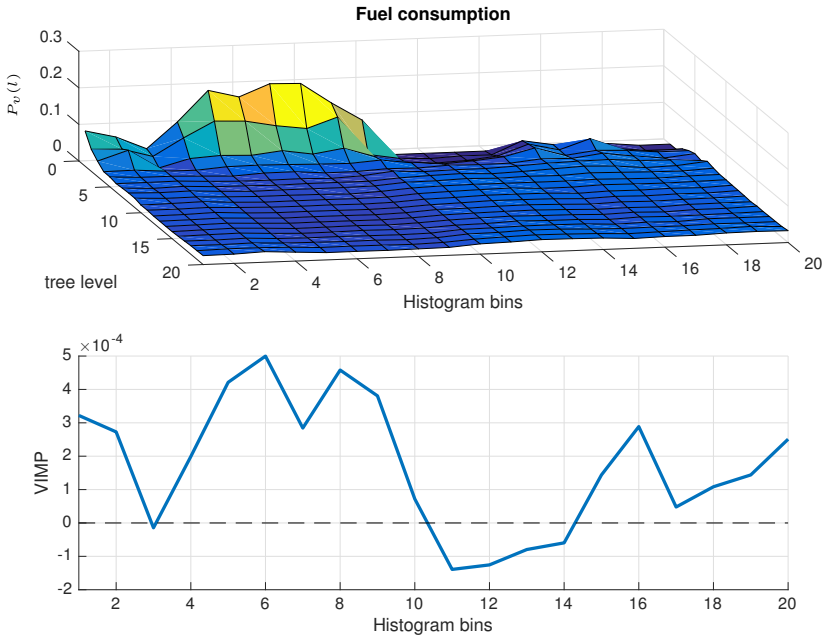


Figure 8: Variable importance analysis of fuel consumption speed histogram variables.

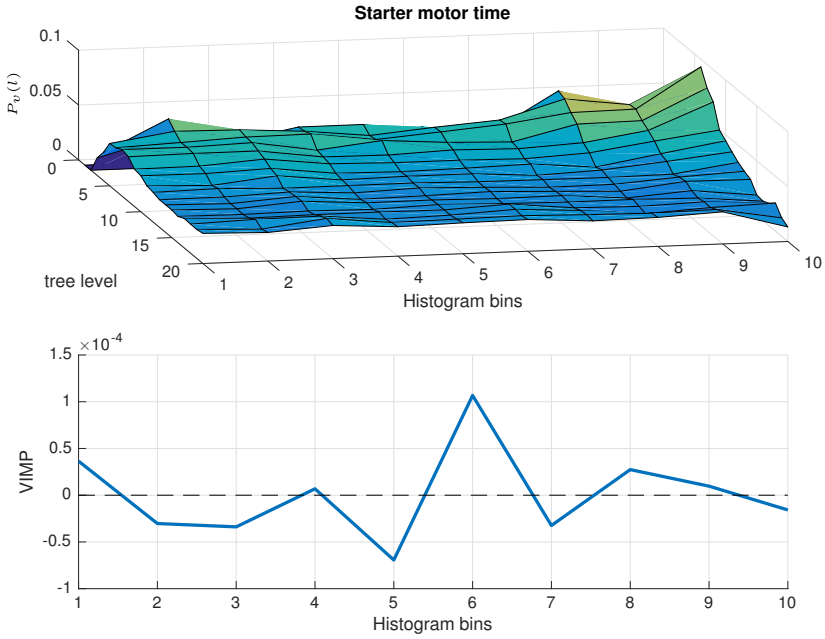


Figure 9: Variable importance analysis of starter motor time histogram variables.

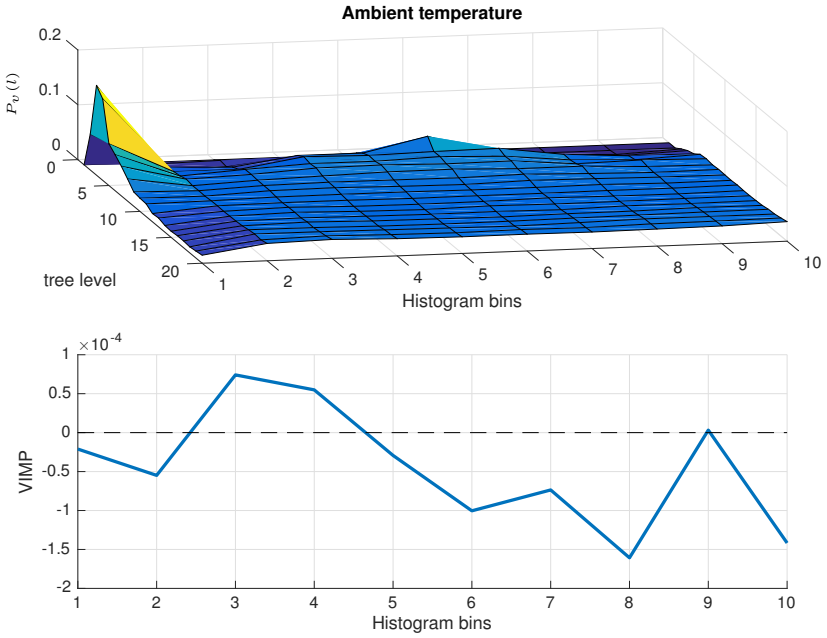


Figure 10: Variable importance analysis of ambient temperature histogram variables.

for battery degradation. When comparing the results with VIMP, the VIMP measure is very close to 0 and even negative in many cases.

Finally, a set of variables describing the vehicle configuration is analyzed in Fig. 11. The variables consider both battery type and position and variables that are related to if the driver sleeps in the truck, for example, if there are any kitchen equipment or beds and thereby use the battery for more purposes than starting the combustion engine. The figure shows that if there is kitchen equipment or not and, the battery position, and if there are any beds in the truck are important variables. This result is understandable since if the driver is using the truck to sleep in it and cook food, the battery will be used, not only for starting the engine but also for powering these auxiliary units. The battery position indicates that some battery positions are correlated to faster battery degradation, for example, increased vibrations. VIMP identifies the kitchen equipment variable but there seems to be no significant importance for the other configuration parameters.

These examples show that the results from (9) can be explained from expert knowledge. Further, the examples indicate that the measure is useful for identifying variables relevant for battery failure prognostics and extracts information

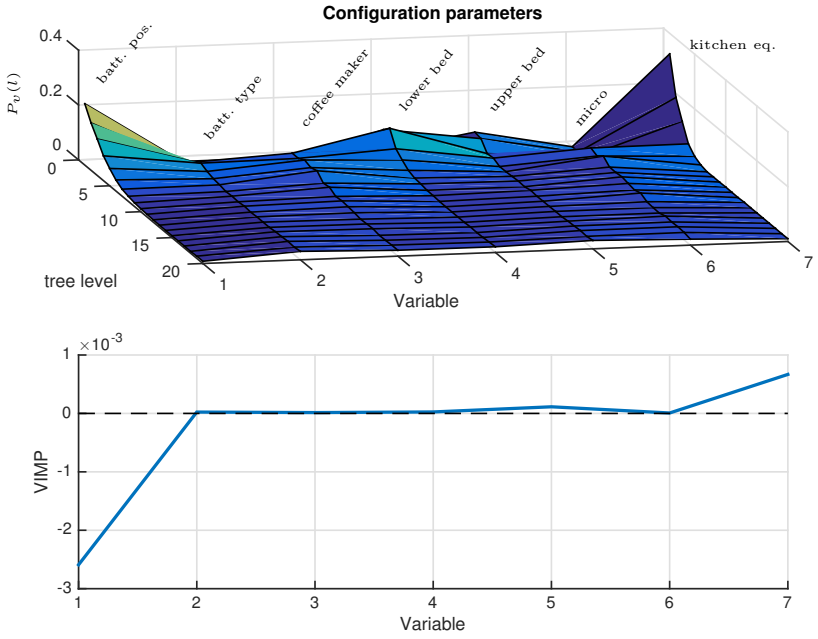


Figure 11: Variable importance of different truck configuration variables.

not obtained from VIMP or minimal depth metrics.

However, the analysis here is performed manually. To automatically select important variables for generating an RSF model, it must be possible to measure the variable importance based on the shape of (9).

6.1 Variable selection using shape of depth distribution

To select a suitable set of variables, the variable importance is measured by computing the skewness and mean of $P_v(d)$ in (9) for each variable v,

$$\begin{aligned} \mu_d &= \mathrm{E}_{P_v}[d] && \text{(mean)} \\ \gamma_d &= \mathrm{E}_{P_v}\left[\left(\frac{d-\mu_d}{\sigma_d}\right)^3\right] && \text{(skewness)} \end{aligned} \tag{10}$$

where σ_d is the standard deviation of d. Fig. 12 shows plotted mean and skewness of $P_v(d)$ for each variable v from vehicle database. Important variables should have a large positive values of skewness and a low mean value, i.e., they should be in the lower right corner of the figure, while noisy variables should be in the upper left corner. Most of the variables are located in the upper left corner, including the injected noisy variables, indicating that many variables are not relevant for battery degradation. However, there is a set of points that are located along the way down to the right indicating their increasing importance.

The corresponding skewness and mean for each of the variables in Fig. 6 are marked in Fig. 12 showing that the variables thought to be important are located down to the right while the noise variable is located up to the left.

7 Evaluating RSF model for battery health prognosis

Based on Fig. 12, a manually selected threshold is defined to select a subset of variables that are most important to generate a new RSF. The performance of the RSF using the reduced set of variables is compared to using all variables. The selected subset of variables includes 50 variables out of 283 variables. For both sets of variables, an RSF is generated with 1000 trees and a minimal terminal node size of 200. The error rate for the case with all features is 0.2011, and for the reduced set 0.2186 which are comparable in size. Note that, as observed in Section 2.2, this does not necessarily mean similar predictor performance.

For the analysis, 10 vehicles with battery failures and 10 without are selected randomly. These vehicles are then used as inputs in the RSF to compute the life functions $\mathcal{B}^\nu(t;t_0)$ and the results are shown on Fig. 13 and Fig. 14 for vehicles with battery problems and healthy ones, respectively. It should be noted that time units were used on x axis for both figures, original time was scaled to

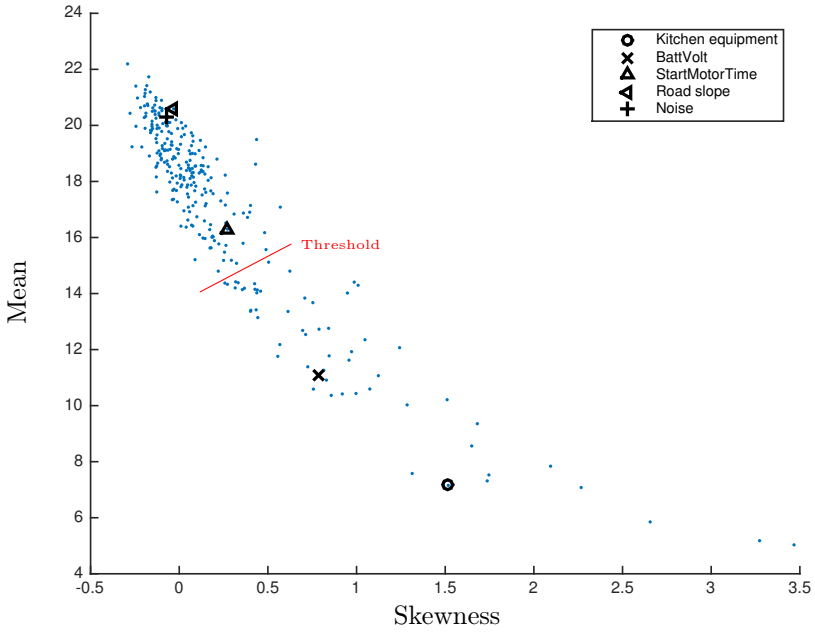
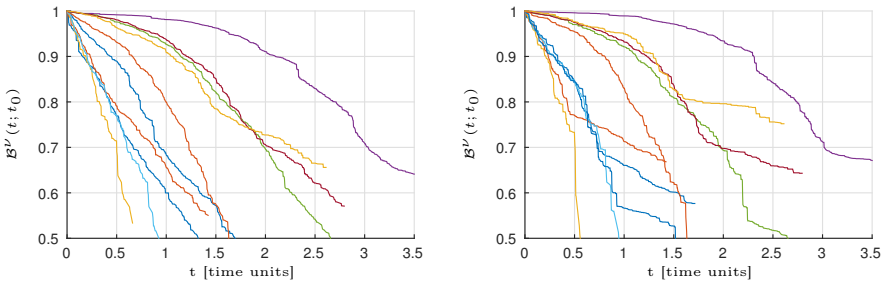


Figure 12: Skewness and mean of (8) plotted for each variable in vehicle database. A manually selected threshold is used to select a subset of important variables which reside down to the right from the threshold.

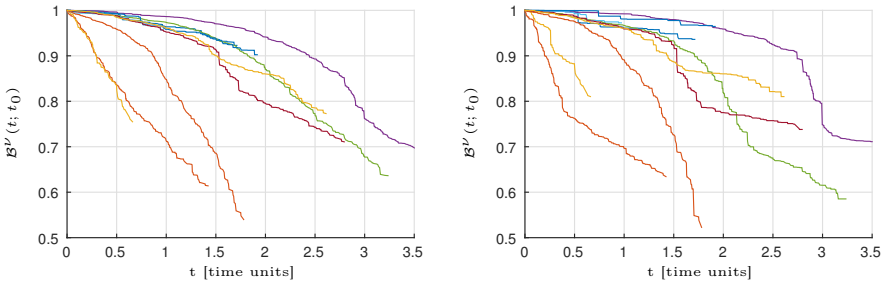
hide true life-time of batteries to not reveal sensitive information for industrial partner.



(a) Broken vehicles with all features (b) Broken vehicles with important features

Figure 13: Lifetime functions $\mathcal{B}^{\nu}(t;t_0)$ for 10 vehicles with battery failures from vehicle database. Two models of RSF compared, namely, with all features and with reduced set of features.

The computed lifetime functions have, in general, higher values for vehicles without battery problems than for vehicles with battery problems, see Fig. 13 and Fig. 14. This is true for the cases with and without feature selection which is expected. Another thing that can be noticed is that the lifetime functions are more less the same for the case with all variables and the case with only the identified important ones. It is difficult to evalaute the quality of the predictions of the two RSF models. However, the results in the example in Section 2 shows that a reduced number of noisy variables should have a positive impact on prediction accuracy even though the error rates are comparable.



(a) Healthy vehicles with all features (b) Healthy vehicles with important features

Figure 14: Lifetime functions $\mathcal{B}^{\nu}(t;t_0)$ for 10 censored vehicles from vehicle database. Two models of RSF compared, namely, with all features and with reduced set of features.

8 Conclusions

A heavy-duty truck battery failure prognosis model is estimated based on truck operational data using random survival forests. The available data have several complicating factors, such as, missing and censored data, varying variable types, etc., which can be handled using random survival forests. Applying variable selection before generating the battery failure prognosis model can help improve the prognosis, but also interpretability, and computational cost. Standard techniques for variables importance measures are evaluated. Since satisfactory performance was not achieved, a new variable importance measure is proposed to identify variables relevant for battery failure prognosis. The analysis is used both to identify which variables are relevant for battery lifetime prediction and to improve prediction performance. The results of the new approach are consistent with expert knowledge, for example, identifying low ambient temperatures and if the driver uses kitchen equipment in the truck as important information. The performance of the proposed variable importance measure promising for this application when compared to existing measures. Training an RSF for the two cases, using all variables and only 18% of important ones, result is comparable in error rates. The introductory example shows that similar error rates still give varying results compared to the truth which indicates that the proposed variable selection method should improve prediction performance. However, more work should be done in this direction to justify the results.

Acknowledgment

The authors acknowledge Scania and VINNOVA (Swedish Governmental Agency for Innovation Systems) for sponsorship of this work.

References

L. Breiman. Random forests. *Machine learning*, 45(1):5–32, 2001.

D. R. Cox and D. Oakes. *Analysis of survival data*, volume 21. CRC Press, 1984.

M. Daigle and K. Goebel. A model-based prognostics approach applied to pneumatic valves. *International Journal of Prognostics and Health Management Volume 2 (color)*, page 84, 2011.

E. Frisk and M. Krysander. Treatment of accumulative variables in data-driven prognostics of lead-acid batteries. In *Proceedings of IFAC Safeprocess'15*, Paris, France, 2015.

E. Frisk, M. Krysander, and E. Larsson. Data-driven lead-acide battery prognostics using random survival forests. In *Proceedings of the Annual Conference of The Prognostics and Health Management Society*, Fort Worth, Texas, USA, 2014.

I. Guyon and A. Elisseeff. An introduction to variable and feature selection. *The Journal of Machine Learning Research*, 3:1157–1182, 2003.

H. Ishwaran and U. Kogalur. Random survival forests for r. *Rnews*, 7/2:25–31, 2007.

H. Ishwaran, U. Kogalur, E. Blackstone, and M. Lauer. Random survival forests. *The Annals of Applied Statistics*, pages 841–860, 2008.

H. Ishwaran, U. Kogalur, E. Gorodeski, A. Minn, and M. Lauer. High-dimensional variable selection for survival data. *Journal of the American Statistical Association*, 105(489):205–217, 2010.

H. Ishwaran, U. Kogalur, X. Chen, and A. Minn. Random survival forests for high-dimensional data. *Statistical Analysis and Data Mining: The ASA Data Science Journal*, 4(1):115–132, 2011.

X. Si, W. Wang, C. Hu, and D. Zhou. Remaining useful life estimation–a review on the statistical data driven approaches. *European Journal of Operational Research*, 213(1):1–14, 2011.

Paper B

Variable selection for heavy-duty vehicle battery failure prognostics using random survival forests*

B

*Published in Proceedings of *European Conference of the PHM Society, Bilbao, Spain, 2016.*

Variable selection for heavy-duty vehicle battery failure prognostics using random survival forests

Sergii Voronov, Daniel Jung, and Erik Frisk

Vehicular Systems, Department of Electrical Engineering, Linköping University, SE-581 83 Linköping, Sweden.

Abstract

Prognostics and health management is a useful tool for more flexible maintenance planning and increased system reliability. The application in this study is lead-acid battery failure prognosis for heavy-duty trucks which is important to avoid unplanned stops by the road. There are large amounts of data available, logged from trucks in operation. However, data is not closely related to battery health which makes battery prognostic challenging. When developing a data-driven prognostics model and the number of available variables is large, variable selection is an important task, since including non-informative variables in the model have a negative impact on prognosis performance. Two features of the dataset has been identified, 1) few informative variables, and 2) highly correlated variables in the dataset. The main contribution is a novel method for identifying important variables, taking these two properties into account, using Random Survival Forests to estimate prognostics models. The result of the proposed method is compared to existing variable selection methods, and applied to a real-world automotive dataset. Prognostic models with all and reduced set of variables are generated and differences between the model predictions are discussed, and favorable properties of the proposed approach are highlighted.

1 Introduction

Prognostics and health management are important parts to prevent unexpected failures by more flexible maintenance planning. The purpose is to replace a failing component before it fails, but avoid changing it too often. Coarsely, there are two main approaches in prognostics, data-driven and model-based techniques, but also hybrid approaches that combine the two are possible. Model-based prognostics uses a model of the monitored system and the fault to monitor to predict the degradation rate and Remaining Useful Life (RUL), see for example (Daigle and Goebel, 2011). Statistical data-driven methods (Si et al., 2011) generate a prediction model based on training data to predict RUL.

One relevant application is lead-acid starter battery prognosis for heavy-duty trucks. Heavy-duty trucks are important for transporting goods, working at mines, or construction sites, and it is vital that vehicles have a high degree of availability. Unplanned stops by the road can result in increased cost for the company due to the delay in delivery, but can also lead to damaged cargo. One cause of unplanned stops is a failure in the electrical power system, and in particular the lead-acid starter battery. The main purpose of the battery is to power the starter motor to get the diesel engine running, but it is also used to, for example, power auxiliary units such as heating and kitchen equipment.

The main contribution in this work is a data-driven method for variable selection when estimating a battery failure prognostics model for automotive lead-acid batteries based on Random Survival Forests (Ishwaran et al., 2008). In particular, two key properties of the application data set are addressed 1) the number of informative variables is assumed to be small, and 2) the data contains highly correlated variables. Both aspects make building a prognostics model more difficult and are the main motivating factors for the proposed approach. Further, variable selection is also important to better understand which factors that are correlated with battery failure rate and also what is causing it. This work is a continuation of (Voronov et al., 2016), where the main focus was to analyze the automotive application case study. Here, the main contribution is an extended analysis of the variable selection problem that results in an augmentation of the decision space with an extra dimension. Further, characteristics of existing variable selection methods for Random Survival Forests are analyzed and compared to the proposed method, in particular for the case where there are many correlated variables in the data set. In addition, a basic variable selection methodology is proposed.

2 Problem formulation

The main objective in this work is to use Random Survival Forests (RSF) (Ishwaran et al., 2008) to identify, from data, which variables are relevant for building RSF models for survival analysis. The problem of identifying important variables is usually referred to as variable selection and is a relevant topic in

data-driven prognostics and machine learning in general (Guyon and Elisseeff, 2003).

The prognostic problem studied here is to estimate the battery lifetime prediction function based on recorded vehicle data. The lifetime prediction function is defined as

$$\mathcal{B}^{\mathcal{V}}(t;t_0) = P(T > t + t_0 \mid T \geq t_0,\ \mathcal{V}) \tag{1}$$

where T is the random variable failure time of the battery and $\mathcal{V}$ the vehicle data at time $t = t_0$ when data is submitted into the model, in our case when a vehicle comes to the workshop. The function $\mathcal{B}^{\mathcal{V}}(t;t_0)$ is a function of t and gives the probability that the battery will function at least t time units after t_0. The data $\mathcal{V}$ is recorded operational data for a specific vehicle.

2.1 Operational data

In this work a vehicle fleet database is provided by an industrial partner, where one snapshot of data is available from each vehicle including information regarding how the truck has been used and the configuration of the specific truck. There is also information if the battery has failed or not. The database contains lots of information from the truck, not always related to battery degradation, meaning that it is not known what available information is relevant for this specific task. Therefore, it is relevant to identify which variables are relevant for battery lifetime prediction. Previous works considering this vehicle data set are presented in (Frisk and Krysander, 2015) and (Frisk et al., 2014).

The main characteristics of the database can be summarized as follows:

- 33603 vehicles from 5 EU markets
- A single snapshot per vehicle
- 284 variables stored for each vehicle snapshot
- Heterogeneous data, i.e., it is a mixture of categorical and numerical data
- Availability of histogram variables
- Censoring rate more than 90 percent
- Significant missing data rate

A main characteristic of the database is that there are no time series available for a vehicle. It means that there is only one snapshot $\mathcal{V}$ of the variables in the database from each vehicle. Information describing how the vehicle has been used is stored as histogram data representing how often specific sensor data is measured within different intervals. As an example, there is a histogram describing how much time the vehicle has been subjected to different ambient temperatures.

Due to the non-specific purpose of the database, it is probable that only a small number of variables from set $\mathcal{V}$ influence prediction of the battery failure rate. Thus, identifying the important variables in order to remove irrelevant variables, should improve the performance of a battery prognosis model.

2.2 Motivation for variable selection

There are several reasons why variable selection is important when working with data-driven models. First, it is possible to improve prediction performance by reducing the number of variables. The second motivation is better interpretability of the results by clearly understanding which factors are important for battery failure. The third motivation is to reduce model generation and prediction time by reducing the number of variables used for generating the RSF.

An example why the quality of predictor may become bad if the number of noisy (non-important) variables is significantly large is given below. Synthetic data is created with the following properties. Let h_0 be a constant nominal hazard rate (Cox and Oakes, 1984) for battery failures. The hazard rate

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

represents the probability of a battery failure at a particular time t. In this example, the hazard rate does not change with time and the nominal hazard rate corresponds to an expected 10 years of battery life. It is assumed that there is one variable v_1 with an impact on battery hazard rate h as

$$h = \begin{cases} 1 \cdot h_0, & \text{if } v_1 = 1 \\ 2 \cdot h_0, & \text{if } v_1 = 2 \\ 3 \cdot h_0, & \text{if } v_1 = 3 \end{cases} \tag{3}$$

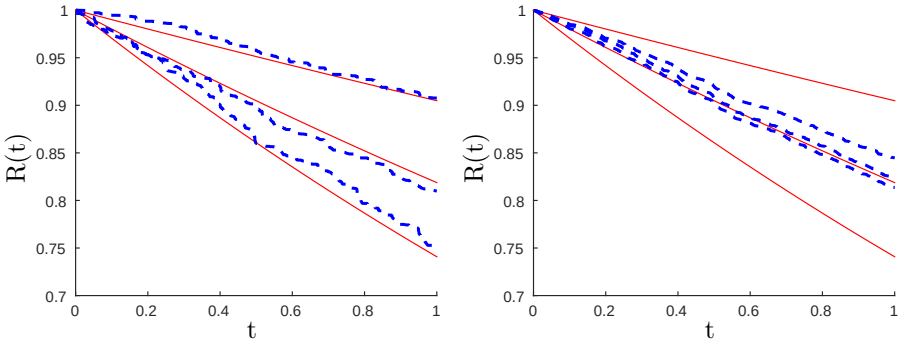
where h_0 is the nominal hazard rate. Data for 3000 vehicles is generated with a censoring rate about 80 percent. Different numbers of noisy variables are included in the synthetic data to observe how they change the RSF output.

First, only two noisy variables are added in addition to v_1. In the second case, 100 noisy variables are added. All noisy variables are sampled from a normal distribution with zero mean and unity variance. After generating two RSF models, one for each set of variables, the reliability functions (Cox and Oakes, 1984)

$$R(t) = P(T \geq t) \tag{4}$$

computed by the two RSF models are compared with the theoretical values of the reliability as shown in Figure 1. One vehicle from each of the three classes was chosen and submitted to the forest to receive the predictions. It is shown in Figure 1 (a) that predictions from RSF for the case of 2 noisy variables, dashed blue curves, are following the theoretical reliability functions, red solid curves, better than the case with 100 noisy variables, see Figure 1 (b). However, the

error rate, which is a common performance measure for the RSF, is similar for both cases. This means that the error rate is not a good measure in prognostic terms. It is worth to notice that in the simulation environment, information about the true reliability curves is available. However, this is not the case for the vehicle fleet database.



(a) 2 noisy variables (b) 100 noisy variables

Figure 1: Predictions from RSF with different number of noisy variables.

The example motivates the relevance of finding the important variables and at the same time removing noisy ones, especially if number of important is small, in a set of data as expected in the vehicle fleet database. The quality of the estimated reliability function from the RSF is significantly improved when the noisy variables are removed.

3 Random survival forests

A brief description of Random Survival Forests and two standard methods for evaluating variable importance are presented. For a more detailed description, the interested reader is referred to, for example, (Ishwaran et al., 2008) and (Ishwaran et al., 2011).

The difference between an ordinary decision tree classifier and a random forest is that there is randomness of two kinds injected into the process of estimating the model. The first source is the usage of a bootstrap procedure. Each tree is grown using its own bag of cases which are sampled from the training set. Second, for each node in a tree, splitting variables are selected from a randomly sampled subset. RSF extends the RF approach to right-censored survival data, i.e., objects in the study without experienced failure.

RSF is a data-driven method that can be used for computing maximum-likelihood estimates of the reliability function (4). It can be used to rewrite the

lifetime prediction function (1) as

$$\mathcal{B}^{\mathcal{V}}(t;t_0) = P(T > t + t_0 \mid T \geq t_0,\ \mathcal{V}) = \frac{R^{\mathcal{V}}(t+t_0)}{R^{\mathcal{V}}(t_0)} \tag{5}$$

The output from each tree $\mathcal{T}$ in the RSF is the Nelson-Aalen estimate of the cumulative hazard rate, see (Cox and Oakes, 1984). Let $t_1^{\mathcal{T}} < t_2^{\mathcal{T}} < \ldots < t_N^{\mathcal{T}}$ be N distinct event times when failures of objects under study occur. Then, the Nelson-Aalen estimate for tree $\mathcal{T}$ and vehicle (data) $\mathcal{V}$ is

$$\hat{H}_{\mathcal{T}}(t|\mathcal{V}) = \sum_{t_j^{\mathcal{T}} \leq t} \frac{f_{j,n_i}}{s_{j,n_i}} \tag{6}$$

where f_{j,n_i} and s_{j,n_i} are number of failures and survived objects in terminal node n_i of a tree $\mathcal{T}$ at event time $t_j^{\mathcal{T}}$ respectively. Terminal node n_i is determined by dropping vehicle $\mathcal{V}$ down through the forest. The cumulative hazard estimate $\hat{H}(t|\mathcal{V})$ for the whole forest is received by averaging over all $\hat{H}_{\mathcal{T}}(t|\mathcal{V})$. Finally, the reliability function $R^{\mathcal{V}}(t)$ from (5) is obtained from the fact (Cox and Oakes, 1984)

$$R^{\mathcal{V}}(t) = e^{-\hat{H}(t|\mathcal{V})} \tag{7}$$

and then $\mathcal{B}^{\mathcal{V}}(t;t_0)$ can be computed from (5).

One measure of prediction error of RSF models proposed in (Ishwaran et al., 2008) is based on pair-wise evaluation of non-censored data, called concordance index (Harrell et al., 1982). In short, the measure takes into consideration if the RSF model correctly predicts which of the two samples that will fail first. However, note that it does not take into consideration how accurate the prediction is with respect to the actual failure time. Therefore, the error rates of the two models in Figure 1 turn out to be more or less equal even though the model with fewer variables is visibly more accurate.

3.1 Variable selection using VIMP

One intuitive measure of variable importance is to measure the increase in prediction error when ignoring a variable in the RSF. This is done by randomizing the sample variable value when used as a splitting variable in the forest (Ishwaran et al., 2008). The idea is that a large increase in prediction error indicates that a variable is important while a low increase (or a decrease) indicates that the variable is not important. This variable importance method is called VIMP and is a candidate tool for variable selection by selecting a subset of the variables with the highest VIMP values. However, previous works, for example (Ishwaran et al., 2011), have shown that VIMP can have problems when there are many correlated variables. If several important variables are correlated they will share importance and VIMP will be low even if the variables are important. Thus, there is a risk that important variables will be lost and result in degraded prediction performance.

3.2 Variable selection using Minimal depth

As an alternative to VIMP, a candidate measure called minimal depth for variable selection in RSF has been proposed, see (Ishwaran et al., 2011) or (Ishwaran et al., 2010). The minimal depth for variable v is defined as the average distance from the root to the closest node where it appears in the RSF. Important variables should have a higher probability to be selected as splitting variables, compared to noisy variables, at low levels close to the root when the trees are generated. Thus, the minimal depth for important variables in the forest should be lower compared to noisy variables. To identify important variables using minimal depth, a threshold that distinguishes important variables from noisy variables is derived in (Ishwaran et al., 2011) based on the distribution for minimal depth D_v of noisy variables as

$$P(D_v = d \mid v \text{ is noisy variable}) = \\ = \left(1 - \frac{1}{p}\right)^{L_d} \left[1 - \left(1 - \frac{1}{p}\right)^{l_d}\right], \; 0 \leq d \leq D(T) - 1 \quad (8)$$

where $D(\mathcal{T})$ is the tree depth, l_d is number of nodes at depth d, $L_d = l_0 + l_1 + \ldots + l_{d-1}$ and p is number of candidate variables chosen from when generating the splitting rule in a node. The threshold can be selected as the mean value for the variable distribution (8). If the minimal depth measure of a variable mean value is less than the threshold, it is treated as important, otherwise as noise. The minimal depth measure is evaluated in (Ishwaran et al., 2011) and (Ishwaran et al., 2010) where it is shown to be successful for finding important variables in problems with few important variables and large number of noisy ones, even when the data samples are relatively small.

4 Variable depth distribution method

VIMP and minimal depth are the standard methods for variable selection in RSF models. However, there are problems connected with them. If many correlated variables are present in the database, as expected in our case, variables share VIMP between each other and it could happen that important variables will be lost if a VIMP based variable selection procedure is applied. The second reason why VIMP can have problems is that it is associated with error rate. As illustrated in Section 2, low error rate does not always correspond to good prognostic performance. Minimal depth does not depend on error rate and the variable selection approach has shown good results when applied to different databases in medical applications, see (Ishwaran et al., 2011) and (Ishwaran et al., 2010). However, it will be shown later that it did not work well when applied to the vehicle database. Taking into account aforementioned reasons, a new method for variable selection called Variable Depth Distribution (VDD) is proposed.

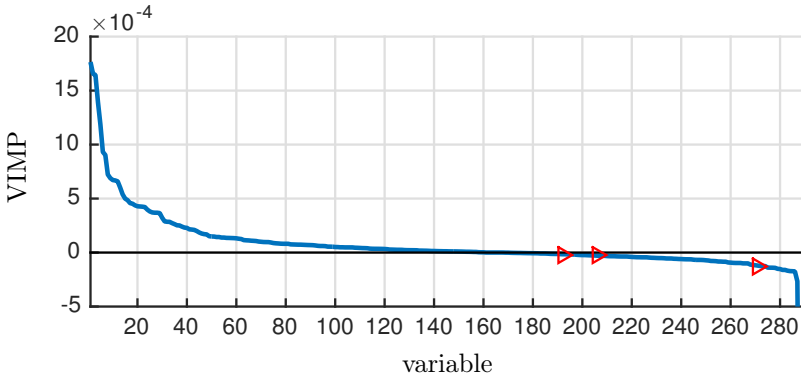


Figure 2: VIMP of variables in vehicle database sorted in ascending order.

The VIMP and minimal depth measures are applied to the vehicle database and the results are shown in Figure 2 and Figure 3, respectively. As a reference, three variables, only containing Gaussian noise, are included in the data set. The computed VIMP is positive for half of the variables, but the VIMP curve starts to flatten out after the first 30 variables with highest VIMP indicating that approximately 10% of the variables are expected to be relevant for battery lifetime prediction. The result of the minimal depth measure is presented in Figure 3 where the x axis is a mean value of the first appearance of the variable in the forest, y axis is a mean value of the second appearance of the variable in the forest, and the red dashed line is the threshold computed based on (8). The figure shows that most variables are identified as important, including the added noisy variables. Since the noisy variables are identified as important, it is an indication that minimal depth is not a suitable method for the vehicle database.

Due to the limitations using the VIMP, as discussed above, and the evaluation of the minimal depth measure in Figure 3, a new measure of variable importance is proposed. The principle of the proposed measure is similar to minimal depth, but considers the probability of a splitting variable being used at different levels of a tree. An important variable should be used more often as a splitting variable at lower tree levels, close to the root, and less at higher tree levels as illustrated in Figure 4. If noisy variables are selected as splitting variables the probability should not change as much between different tree levels, maybe increase slightly for higher levels.

Let $d = 1, 2, \ldots, \max(D(\mathcal{T}))$, where $D(\mathcal{T})$ is the tree depth, be all possible tree levels in a RSF and $v \in \nu$ is a splitting variable. Consider two random events, namely, choosing at random level d in a tree and picking a variable v as splitting in a tree. First event is similar to the problem of drawing a one ball from the boxes of enumerated balls. First, define $P(v, d)$ which describes the joint probability that v is selected as a splitting variable in a node at a tree level

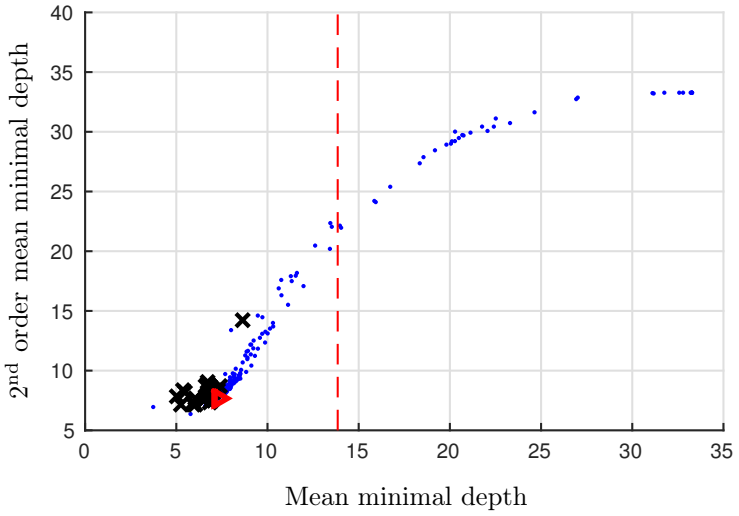


Figure 3: Minimal depth analysis of vehicle data. Black crosses correspond to 30 variables with highest VIMP and red triangles to added noise variables. Red dashed line is a threshold. Noisy variables should be located to the right of the red dashed line.

d. Then, according to Bayes rule

$$P(d|v) = \frac{P(v|d)P(d)}{P(v)} \tag{9}$$

where, $P(v|d)$ denotes the conditional probability that v is selected as a splitting variable in a node given tree level d. The probability $P(d)$ is a prior probability to select a specific level in a tree, independent of splitting variable, and $P(v)$ is the probability of selecting v as a splitting variable for the whole tree. It is assumed that there is no prior knowledge of $P(d)$, therefore, the probability is set equal for all levels, i.e., $P(d) = \frac{1}{\max(D(T))}$, $\forall d$. The conditional probability $P(d|v)$ can be interpreted as the a posterior probability of selecting a tree level given that v is used as a splitting variable. The posterior distribution (9) is here considered a relevant measure of the importance of the splitting variable v in the RSF. The measure avoids the problem that, for example, VIMP has where the importance will be shared between the correlated variables. This is because (9) considers the probability of selecting different tree levels conditioned that a splitting variable is selected and does not depend on the probability of selecting v which is reduced if variables are correlated.

The conditional probability (9) will be used as a variable importance measure. However, the true probability is not known because it depends on many different factors, for example, the parameters when generating the RSF. It can be noticed

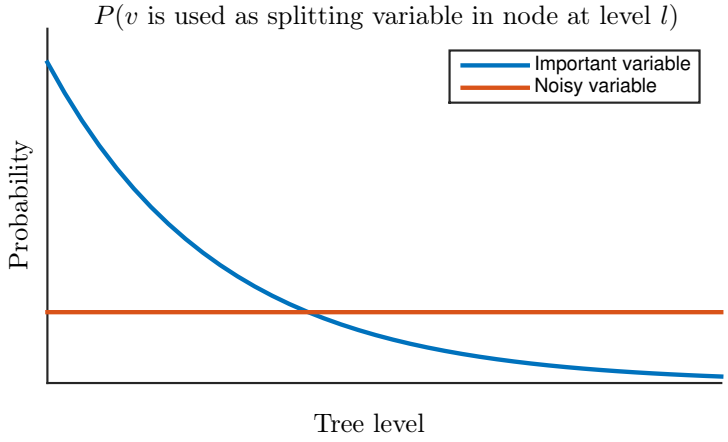


Figure 4: Illustrative example of the probability that a given splitting variable is used in a node at different tree levels.

from (9) that $P(d|v) \propto P(v|d)$ and if $P(v|d)$ is known the value $P(d|v)$ could be found as well. After growing the forest, $P(v|d)$ can be estimated by first computing

$$\phi_v(d) = \frac{\sum_{\mathcal{T}} \frac{l_{d,v}}{l_d}}{\#\text{ of trees in RSF}}$$

where $l_{d,v}$ is number of nodes at level d where v is splitting variable. Equation (10) is then used to compute the estimate

$$P_v(d) = \frac{\phi_v(d)}{\sum_k \phi_v(k)}. \tag{10}$$

which will be used when analyzing the RSF. An example of different distributions $P_v(d)$ are shown in Figure 5. Four variables from the vehicle data and one added noise variable are analyzed how they are used in a RSF generated from the vehicle fleet database. The distribution $P_v(d)$ of the noise variable is almost evenly distributed between levels 3 to 30, while variables related to battery usage, such as, if there is kitchen equipment in the truck and information about the battery voltage are significantly skewed to the left, indicating that these variables are important for prognostics of the battery health. The starter motor time variable has a higher probability mass at lower tree levels compared to the noisy variable but not as much as the kitchen equipment and battery voltage variables. The real data in Figure 5 resembles Figure 4 and the level of importance appears to increase with increased probability mass at lower tree levels.

Instead of comparing the whole distribution $P_v(d)$ for each variable v, two

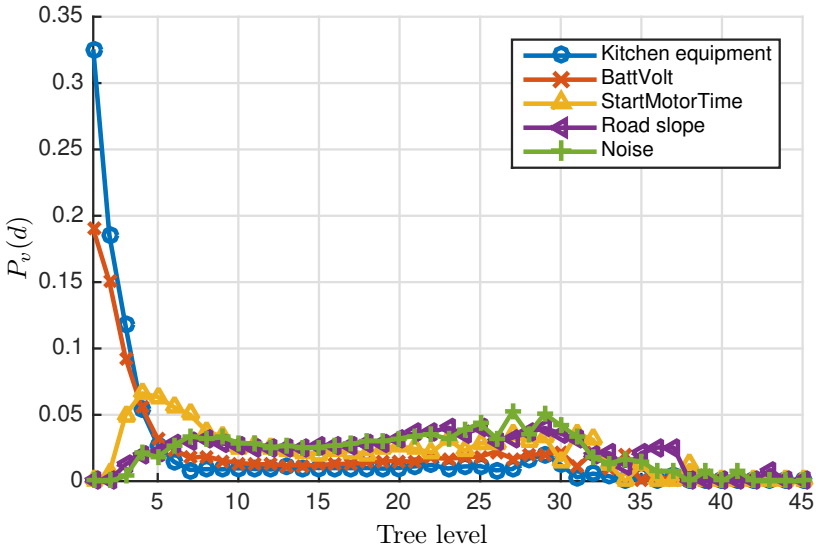


Figure 5: Examples of the estimated $P_v(d)$ for five different variables including one known noisy variable.

representative features are considered, mean and skewness,

$$\begin{aligned} \mu_d &= \mathrm{E}_{P_v}\left[d\right] && \text{(mean)} \\ \gamma_d &= \mathrm{E}_{P_v}\left[\left(\frac{d-\mu_d}{\sigma_d}\right)^3\right] && \text{(skewness)} \end{aligned} \tag{11}$$

According to Figure 4 and Figure 5, an important variable should have high positive value of skewness and low value of mean. These two features can be used alone to identify which variables that are important. There is one drawback with this approach, namely, it is possible that a noisy variable will be selected by random at low level of a tree. It means it will have values of skewness and mean as important variable. However, for a noisy variable, this is likely to be a rare event. Therefore, introducing information about how often a variable is used as a third dimension can help to filter out noisy variables in the area where important one should reside. Two possible candidates to express this information are:

- The probability that v is used as a splitting variable in each node $P(v)$.
- The probability that v is used as a splitting variable in a tree.

The first candidate can be estimated by counting the fraction of nodes a variable is used in a tree and taking the average over the whole forest. The second

feature only considers if a variable is used at all in a tree and can be estimated by counting the number of trees in the forest where a variable is used. As it is shown below, the third dimension, which take into account how often variable is selected, can help identify important variables more efficiently than if only mean and skewness is used as in (Voronov et al., 2016).

4.1 Real data case study

The result of applying the first candidate to the vehicle data as the third dimension together with mean and skewness (11) is shown in Figure 6 where each dot represents one variable. For comparison in the analysis, the 30 most important variables according to VIMP, are highlighted as black crosses and variables *rejected* by minimal depth are highlighted as green triangles pointing up. Also for the analysis the three added noisy variables are highlighted as red triangles pointing right.

Note that the 30 variables with highest VIMP have similar properties in Figure 6. They have low mean, high skewness, and are used in a relatively large fraction of the nodes. This can be interpreted as variables with high VIMP are used as splitting variables in many nodes close to the root of each tree. The noisy variables are also used in many of the nodes, but are located further away from the root node, thus having high mean and low skewness. There is also a number of variables with low mean and high skewness but are used in a smaller fraction of the nodes. Some of these variables are binary, meaning that they cannot be used as splitting variables more than once in a branch. Thus, they can be relevant for the problem but will not be used in many nodes. Note that the variables that are only used in a low fraction of nodes are variables rejected by minimal depth in Figure 3.

Comparing with the results using minimal depth in Figure 3 the results in Figure 6 looks promising because it is possible to find threshold to separate most of the important variables given by VIMP from noisy ones. Here, it is assumed that there are important variables among the 30 best given by VIMP, but it does not mean that all are important. The minimal depth method maps most of the variables below the threshold, including the known noisy variables, which indicates that it has difficulties with this data set. Note that Figure 6 clearly illustrates what properties are important in this case study according to VIMP and Minimal depth.

5 Analysis

Before continuing the analysis of the vehicle fleet data using the new variable selection method in the prognostic algorithm, the properties of the proposed measure in Section 4 are further analyzed. As mentioned in Section 4, there is no knowledge which variables are important in the vehicle database. There is an

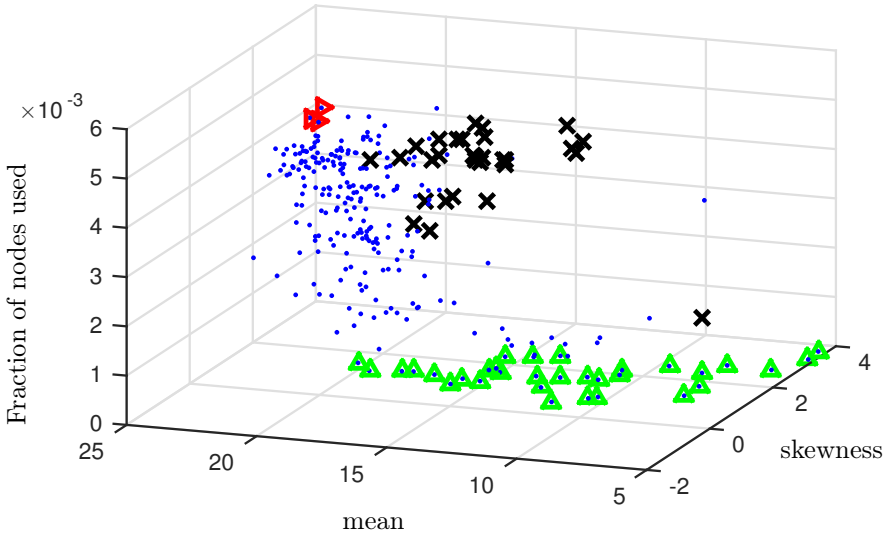


Figure 6: Skewness and mean of (10) of vehicle data combined with fraction of nodes. Black crosses correspond to 30 variables with highest VIMP, green triangles are variables rejected by minimal depth, and red triangles are added noise variables.

intuition that some of them could be informative, but it is not clear how many they are and what their influence is on the battery hazard rate.

In this section, two case studies are performed, namely, understanding the properties of the VDD method in a simulated environment and how to select important variables using an ad-hoc threshold based on simulations. First, a simple model is considered where only one important variable influences the life of the battery. Then, another example with a large number of correlated variables is considered. A third example using the simulated environment shows when the VDD method can be more advantageous than VIMP. Finally, the VDD method is applied to the vehicle fleet database where a set of important variables is selected using on the proposed methodology.

5.1 Case study in simulated environment

To analyze the properties of the measure discussed in Section 4, simulated battery failure data is generated which should resemble the general characteristics of the real vehicle database. Similar to the example from Section 2, it is assumed that the average battery lives for 10 years which is defined by a constant hazard rate h_0. One important variable v_1 changes the hazard rate h_0 by a factor h_1 defined as

$$h_1 = \begin{cases} 1, & \text{if } v_1 = 1 \\ 1.5, & \text{if } v_1 = 2 \\ 2.5, & \text{if } v_1 = 3 \\ 2.9, & \text{if } v_1 = 4 \\ 3.4, & \text{if } v_1 = 5 \end{cases} \tag{12}$$

Thus, the hazard rate for a randomly generated vehicle would be $h_1 \cdot h_0$. After generating hazard rates for all vehicles, simulated battery lifetimes are generated sampling from an exponential distribution with mean $\mu = \frac{1}{h_1 \cdot h_0}$. Censoring is done by sampling censored times from a gamma distribution, with shape parameter $k = \frac{1}{7 \cdot h_0}$ and scale parameter $\theta = 1$, and comparing achieved time values with failure ones. If the battery lifetime is less than the censored time the battery experienced failure, otherwise it is censored. The selected gamma distribution gives a censoring rate of approximately 80 percent which is similar to the vehicle database.

In the first example, data from 10 000 vehicles is generated and one hundred noisy variables are added to simulate non-important variables. Half of them are normally distributed with zero mean and unit variance and the other half are discrete uniformly distributed numbers from 1 to 10. The result of applying the proposed method is shown in Figure 7. The known important feature is highlighted as a black cross and noisy variables are shown as blue dots.

Figures 8 and 9 show the results for the same problem, but using VIMP and minimal depth respectively. Using VIMP, it is easy to identify the important variable, therefore, VIMP and the method proposed in the paper gives similar results in this case. In Figure 9, the red dashed line is the threshold that

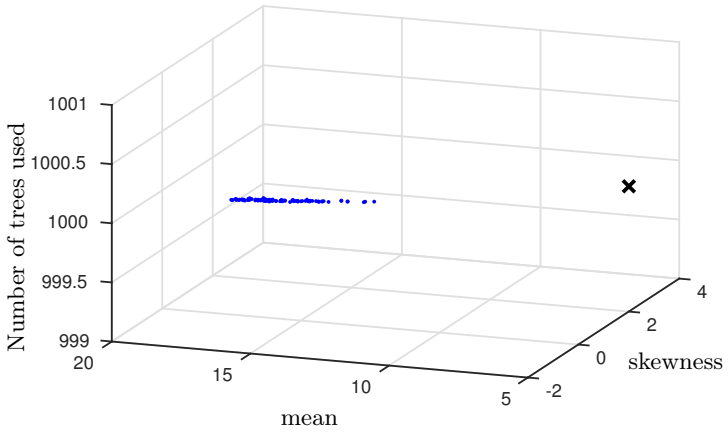


Figure 7: Simulated data from 10000 generated vehicles with censoring rate 80 percent. The important variable is marked with a cross.

separates important and non-important variables according to (Ishwaran et al., 2011). Variables to the left of the threshold should be important and variables to the right are not. Figure 9 shows that the specific threshold is not able to distinguish important variables in this case. However, it is visible that it is possible to manually select a threshold that could to that.

When using VIMP, correlated variables will share importance. Therefore, there is a risk that they will be missed when choosing a set of important variables since their individual importance will be low. In the proposed method, skewness and mean of strongly correlated variables should be similar to each other, because they should be chosen in a tree at the same levels. To illustrate this, 20 correlated variables to the important one from the previous example are added to the simulated database. The number of vehicles in the simulated database is kept unchanged as well as number of noisy variables and censoring rate. Results are presented in Figures 10 - 12. Note that the gap between important and non-important variables using VIMP has almost vanished compared to the previous example in Figure 8. At the same time, skewness and mean of the 21 important variables are similar to the single variable case in Figure 10 and Figure 7, respectively. The main difference is that the number of trees where each variable is chosen has decreased. The minimal depth approach fails in this case and is treating all important variables as non-important, see Figure 12 which is consistent with the observation in Figure 6. The proposed VDD method, as can be seen above, does not suffer of problems with correlated variables like VIMP do.

An example showing why the VDD method could be more advantageous in some situations with respect to VIMP is presented below. The case of one

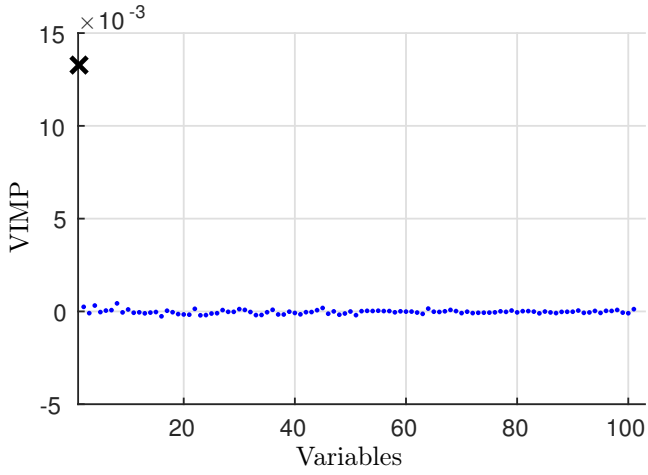


Figure 8: Computed VIMP of simulated data from 10 000 generated vehicles with censoring rate 80 percent. The important variable is marked with a cross.

important variable and 20 correlated is considered. The number of vehicles was reduced to 500 but keeping censoring rate unchanged. The number of noisy variables is also increased to 400, equally splitting between discrete and continuous noise. Results are shown in Figure 13 - Figure 15. Note that the added third dimension helps to separate important variables from noise in Figure 13. VIMP performs worse than VDD, see Figure 14, where the level of importance for some noisy variables is higher than for important ones. The Minimal depth still have problems identifying the important variables as shown in Figure 15.

5.2 Strategy for variable selection

As it was shown above, it is possible to set up a threshold that separates important variables from noisy, however, it is not straightforward. Further studies are required to understand how information contained in the three dimensions could be used to build a consistent and automatic algorithm for variable selection. However, it is possible, using the results in the paper and experience from simulated data, to suggest an ad-hoc strategy.

Variables from the vehicle database are plotted in Figure 16 where the number of trees a variable is used in is used as the third dimension. Important variables should be used in most of the trees. Therefore, selecting a threshold that sorts out variables that are not used in many trees, for example 800, should give a first set of candidates of important variables. It could be the case that important variables are used less if there are many correlated variables, however, in that case it is expected that skewness and mean would be similar for those

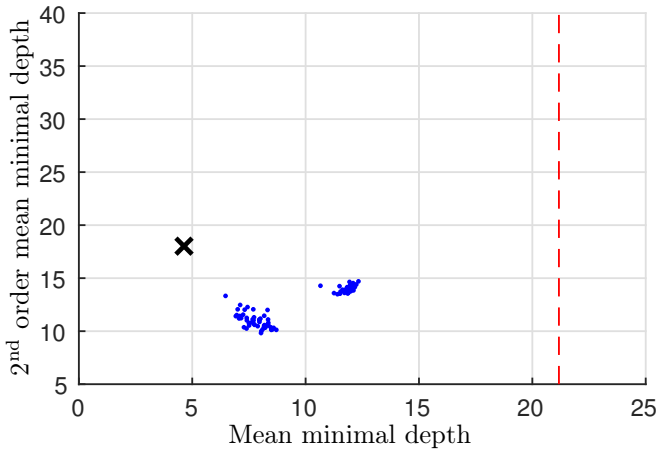


Figure 9: Minimal depth of simulated data from 10 000 generated vehicles with censoring rate 80 percent. The important variable is marked with a cross.

variables, Section 5.1. Then, there should be variables that are grouped in the skewness-mean plane which is not observed for the variables with values of number of trees less than 800. Therefore, it is assumed that there are no important variables in that area. It was shown in Section 5.1 that for some difficult cases, noisy variables are used as splitting variables more often than important variables, see Figure 13. Setting up a threshold with aforementioned strategy will not work for that case. This situation is not considered in this case study, but a more general strategy for selecting the threshold is required for a final version of the variable selection algorithm.

The second step is to project candidate important variables into the skewness-mean plane and to set up a new threshold to remove noisy variables. This step is illustrated in Figure 17. Important variables should have high positive value of skewness and low value of mean. The threshold is manually selected to reject the cloud of variables which are treated as noisy. This step is similar to approach in (Voronov et al., 2016). However, number of variables that are considered to be important is less than in the previous paper due to the augmentation of two dimensional space with the extra dimension. The methodology for variable selection could be summarized in the following steps:

1. Set up threshold in the number of trees dimension to filter out noisy variables which are seldom used.

2. Project remaining variables in the skewness-mean plane and set up a threshold that distinguishes important variables as the subset of variables with high positive value of skewness and low value of mean.

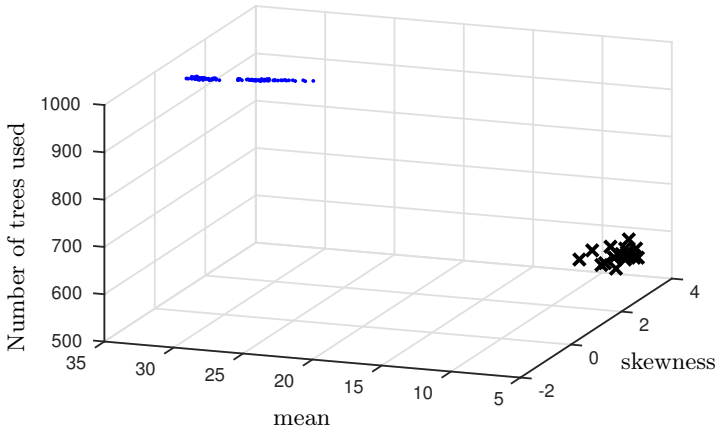


Figure 10: Simulated data from 10000 generated vehicles with censoring rate 80 percent. The important strongly correlated variables are marked as crosses.

6 Case study: Battery failure prognostics

Using the manually chosen thresholds as described in Section 5 and demonstrated with the means of Figures 16-17, 34 of the variables, i.e. about 12 percent, are selected and treated as important. The performance of the RSF using the reduced set of variables is compared to using all variables. The performances of the generated RSF models are evaluated using error rate. However, as discussed earlier in Section 2, the error rate is not an optimal measure since the two models in Figure 1 achieves similar error rates while their prediction quality is significantly different.

An RSF is generated with 1000 trees and a minimal terminal node size of 200 for both variable sets, the 34 selected variables and all variables. The error rate for the case with all variables is 0.2011, and for the reduced set, 0.2177, which are comparable in magnitude. It is worth to emphasize that node size 200 is here used for growing the forest for predictive purposes and node size 2 for variable selection.

For the analysis, 10 vehicles with battery failures and 10 without are selected randomly as validation data. These vehicles are then used as inputs in the RSF to compute the lifetime prediction functions $\mathcal{B}^{\mathcal{V}}(t;t_0)$ and the results are shown in Figures 18 and 19, respectively, for vehicles with battery problems and healthy ones.

In Figure 18 (b), vehicles are clearly more grouped compared to Figure 18 (a) where most vehicles have faster decaying lifetime prediction. The result seems reasonable since lifetime of the batteries of grouped vehicles with fast decaying lifetime prediction functions $\mathcal{B}^{\mathcal{V}}(t;t_0)$ in Figure 18 (b) are within 2 to 3 time units which is quite long life for batteries. Therefore, fast decaying lifetime

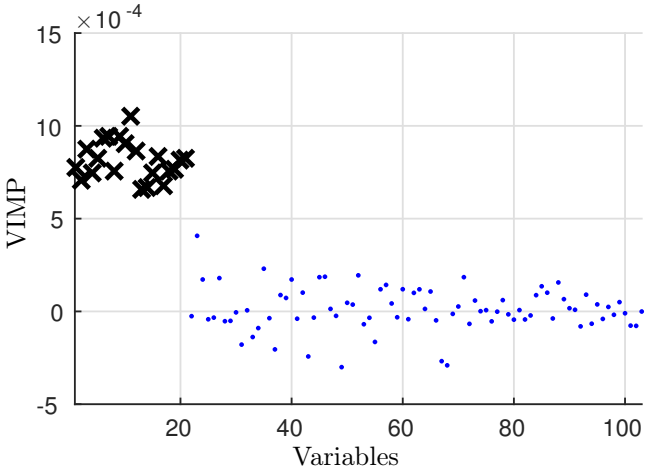


Figure 11: Computed VIMP for simulated data from 10000 generated vehicles with censoring rate 80 percent. The important strongly correlated variables are marked as crosses.

prediction functions for those vehicles should be expected. Battery lifetime of the vehicle corresponding to the purple curve in Figure 18 is about 0.14 time units. However, the vehicle failed early and value of lifetime function would not allow to predict the failure, but it is possible that the cause of the battery problem is not so common in the vehicles from the database. In general, it could be seen that vehicles that lived longer are well separated from the ones that lived shorter. Of cause, it could not be used as the evaluation of the method, but as a positive sign. Note that the lifetime function of the vehicle which corresponds to the green curve in Figure 18 has changed significantly between the two figures. This vehicle operated for about 2.5 time units. It has not yet failed, but should be likely to fail soon. That is why the lifetime prediction function decays faster than for the other vehicles. It should be noticed that we need a measure for assessing predictive performance of RSF, and, when it is available, more can be said about the influence of variable selection on prognostic capabilities of the model.

7 Conclusions

A method for variable selection and variable importance analysis using random survival forests is proposed and analyzed. Main motivating factors for the approach are 1) small number of informative variables, and 2) highly correlated variables in the data set. Analyzing the feature space in Figure 6 indicates that it is possible to distinguish how VIMP and Minimal depth determines which variables that are considered important and this should be analyzed further. The

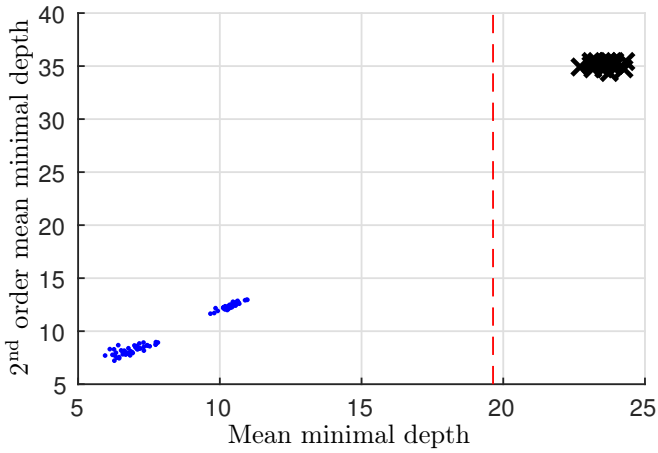


Figure 12: Minimal depth of simulated data from 10000 generated vehicles with censoring rate 80 percent. The important strongly correlated variables are marked as crosses.

proposed method is evaluated in the industrially relevant problem of heavy-duty vehicle battery failure prognostics and evaluated using real vehicle fleet data and simulated data. Simulated data shows that important variables can be distinguished from noisy variables even in difficult cases. The case study using real data shows that a prognosis model with 12% of the available variables achieves comparable error-rate with using all variables.

Acknowledgment

The authors acknowledge Scania and VINNOVA (Swedish Governmental Agency for Innovation Systems) for sponsorship of this work.

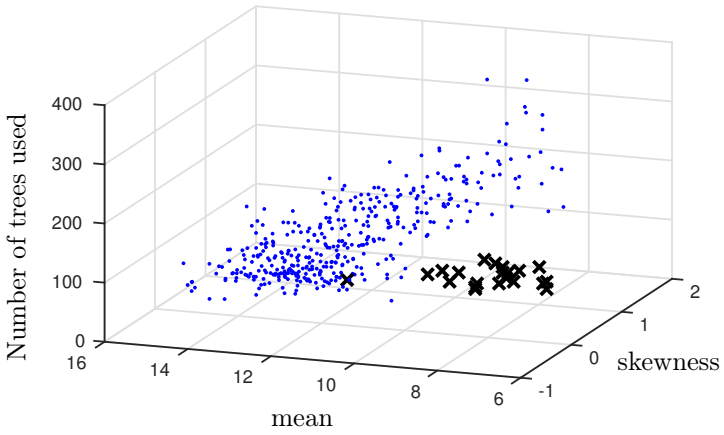


Figure 13: Simulated data from 500 generated vehicles with 21 strongly correlated important variables, 400 noisy variables, and censoring rate 80 percent. Important variables are highlighted with crosses.

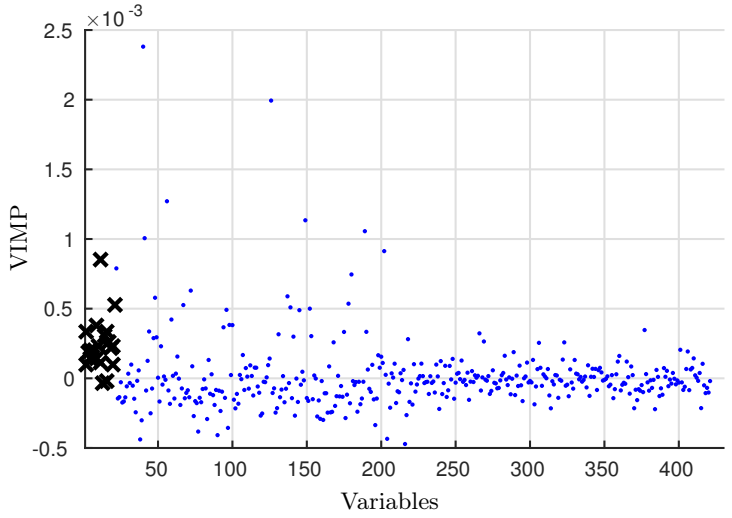


Figure 14: Computed VIMP of simulated data from 500 generated vehicles with with 21 strongly correlated important variables, 400 noisy variables, and censoring rate 80 percent. Important variables are highlighted with crosses.

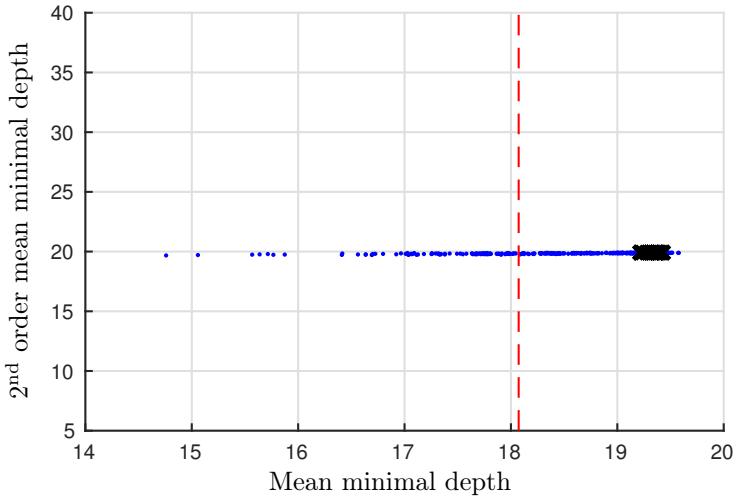


Figure 15: Minimal depth of simulated data from 500 generated vehicles with with 21 strongly correlated important variables, 400 noisy variables, and censoring rate 80 percent. Important variables are highlighted with crosses.

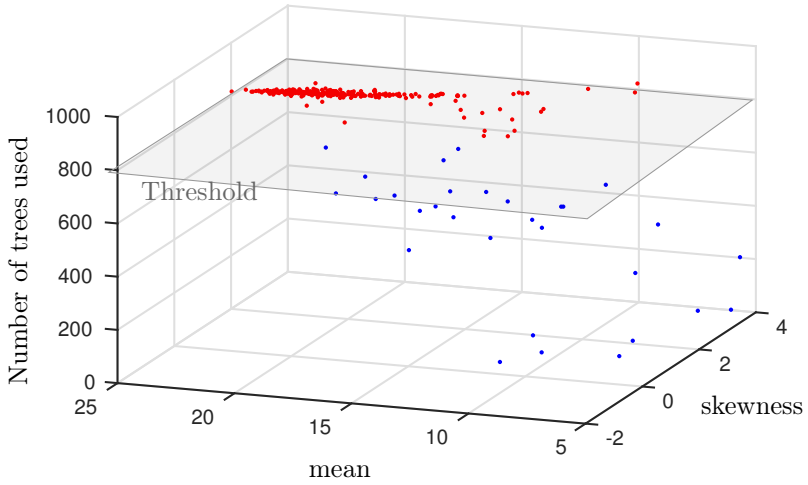


Figure 16: Setting up threshold for the vehicle database. x and y axis are skewness and mean of (10) respectively, and z axis is the number of trees in forest variable was chosen. Red points are candidates for important variables, blue dots - noisy variables, gray plane - threshold value.

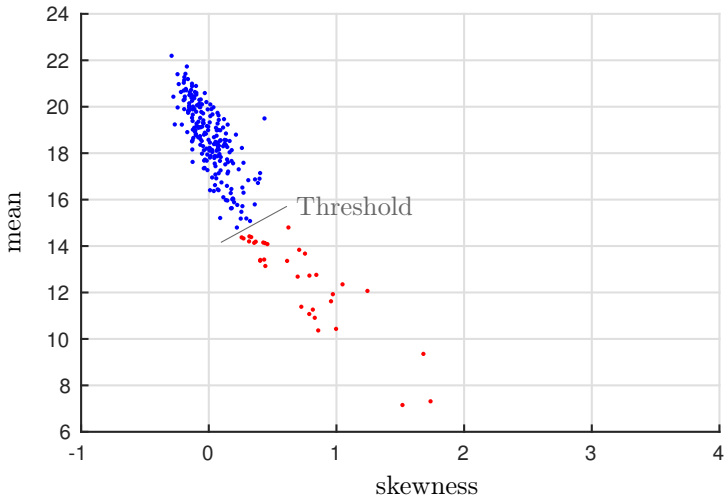
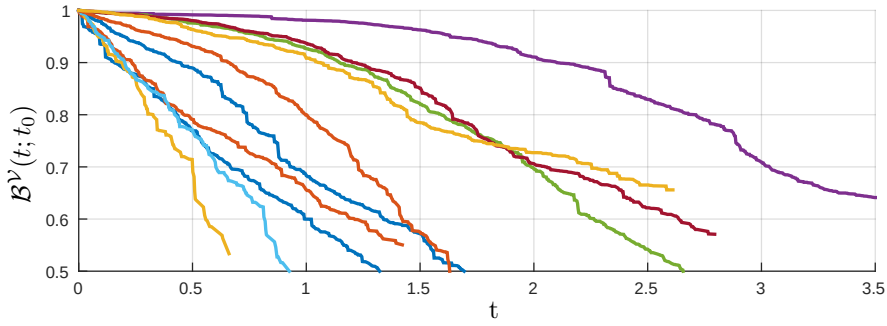
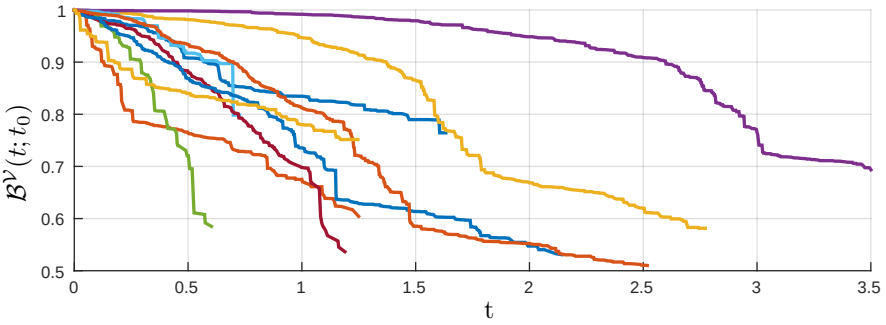


Figure 17: Setting up threshold for the vehicle database. x and y axis are skewness and mean of (10) respectively. Red points correspond to important variables, blue dots - to noisy.

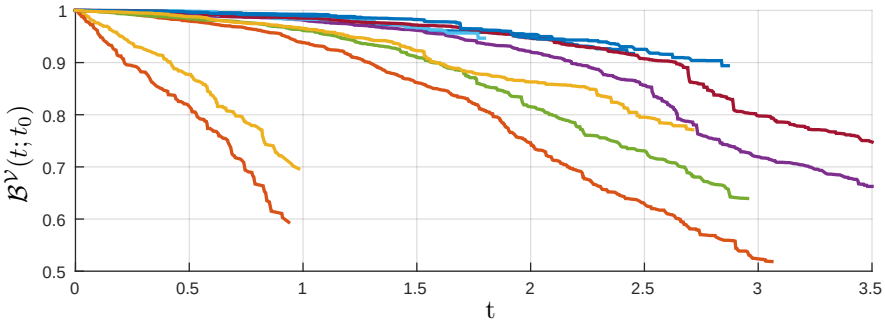


(a) RSF using all variables

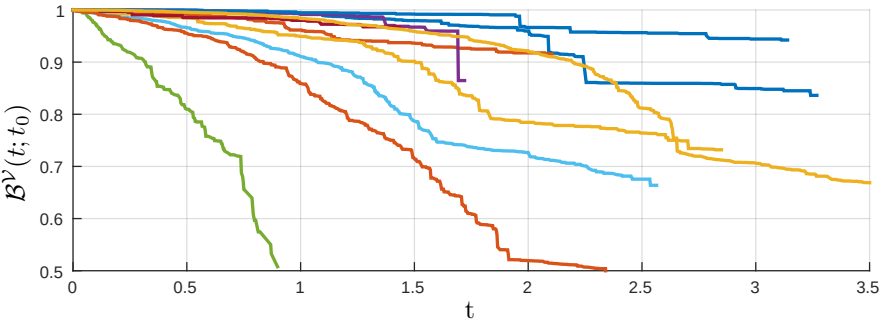


(b) RSF using selected subset of variables

Figure 18: Lifetime prediction function $\mathcal{B}^{\mathcal{V}}(t;t_0)$ for vehicles with battery failures.



(a) RSF using all variables



(b) RSF using selected subset of variables

Figure 19: Lifetime prediction function $\mathcal{B}^{\mathcal{V}}(t;t_0)$ for censored vehicles.

References

D. R. Cox and D. Oakes. *Analysis of survival data*, volume 21. CRC Press, 1984.

M. Daigle and K. Goebel. A model-based prognostics approach applied to pneumatic valves. *International Journal of Prognostics and Health Management Volume 2 (color)*, page 84, 2011.

E. Frisk and M. Krysander. Treatment of accumulative variables in data-driven prognostics of lead-acid batteries. In *Proceedings of IFAC Safeprocess'15*, Paris, France, 2015.

E. Frisk, M. Krysander, and E. Larsson. Data-driven lead-acide battery prognostics using random survival forests. In *Proceedings of the Annual Conference of The Prognostics and Health Management Society*, Fort Worth, Texas, USA, 2014.

I. Guyon and A. Elisseeff. An introduction to variable and feature selection. *The Journal of Machine Learning Research*, 3:1157–1182, 2003.

F. Harrell, R. Califf, D. Pryor, K. Lee, and R. Rosati. Evaluating the yield of medical tests. *Jama*, 247(18):2543–2546, 1982.

H. Ishwaran, U. Kogalur, E. Blackstone, and M. Lauer. Random survival forests. *The Annals of Applied Statistics*, pages 841–860, 2008.

H. Ishwaran, U. Kogalur, E. Gorodeski, A. Minn, and M. Lauer. High-dimensional variable selection for survival data. *Journal of the American Statistical Association*, 105(489):205–217, 2010.

H. Ishwaran, U. Kogalur, X. Chen, and A. Minn. Random survival forests for high-dimensional data. *Statistical Analysis and Data Mining: The ASA Data Science Journal*, 4(1):115–132, 2011.

X. Si, W. Wang, C. Hu, and D. Zhou. Remaining useful life estimation–a review on the statistical data driven approaches. *European Journal of Operational Research*, 213(1):1–14, 2011.

S. Voronov, D. Jung, and E. Frisk. Heavy-duty truck battery failure prognostics using random survival forests. In *Proceedings of Advances in Automotive Control, (Accepted for publication)*, Norrköping, Sweden, 2016.

Paper C

Data-driven battery lifetime prediction and confidence estimation for heavy-duty trucks*

C

*Submitted to a journal for review.

Data-driven battery lifetime prediction and confidence estimation for heavy-duty trucks

Sergii Voronov, Erik Frisk, and Mattias Krysander

Vehicular Systems, Department of Electrical Engineering, Linköping University, SE-581 83 Linköping, Sweden.

Abstract

Maintenance planning is important in the automotive industry as it will allow fleet owners or regular customers to avoid unexpected failures of the components. One cause of unplanned stops of heavy-duty trucks is failures in the lead-acid starter battery. High availability of the vehicles can be achieved by changing the battery frequently, but such an approach is expensive both due to the frequent visits to a workshop and also due to the component cost. Here, a data-driven method based on Random Survival Forest (RSF) is proposed for predicting the reliability of the batteries. The data set, covering more than 50000 trucks, available for the study has two important properties. First, it does not contain measurements related directly to the battery health, in addition there are no time series of measurements for every vehicle. It is shown in the paper that the RSF method is used to predict the reliability function for a particular vehicle using data from the fleet of vehicles given that only one set of measurements per vehicle is available. A theory behind confidence bands for the RSF method is developed which is the extension of the existing technique for estimating variance of Random Forest method. Adding confidence bands to the RSF method gives an opportunity to an engineer to evaluate the confidence of the model prediction. Some aspects of the confidence bands are considered: a) their asymptotic behavior and b) usefulness in model selection. A problem of including time related variables is addressed in the paper with arguments why it is a good choice not to add them into the model. Metrics for performance evaluation are suggested which show that the model can be used to schedule and optimize the cost of the battery replacement.

1 Introduction

In order to transport goods efficiently by heavy-duty trucks, it is important that vehicles have a high degree of availability and in particular avoid becoming standing by the road unable to continue the transport mission. An unplanned stop by the road does not only cost due to the delay in delivery, but can also lead to damaged cargo. Therefore, maintenance planning becomes important in the automotive industry and in the near future car or truck manufactures do not only produce and deliver cars and trucks, but also provide maintenance services that will allow fleet owners or regular customers to avoid unexpected failures. High availability can be achieved by changing components frequently, but such an approach is expensive both due to the frequent visits to a workshop and also due to the component cost. Therefore, failure prognostics and flexible maintenance has significant potential in the automotive field for both manufacturers, commercial fleet owners, and private customers.

In heavy-duty trucks, one cause of unplanned stops are failures in the electrical power system, and in particular, the lead-acid starter battery. The main purpose of the battery is to power the starter motor to get the diesel engine running, but it is also used to, for example, power auxiliary units such as cabin heating and kitchen equipment. Detailed physical models of battery degradation is inherently difficult and requires, in addition to battery health sensing which is not available in the given study, detailed knowledge of battery chemistry and how degradation depends on the vehicle and battery usage profiles.

Methods for lifetime prognostics of system components can coarsely be split into two categories: model based and data-driven methods (Roemer et al., 2005). Model based methods rely on physical laws and equations that describe degradation of the components and for accurate predictions, accurate degradation models are required. However, it is sometimes hard to develop an accurate degradation model for a particular system, and then data-driven methods can be an alternative. It is common for both approaches to estimate the Remaining Useful Life (RUL) which is either the remaining time until component failure or to the point where it can no longer fulfill its function. In general, RUL is estimated using sensors that give health related information of the component, meaning, there is a possibility to track and predict the state of the health related parameters during the lifetime of the component. Examples of model-based prognostics are given in (Daigle and Goebel, 2011; Hanachi et al., 2015; Saha and Goebel, 2009) where detailed physics-based degradation models are developed and used. Data-driven methods use machine learning algorithms to either estimate RUL, or health of the component, and can be categorized into parametric and non-parametric methods. A parametric approach assumes that the underlying degradation can be well described by a parametric distribution where the parameters of interest are estimated through the observations, see for example (Medjaher et al., 2012; Cox, 1972). In turn, non-parametric data-driven models use machine learning methods that do not have any basic assumption regarding underlying degradation distribution (Ishwaran et al., 2008). Nowadays,

hybrid methods that fuse predictions both from the model-based and data-driven approaches are proposed, see (Zhao et al., 2015; Liao and Kottig, 2014). Unlike the aforementioned methods, where in most cases time series of sensor data is available, the data set under study only contains static information, i.e., one can not observe a battery degradation during its lifetime. For this reason, an alternative approach is adopted here where a conditional probability distribution of the battery lifetime, referred to as the battery lifetime function, is estimated instead of the RUL.

Given snapshots from a fleet of the vehicles coming into a workshop, the problem of estimating the lifetime function of the lead-acid battery, using a non-parametric approach, for the vehicles is considered in order to decide when to replace its battery. A distinctive feature of the data set is lack of information directly related to the battery health. Therefore, battery health must be estimated using available information in variables correlated with battery usage. Taking all into account and the fact that the degradation profiles of the batteries are not available, a non-parametric method called Random Survival Forest (RSF) (Ishwaran et al., 2008) is selected to estimate the reliability function of a particular battery and subsequently the lifetime function. Contributions in this paper are the following: a) the lifetime function is used instead of the RUL and the RSF model is proposed to estimate the lifetime function, b) a variance estimate of the predictor is suggested which uses the structure of the RSF model allowing to judge the quality of the prediction and c) an analysis of the predictive capabilities of the RSF model with different sets of input variables.

2 Problem motivation

Prognostics for flexible maintenance of batteries in heavy-duty vehicles is the topic of this study. To illustrate the potential for flexible maintenance in the case under study, consider the distribution of failure and censoring times, times when a vehicles leaves the study without experiencing a battery failure, in Fig. 1 (time is scaled to avoid revealing sensitive information). The shape of the distribution of failed vehicles, red curve in the figure, is such that it is impossible to set up an efficient maintenance point to replace the battery. If the maintenance point is scheduled, for instance, around 0.5 time units, then majority of the batteries are replaced before failure, however, the batteries are not used efficiently, in addition, customers will not be happy with the quality of the batteries if they are changed too soon and will shift to another battery manufacturer who can deliver better service. On the other hand, if the maintenance point is scheduled around 5 time units, majority of the batteries are used till the end of their lives, but many battery failures are missed. Therefore, the figure motivates the need for a vehicle specific prognostic model described in the paper. Before the studied problems are explicitly stated, the vehicle fleet data is introduced which is used to build the model.

2.1 Vehicle fleet data

The data source is a vehicle fleet database from an industrial partner, Scania CV a heavy-duty truck manufacturer in Sweden. Each vehicle has a record, called snapshot, in the database which tells how the vehicle was used during its complete lifetime until the snapshot time. The snapshot is comprised of variables where a subset of the variables correspond to the vehicle configuration, i.e., are fixed for the complete life of the vehicle. Other variables are related to usage of the vehicle and will therefore change over time. Information is logged in the database when a vehicle comes to a workshop and it is noted in the data if the vehicle has had a battery problem and a time stamp of the event. A snapshot with no indication on battery problems is called censored, since the future time of failure is not known. Information present in the database is general purpose, meaning it is not designed for battery prognostics and there is no specific battery health indicator in the data. In addition, there are relatively few variables that are directly related to the battery usage.

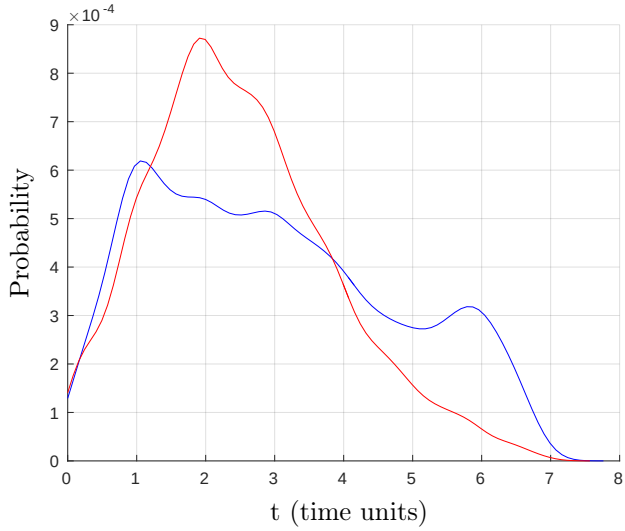


Figure 1: Distribution of failure times for censored, blue curve, and failed vehicles, red curve.

Main characteristics of the database are:

- 56,163 vehicles from 5 EU markets
- 536 variables stored in each vehicle snapshot
- One single snapshot per vehicle

- Heterogeneous data, i.e., a mixture of categorical and numerical data
- Histogram variables
- Censoring rate about 80 percent
- Significant missing data rate about 40 percent

The data set includes both categorical and numerical variables where categorical variables have a limited number of possible values. For example, the battery position variable has three possible values (right, left hand side and rear frame end). Numerical data is mostly organized in the form of histograms but there are, so called, accumulative variables such as mileage and age which increase with time. As an example, one of the histogram data is a voltage histogram that has ten bins, each showing what fraction of time the battery of the vehicle have been operating in a particular voltage range. Here, every bin of the histograms is treated as a separate variable and then the voltage histogram contributes with 10 variables to the study. The censoring rate is another distinctive property. Only a fraction of the vehicles have problems with batteries while all others do not, meaning that the failure times are censored. Missing data is also an essential characteristic of many real life data sources and the main reason in our case is the fact that variables introduced for one type of vehicle are not relevant for another type. The missing data rate is about 40 % and it should be noted that missing values are not uniformly distributed among variables. Specific variables can have significantly higher missing rate than others. Thus, systematic handling of missing data is important in the proposed approach.

Another thing to notice is that there are no time series of snapshots for the vehicles and therefore it is not possible to track degradation of the battery over time for a given vehicle. All characteristics of the database mentioned above significantly influence the choice of techniques in the proposed approach.

2.2 Battery lifetime function

A probabilistic framework is used to describe the battery prognostic information corresponding to the battery health. In model based prognostics a health indicator is generally measured or modeled, and it is possible then to track the health indicator during the whole life of a battery. Here, there are no variables in the database of study which correspond directly to battery health, in addition, properties of the database such as missing data rate and censoring will add uncertainties to the predictor. Therefore, a probabilistic model is used since it is then possible to explicitly represent the inherent uncertainty in the model.

Let a random variable T be the battery failure time, $\mathcal{V}$ the snapshot of variables for a given vehicle taken at time point t_0. The main objective is to estimate the conditional probability function, here referred to as lifetime prediction function, of the battery defined as

$$\mathcal{B}^{\mathcal{V}}(t; t_0) = P(T > t + t_0 \mid T \geq t_0,\ \mathcal{V}). \tag{1}$$

The function states the probability that failure time T for a battery of interest is greater than $t+t_0$ time units given that it has survived t_0 time units conditioning on snapshot data $\mathcal{V}$. Prediction of battery lifetime can be made, for example, in the workshop when data is retrieved from the vehicle. The lifetime function $\mathcal{B}^{\mathcal{V}}(t;t_0)$ can be expressed using the established reliability function $R^{\mathcal{V}}(t) = P(T \geq t \mid \mathcal{V})$, see (Cox and Oakes, 1984), as

$$\mathcal{B}^{\mathcal{V}}(t;t_0) = \frac{R^{\mathcal{V}}(t+t_0)}{R^{\mathcal{V}}(t_0)}. \tag{2}$$

2.3 Estimate confidence of a predictor model

As mentioned in Section 2.2, the objective is to estimate the battery lifetime prediction function (2). To evaluate if an estimate is reliable or not, some measure of confidence is needed. A common approach is to add a standard deviation forming confidence bands of the estimator. Fig. 2 shows a prediction with a corresponding confidence band, a 95% confidence band based on a Gaussian assumption, i.e., assumption that the estimator is normally distributed. The figure is based on a synthetic data set to show the advantage of adding confidence bands. There are 5 classes of the vehicles with different degradation profiles of the batteries in the synthetic data set. Fig. 2 demonstrates estimation of the true reliability for one of the classes. Information about the true reliabilities is not available in the real data set, and, therefore, the synthetic data set is used to show statistical properties of the estimator. When classes of vehicles with true degradation profiles are known, it is possible to compute a Kaplan-Meier estimate, a maximum likelihood estimate of the reliability function (Cox and Oakes, 1984), shown as a green curve in Fig. 2, and the 95 % confidence bands with Gaussian assumption using a standard deviation estimated by a the Greenwood formula (Cox and Oakes, 1984), dashed blue curves in Fig. 2. A main problem studied in the paper is how to estimate standard errors and confidence intervals for a battery lifetime function estimator. A main difference between this case and the basic survival analysis case is that in the battery data case there are not a set of distinct degradation classes but rather a continuum of degradation profiles and therefore, the Kaplan-Meier and Greenwood formula are not directly applicable.

2.4 Summary

Maintenance planning is based on the estimation of the battery lifetime function together with confidence bands. The main objective is to estimate vehicle individual battery lifetime functions together with variance estimates of the predictor. Analysis regarding the predictive capabilities of the RSF models with different type of variables is carried out and properties of the estimator are analyzed on both the real data set and synthetic data where the ground truth is known.

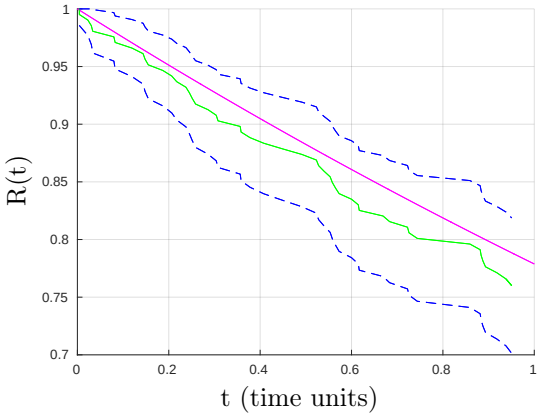


Figure 2: True reliability function for a vehicle battery, magenta curve, Kaplan-Meier estimate for the given class of vehicles, green curve, and 95 % confidence bands with Gaussian assumption where variance estimated using Greenwood formula, blue dashed curves.

3 Lifetime prediction function model

In the selection of a suitable data-driven method for the studied battery prognostic problem, a few things need to be taken into account. The underlying baseline hazard functions, i.e., the failure rates, are not known in the data set under study. In addition, it is not clear how to estimate parameters in the case of a parametric model such as in a Cox regression model, see (Cox, 1972). This is the main reason for choosing a non-parametric approach and Random Survival Forest (RSF) (Ishwaran et al., 2008) is a suitable method that gives the ability to handle different types of data, direct applicability of the method to survival analysis, and automatic missing data imputation. The output from the RSF model is a reliability function which can be directly used to estimate the lifetime function (2). The basic idea of the RSF model is to group vehicles with similar battery degradation characteristics and estimate the reliability function for that particular group of vehicles.

Next, Random Survival Forest is briefly summarized in Section 3.1 and then the approach is applied to the battery prognostic case in Section 3.2.

3.1 Random survival forests

Classification and regression trees are machine learning techniques that maps/predicts a feature or variable space X into a space of outcomes Y by means of binary trees (Breiman et al., 1984) where features and outcome for a particular case are considered as a pair (x_i, y_i) . Target values y_i from the outcome space could

be continuous valued in case of regression and discrete in case of a classification problem. A decision tree is a non-linear estimator

$$\hat{\theta}(x_i) = \hat{y}_i \tag{3}$$

where $\hat{\theta}(x)$ is built by partitioning the feature space X into disjoint regions R_m with some fitting model for each region. For a regression problem a fitting model is a real value that fits data in a region R_m best, for instance the mean, while for the classification fitting value is, for example the majority class among all classes in the given region.

The aforementioned partitioning process happens at every node of the tree. For a basic decision tree the best splitting variable and splitting value is determined in a greedy manner, namely, all variables and every possible splits are accessed based on a cost function. The split with the lowest value of the cost function is then selected. Decision trees can be applied to data sets with different types of variables and another advantage is interpretability as rules can be built from a single decision tree. A decision tree is a weak classifier and generally perform well on the training data, however, they usually generalize poorly on unseen data.

Therefore, ensemble of trees, a Random Forest (RF) model, was introduced by Breiman (2001). There are different implementations of ensemble of trees such as (Dietterich, 2000) and (Ho, 1998), however, the basic Breiman model is described here since the RSF model is an extension of RF. There are two techniques that are the distinctive features of the RF method, namely, bootstrap aggregation, also known as bagging, and a step that reduces correlation between trees in the forest. When number of data samples is small, bootstrap is a powerful method for estimating statistics. By sampling from the given data samples with replacement one can construct a significantly large set of new samples that can be used to estimate target statistics. Bootstrap aggregation is an ensemble method that combines predictions from different machine learning models. In the case of trees, a number of sets of bootstrap samples are created and then a classification or regression tree model is fitted for each of bootstrap sample. As mentioned, a single tree model is sensitive to unseen data, but by combing outputs from a set of trees, grown on different bootstrap samples, the resulting output has reduced variance of a predictor compared to the single tree model. In regression, the output from a bootstrap aggregation model is the mean of outputs of all trees

$$\hat{\theta}_{\mathrm{BAGG}}(x) = \frac{1}{B}\sum_{i=1}^{B}\hat{\theta}_i(x) \tag{4}$$

where $\hat{\theta}_i(x)$ is a tree model fitted to the i^{th} bootstrap sample, and B is the number of trees/bootstrap samples. It was suggested by Breiman (Breiman, 2001) that introducing randomness into the procedure of choosing variables for splitting reduces correlation between trees and increase performance of the

aggregated model. Therefore, instead of choosing all m available variables for split at each node, only a fraction p of them is considered. This step also increases speed of the algorithm as it requires less variables to check at each split.

A Random Survival Forest (RSF) model is a RF model modified for the purpose of survival analysis (Ishwaran et al., 2008). Structurally, an RSF model is similar to a RF except for the following changes. The cost function used for splitting is so called log-rank test (Ciampi et al., 1986). It is a hypothesis test which compares survival distributions of samples that are formed by dividing data available at the splitting node into two samples which will be the part of the two child nodes. The best split corresponds to a variable with a value under which two samples have as distinctive degradation profiles as possible. The log-rank test is non-parametric and designed for censored data, a type of data encountered in survival analysis. At each terminal node, a node at which splitting no longer is performed, the Nelson-Aalen estimate of the cumulative hazard rate $H(t)$ is computed (Cox and Oakes, 1984). The estimated cumulative hazard rate $\hat{H}(t)$ of the whole forest is computed by averaging over tree hazard rates. Since the cumulative hazard rate is the negative logarithm om the reliability function $R(t)$ Cox and Oakes (1984), the estimate $\hat{R}(t)$ of the reliability function is computed as

$$\hat{R}(t) = e^{-\hat{H}(t)}. \tag{5}$$

The estimate $\hat{R}(t)$ of the reliability function is the forest output.

3.2 Battery prediction model

The output from the RSF is an estimate of reliability function as in (5). Then, an estimate of the lifetime function $\hat{\mathcal{B}}^{\mathcal{V}}(t, t_0)$ can be expressed directly from (2) as

$$\hat{\mathcal{B}}^{\mathcal{V}}(t, t_0) = \frac{\hat{R}^{\mathcal{V}}(t + t_0)}{\hat{R}^{\mathcal{V}}(t_0)} \tag{6}$$

4 Confidence estimate for the battery lifetime prognostics function

Assume a bagged predictor (4). Such an estimator is complex, nonlinear, and deriving an explicit expression for the estimation covariance is infeasible. Then, one option is to use a bootstrap technique. Since the estimator already uses a bootstrap technique, a bootstrap strategy for estimating the variance would require to compute bootstrap of bootstraps which is not computationally feasible, (Efron, 2014). Thus, an approach that uses the original bootstrap samples used when building the model also for estimating variance is desired. One possibility for such an approach is the Infinitesimal Jackknife (IJ) variance estimate suggested in (Efron et al., 2014) for random forests. The basic approach

is described in Section 4.1 and then the technique will be extended to RSF and the battery lifetime function in Section 4.2.

4.1 Theoretical background on IJ variance estimation

To summarize results from (Efron et al., 2014), consider the i^{th} bootstrap sample $\boldsymbol{Y_i^*} = (y_{i1}^*, y_{i2}^*, \ldots, y_{in}^*)$ which is sampled from the initial data set $\boldsymbol{Y} = (y_1, y_2, \ldots, y_n)$ where y_{ij}^* represents the number of times a particular data point, a snapshot of the vehicle from the data set in the given study, is included in the bootstrap sample. Introduce a resampling vector as

$$\boldsymbol{P} = (p_1, p_2, \ldots, p_n) \tag{7}$$

where p_i denotes probability of selecting y_i in the bootstrap sample. This vector belongs to a set such that

$$\mathcal{L}_n = \left\{ \boldsymbol{P} : P_i \geq 0, \sum_{i=1}^{n} P_i = 1 \right\}. \tag{8}$$

The resampling vector represents the weight each data point y_i from the initial sample $\boldsymbol{Y} = (y_1, y_2, \ldots, y_n)$ has in the i^{th} bootstrap sample. For example, the resampling vector $\boldsymbol{P}^0 = (\frac{1}{n}, \ldots, \frac{1}{n})$ is associated with an initial sample $\boldsymbol{Y}$ where each element of the sample has equal weight. The infinitesimal jackknife (IJ) variance estimate is based on a linearization approach and the variance estimate $\hat{V}_{\mathrm{IJ}}$ of the true variance $\mathrm{var}\left[\hat{\theta}_{\mathrm{BAGG}}\right]$ of the bagged predictor is

$$\hat{V}_{\mathrm{IJ}} = \frac{1}{n^2} \sum_{i=1}^{n} U_i^2 \tag{9}$$

where n is the size of the sample and U_i are the directional derivatives

$$U_i = \lim_{\epsilon \to 0} \frac{\hat{\theta}_{\mathrm{BAGG}}(\boldsymbol{P}^0 + \epsilon(\boldsymbol{\delta}_i - \boldsymbol{P}^0)) - \hat{\theta}_{\mathrm{BAGG}}(\boldsymbol{P}^0)}{\epsilon}, \qquad i = 1, \ldots, n \tag{10}$$

with $\boldsymbol{\delta}_i$ being the i^{th} coordinate vector. For a bagged estimator, it turns out that there exists an explicit expression for the asymptotic, with respect to number of bootstrap samples B, expression of the directional derivatives

$$\hat{V}_{\mathrm{IJ}} = \sum_{i=1}^{n} \widehat{\mathrm{Cov}}_i^{\,2} \tag{11}$$

where

$$\widehat{\mathrm{Cov}}_i = \frac{1}{B} \sum_{b=1}^{B} (y_{bi}^* - 1)(t_b^* - \bar{t}).$$

Here, the b^{th} tree grown on the b^{th} bootstrap sample is built with the Breiman procedure and $\bar{t}$ is the RF output. As stated in Efron et al. (2014), estimator (11) is biased and an improved unbiased estimator can be derived as

$$\hat{V}_{\text{IJ-U}} = \hat{V}_{\text{IJ}} - \frac{n}{B^2}\sum_{b=1}^{B}(t_b^* - \bar{t})^2 \tag{12}$$

4.2 IJ variance estimate for the lifetime function

There are two main differences between IJ variance estimate of the RF model compared to variance estimate of lifetime function (6). First, the output of the RF model is either a class or regression value, but in the RSF case the output is a time dependent function, and secondly, the lifetime function is a ratio of the reliability estimates $\hat{R}^{\mathcal{V}}(t)$ as in (2).

For the first difference mentioned above, the reliability function is computed on a predefined grid of time points, the variance estimate $\hat{V}_{\text{IJ}}^{\text{RSF}}(t)$ of the true forest variance var $\left[\hat{\theta}_{\text{RSF}}\right]$ becomes

$$\hat{V}_{\text{IJ}}^{\text{RSF}}(t) = \sum_{i=1}^{n}\widehat{\text{Cov}}_i^{\,2}(t) \tag{13}$$

where

$$\widehat{\text{Cov}}_i(t) = \frac{1}{B}\sum_{b=1}^{B}(y_{bi}^* - 1)(\hat{R}_b^{\mathcal{V}}(t) - \hat{R}^{\mathcal{V}}(t)). \tag{14}$$

Here, the reliability $\hat{R}_b^{\mathcal{V}}(t)$ is the output reliability from the b^{th} tree for a particular vehicle with data $\mathcal{V}$ and $\hat{R}^{\mathcal{V}}(t)$ is the output from the forest. These values correspond to t_b^* and $\bar{t}$ in (12) respectively. An unbiased IJ variance estimate $\hat{V}_{\text{IJ-U}}^{\text{RSF}}$ in analogy with Efron's estimate is then

$$\hat{V}_{\text{IJ-U}}^{\text{RSF}}(t) = \hat{V}_{\text{IJ}}^{\text{RSF}}(t) - \frac{n}{B^2}\sum_{b=1}^{B}(\hat{R}_b^{\mathcal{V}}(t) - \hat{R}^{\mathcal{V}}(t))^2. \tag{15}$$

For the second property, the variance estimate for the lifetime function $\hat{\mathcal{B}}^{\mathcal{V}}(t, t_0)$ from (6), which is a ratio of the outputs of the random survival forest, is estimated and summarized next.

Theorem 1. *Let $\mathcal{B}^{\mathcal{V}}(t, t_0)$ in (2) be the battery lifetime function. Then*

$$\hat{\mathcal{B}}^{\mathcal{V}}(t, t_0) = \frac{\hat{R}^{\mathcal{V}}(t + t_0)}{\hat{R}^{\mathcal{V}}(t_0)}$$

is the RSF estimate of $\mathcal{B}^{\mathcal{V}}(t, t_0)$ and a first order IJ variance estimate is given by

$$var\left[\hat{\mathcal{B}}^{\mathcal{V}}(t, t_0)\right] \approx \left(\frac{\mu_X}{\mu_Y}\right)^2 \cdot \left(\frac{var[X]}{\mu_X^2} + \frac{var[Y]}{\mu_Y^2} - 2\frac{cov[X, Y]}{\mu_X \mu_Y}\right) \tag{16}$$

where the random variable X is the reliability function $\hat{R}^{\mathcal{V}}(t+t_0)$, the random variable Y is the reliability function $\hat{R}^{\mathcal{V}}(t_0)$, and

$$\begin{aligned} \mu_X &\approx \hat{R}^{\mathcal{V}}(t+t_0) \\ \mu_Y &\approx \hat{R}^{\mathcal{V}}(t_0) \\ var\,[X] &= \hat{V}^{RSF}_{IJ\text{-}U}(t+t_0) \\ var\,[Y] &= \hat{V}^{RSF}_{IJ\text{-}U}(t_0) \\ cov\,[X,Y] &= cov_{Bias}\,[X,Y] - Bias. \end{aligned}$$

The random forest directly gives μ_X, μ_Y above, and the infinitesimal jackknife estimator (12) gives $\text{var}\,[X]$ and $\text{var}\,[Y]$. A result for the estimation of $\text{cov}\,[X,Y]$ is given in Lemma 1.

Proof. From (2), the lifetime function can be expressed as the ratio of the reliability functions $\hat{R}^{\mathcal{V}}(t)$ and $\hat{R}^{\mathcal{V}}(t+t_0)$. Assume that $\hat{R}^{\mathcal{V}}(t+t_0)$ is a random variable X and $\hat{R}^{\mathcal{V}}(t_0)$ is a random variable Y. Then, the variance of the lifetime function can be estimated using a Taylor series expansion as (16) where instead of μ_X and μ_Y the outputs from the forest $\hat{R}^{\mathcal{V}}(t+t_0)$ and $\hat{R}^{\mathcal{V}}(t_0)$ are used at time $t+t_0$ and t_0 respectively. The variances $\text{var}\,[X]$ and $\text{var}\,[Y]$ correspond to IJ variance estimates $\hat{V}^{\text{RSF}}_{\text{IJ-U}}(t)$ computed at time $t+t_0$ and t_0 respectively. Covariance $\text{cov}\,[X,Y] = \widehat{\text{cov}}\left[\hat{R}^{\mathcal{V}}(t+t_0), \hat{R}^{\mathcal{V}}(t_0)\right]$ is a covariance between two random variables which are represented by the values of two points from the reliability curve $\hat{R}^{\mathcal{V}}(t)$ at time $t+t_0$ and t_0. □

The missing part and a main contribution is the derivation of $\text{cov}\,[X,Y] = \widehat{\text{cov}}\left[\hat{R}^{\mathcal{V}}(t+t_0), \hat{R}^{\mathcal{V}}(t_0)\right]$ using an infinitesimal jackknife approach also here. The result is summarized in the lemma below.

Lemma 1. *Let $\hat{R}^{\mathcal{V}}(t)$ be an RSF model with B trees grown on the original sample $\mathbf{Y} = (y_1, y_2, \ldots, y_n)$ with size n. Assume that the output, $\hat{R}^{\mathcal{V}}_b(t)$, from tree b is independent from one data point y^*_{ij} from the i^{th} bag, then an asymptotic expression of the infinitesimal jackknife estimate of $\widehat{cov}\left[\hat{R}^{\mathcal{V}}(t+t_0), \hat{R}^{\mathcal{V}}(t_0)\right]$ and the corresponding bias correction are*

$$cov\,[X,Y] = cov_{Bias}\,[X,Y] - Bias \tag{17}$$

where

$$cov_{Bias}\,[X,Y] \;=\; \widehat{cov}\left[\hat{R}^{\mathcal{V}}(t+t_0), \hat{R}^{\mathcal{V}}(t_0)\right] \;=\; \sum_{i=1}^{n} \widehat{Cov}_i(t_0)\widehat{Cov}_i(t+t_0) \tag{18}$$

and

$$Bias \;=\; \frac{n}{B^2}\sum_{i=1}^{B}(\hat{R}^{\mathcal{V}}_i(t_0) \;-\; \hat{R}^{\mathcal{V}}(t_0))(\hat{R}^{\mathcal{V}}_i(t \;+\; t_0) \;-\; \hat{R}^{\mathcal{V}}(t \;+\; t_0)) \tag{19}$$

as the sample size $n \to \infty$, the number of trees $B \to \infty$, and n tends to infinity faster than B.

Proof. According to the definition of the covariance

$$\widehat{\mathrm{cov}}_{\mathrm{Bias}}\left[\hat{R}^{\mathcal{V}}(t+t_0), \hat{R}^{\mathcal{V}}(t_0)\right] = \\ = E\left[\left(\hat{R}^{\mathcal{V}}(t+t_0) - E\left[\hat{R}^{\mathcal{V}}(t+t_0)\right]\right) \cdot \left(\hat{R}^{\mathcal{V}}(t_0) - E\left[\hat{R}^{\mathcal{V}}(t_0)\right]\right)\right] \quad (20)$$

Now, let us write the estimate from the forest for a particular time point t as $\hat{\theta}_{\mathrm{RSF}}(\boldsymbol{P}, t) = \hat{R}^{\mathcal{V}}(t)$ which corresponds to one point on the reliability function. An expansion of nonlinear estimator $\hat{\theta}_{\mathrm{RSF}}(\boldsymbol{P}, t)$ using directional derivatives around resampling vector $\boldsymbol{P}^0$ keeping only a linear term gives

$$\hat{\theta}_{\mathrm{RSF}}(\boldsymbol{P}, t) = \hat{\theta}_{\mathrm{RSF}}(\boldsymbol{P}^0) + (\boldsymbol{P} - \boldsymbol{P}^0) \cdot \boldsymbol{U} + \mathcal{O}((\boldsymbol{P} - \boldsymbol{P}^0) \cdot (\boldsymbol{P} - \boldsymbol{P}^0)') \quad (21)$$

where $\boldsymbol{U}(t)$ is a column vector of directional derivatives

$$U_i(t) = \lim_{\epsilon \to 0} \frac{\hat{\theta}_{\mathrm{RSF}}(\boldsymbol{P}^0 + \epsilon(\boldsymbol{\delta}_i - \boldsymbol{P}^0), t) - \hat{\theta}_{\mathrm{RSF}}(\boldsymbol{P}^0, t)}{\epsilon}, \quad i = 1, \ldots, n \quad (22)$$

Taking the result in (21) into account, covariance of reliabilities in (20) becomes

$$\widehat{\mathrm{cov}}_{\mathrm{Bias}}\left[\hat{R}^{\mathcal{V}}(t+t_0), \hat{R}^{\mathcal{V}}(t_0)\right] = \\ = E\left[\left(\hat{\theta}_{\mathrm{RSF}}(\boldsymbol{P}, t+t_0) - E\left[\hat{\theta}_{\mathrm{RSF}}(\boldsymbol{P}, t+t_0)\right]\right) \cdot \left(\hat{\theta}_{\mathrm{RSF}}(\boldsymbol{P}, t_0) - E\left[\hat{\theta}_{\mathrm{RSF}}(\boldsymbol{P}, t_0)\right]\right)\right] = \\ = E\left[\left((\boldsymbol{P} - \boldsymbol{P}^0)\boldsymbol{U}(t+t_0)\right)\left((\boldsymbol{P} - \boldsymbol{P}^0)\boldsymbol{U}(t_0)\right)\right] \quad (23)$$

It could be seen that a resampling vector for each tree has a rescaled multinominal distribution

$$\boldsymbol{P} \sim \frac{\mathrm{Mult}_n(n, \boldsymbol{P}^0)}{n}$$

with mean and covariance matrices

$$\left(\boldsymbol{P}^0, \frac{\boldsymbol{I}}{n^2} - \frac{\boldsymbol{P}^{0'}\boldsymbol{P}^0}{n}\right)$$

Covariance expression with the directional derivatives becomes

$$\begin{aligned}
&\widehat{\mathrm{cov}}\left[\hat{R}^{\mathcal{V}}(t+t_0), \hat{R}^{\mathcal{V}}(t_0)\right] = \\
&\quad = E\left[\left(\sum_{i=1}^{n}(p_i - \frac{1}{n})U_i(t+t_0)\right)\left(\sum_{j=1}^{n}(p_j - \frac{1}{n})U_j(t_0)\right)\right] = \\
&\quad = E\left[\sum_{i=1}^{n}\left(p_i - \frac{1}{n}\right)^2 U_i(t_0)U_i(t+t_0) + \right. \\
&\quad + \sum_{i\neq j}\left(p_i - \frac{1}{n}\right)\left(p_j - \frac{1}{n}\right)U_i(t_0)U_j(t+t_0) + \\
&\quad \left. + \sum_{i\neq j}\left(p_i - \frac{1}{n}\right)\left(p_j - \frac{1}{n}\right)U_i(t+t_0)U_j(t_0)\right] = \\
&\quad = \sum_{i=1}^{n}\frac{1}{n^2}\left(1 - \frac{1}{n}\right)U_i(t_0)U_i(t+t_0) + \\
&\quad + \sum_{i\neq j}\left(-\frac{1}{n^3}\right)U_i(t_0)U_j(t+t_0) + \\
&\quad + \sum_{i\neq j}\left(-\frac{1}{n^3}\right)U_i(t+t_0)U_j(t_0) = \\
&= \frac{1}{n^2}\sum_{i=1}^{n}U_i(t_0)U_i(t+t_0) - \frac{1}{n^3}\left[\left(\sum_{i=1}^{n}U_i(t_0)\right)\left(\sum_{j=1}^{n}U_j(t+t_0)\right)\right] \qquad (24)
\end{aligned}$$

Now, let us show that the sum of directional derivatives $U_i(t)$ is 0. First, gradient vector D is defined as

$$D = \begin{pmatrix} D_1 \\ \vdots \\ D_n \end{pmatrix} \qquad \text{where} \quad D_i = \frac{\partial}{\partial p_i}\hat{\theta}_{\mathrm{RSF}}(\boldsymbol{P}, t)\Big|_{\boldsymbol{P}=\boldsymbol{P}^0}$$

Therefore, according to the definition of directional derivative $U_i(t)$ in (10) could be expressed as

$$U_i(t) = (\boldsymbol{\delta}_i - \boldsymbol{P}^0)\cdot D$$

where $\boldsymbol{\delta}_i$ has 1 at the i^{th} position and 0 at all others. Rewriting $U_i(t)$ using

knowledge about the vectors' structure gives

$$U_i(t) = \left(\underbrace{-\frac{1}{n}, \ldots, -\frac{1}{n}}_{\text{i - 1}}, 1 - \frac{1}{n}, -\frac{1}{n}, \ldots, -\frac{1}{n} \right) \cdot \begin{pmatrix} D_1 \\ \vdots \\ D_n \end{pmatrix} =$$

$$= \sum_{j \neq i} \left(-\frac{1}{n} \right) \cdot \frac{\partial}{\partial p_j} \hat{\theta}_{\mathrm{RSF}}(\boldsymbol{P}, t) \Big|_{\boldsymbol{P}=\boldsymbol{P}^0} +$$

$$+ \left(1 - \frac{1}{n} \right) \cdot \frac{\partial}{\partial p_i} \hat{\theta}_{\mathrm{RSF}}(\boldsymbol{P}, t) \Big|_{\boldsymbol{P}=\boldsymbol{P}^0}$$

It could be seen that if the sum of $U_i(t)$ is considered then a factor next to every partial derivative will consist of a sum of one summand $\left(1 - \frac{1}{n}\right)$ and all others being $\left(-\frac{1}{n}\right)$. Therefore, the following could be written

$$\sum_{i=1}^{n} U_i(t) = \sum_{i=1}^{n} \left(\left(1 - \frac{1}{n} \right) + \sum_{j=1}^{n-1} \left(-\frac{1}{n} \right) \right) \cdot \frac{\partial}{\partial p_i} \hat{\theta}_{\mathrm{RSF}}(\boldsymbol{P}) \Big|_{\boldsymbol{P}=\boldsymbol{P}^0} =$$

$$= \sum_{i=1}^{n} \left(\left(1 - \frac{1}{n} \right) - \left(\frac{n-1}{n} \right) \right) \cdot \frac{\partial}{\partial p_i} \hat{\theta}_{\mathrm{RSF}}(\boldsymbol{P}) \Big|_{\boldsymbol{P}=\boldsymbol{P}^0} = 0$$

Thus by substituting zeroes for the sums of directional derivatives in (24) we get

$$\widehat{\mathrm{cov}}_{\mathrm{Bias}} \left[\hat{R}^{\mathcal{V}}(t + t_0), \hat{R}^{\mathcal{V}}(t_0) \right] = \frac{1}{n^2} \sum_{i=1}^{n} U_i(t_0) U_i(t + t_0) \tag{25}$$

Following the same steps as in (Efron et al., 2014) it could be shown that

$$U_i(t) = n \widehat{\mathrm{Cov}}_i(t)$$

which proves (18). Bias from (19) of $\widehat{\mathrm{cov}}_{\mathrm{Bias}} \left[\hat{R}^{\mathcal{V}}(t + t_0), \hat{R}^{\mathcal{V}}(t_0) \right]$ estimate can be found as follows

$$\mathrm{Bias} \;=\; E \left[\widehat{\mathrm{cov}}_{\mathrm{Bias}} \left[\hat{R}^{\mathcal{V}}(t + t_0), \hat{R}^{\mathcal{V}}(t_0) \right] \right] \;-\; \mathrm{cov} \left[\hat{R}^{\mathcal{V}}(t + t_0), \hat{R}^{\mathcal{V}}(t_0) \right] \tag{26}$$

Here, $\mathrm{cov} \left[\hat{R}^{\mathcal{V}}(t + t_0), \hat{R}^{\mathcal{V}}(t_0) \right]$ is a covariance between reliabilities when number

of trees in the forest $B \to \infty$. One can rewrite (26) as

$$\begin{aligned}\text{Bias} &= \sum_{j=1}^{n} \left(E\left[\widehat{\text{Cov}}_i(t_0)\widehat{\text{Cov}}_i(t+t_0)\right] - \text{Cov}_i(t_0)\text{Cov}_i(t+t_0)\right) = \\ &= \sum_{j=1}^{n} \left(E\left[\widehat{\text{Cov}}_i(t_0)\widehat{\text{Cov}}_i(t+t_0)\right] - E\left[\widehat{\text{Cov}}_i(t_0)\right] E\left[\widehat{\text{Cov}}_i(t+t_0)\right]\right) = \\ &= \sum_{j=1}^{n} \text{cov}\left[\widehat{\text{Cov}}_i(t_0); \widehat{\text{Cov}}_i(t+t_0)\right] \quad (27)\end{aligned}$$

Taking into account expression for $\widehat{\text{Cov}}_i(t_0)$ in (14) bias becomes

$$\begin{aligned}\text{Bias} &= \sum_{j=1}^{n} \text{cov}\left[\frac{1}{B}\sum_{b=1}^{B}(y^*_{ij}-1)(\hat{R}^{\mathcal{V}}_b(t_0) - \hat{R}^{\mathcal{V}}(t_0)); \right. \\ &\qquad \left. \frac{1}{B}\sum_{b=1}^{B}(y^*_{bj}-1)(\hat{R}^{\mathcal{V}}_b(t+t_0) - \hat{R}^{\mathcal{V}}(t+t_0))\right] = \\ &= \frac{1}{B^2}\sum_{j=1}^{n}\sum_{i=1}^{B}\sum_{b=1}^{B}\left(E\left[(y^*_{bj}-1)(y^*_{ij}-1)(\hat{R}^{\mathcal{V}}_i(t_0) - \hat{R}^{\mathcal{V}}(t_0))(\hat{R}^{\mathcal{V}}_b(t+t_0) - \hat{R}^{\mathcal{V}}(t+t_0))\right] - \right. \\ &\quad \left. - E\left[(y^*_{ij}-1)(\hat{R}^{\mathcal{V}}_i(t_0) - \hat{R}^{\mathcal{V}}(t_0))\right] \cdot E\left[(y^*_{bj}-1)(\hat{R}^{\mathcal{V}}_b(t+t_0) - \hat{R}^{\mathcal{V}}(t+t_0))\right]\right) \quad (28)\end{aligned}$$

Assuming that original sample $\boldsymbol{Y}$ is large enough it becomes possible to suppose that $\hat{R}^{\mathcal{V}}_i(t)$ and y^*_{ij} are independent. This make simplifications in bias computations.

$$\begin{aligned}\text{Bias} = \frac{1}{B^2}\sum_{j=1}^{n}\sum_{i=1}^{B}\sum_{b=1}^{B}\big(& E\left[(y^*_{bj}-1)(y^*_{ij}-1)\right] \cdot \\ & E\left[(\hat{R}^{\mathcal{V}}_i(t_0) - \hat{R}^{\mathcal{V}}(t_0))(\hat{R}^{\mathcal{V}}_b(t+t_0) - \hat{R}^{\mathcal{V}}(t+t_0))\right] - \\ & - E\left[(y^*_{ij}-1)\right] E\left[(\hat{R}^{\mathcal{V}}_i(t_0) - \hat{R}^{\mathcal{V}}(t_0))\right] \cdot \\ & E\left[(y^*_{bj}-1)\right] E\left[(\hat{R}^{\mathcal{V}}_b(t+t_0) - \hat{R}^{\mathcal{V}}(t+t_0))\right]\big) \quad (29)\end{aligned}$$

Random variable y_{ij}^* has the following properties

$$E\left[(y_{ij}^*-1)(y_{bj}^*-1)\right]=\operatorname{cov}\left[y_{ij}^*;y_{bj}^*\right]=\frac{1}{n}\to 0,$$
$$n\to\infty, b\neq j$$
$$E\left[y_{ij}^*-1\right]=E\left[y_{bj}^*-1\right]=0$$
$$E\left[(y_{ij}^*-1)^2\right]=\operatorname{var}\left[y_{ij}^*\right]=1-\frac{1}{n}\to 1,$$
$$n\to\infty, b=j$$

Therefore, if it is assumed that size of sample $n\to\infty$ and n tends to infinity faster than B, bias becomes

$$\text{Bias}=\frac{1}{B^2}\sum_{j=1}^{n}\sum_{i=1}^{B}\Big(E\left[(y_{ij}^*-1)^2\right]\cdot$$
$$E\left[(\hat{R}_i^{\mathcal{V}}(t_0)-\hat{R}^{\mathcal{V}}(t_0))(\hat{R}_i^{\mathcal{V}}(t+t_0)-\hat{R}^{\mathcal{V}}(t+t_0))\right]\Big)=$$
$$=\frac{n}{B^2}\sum_{i=1}^{B}(\hat{R}_i^{\mathcal{V}}(t_0)-\hat{R}^{\mathcal{V}}(t_0))(\hat{R}_i^{\mathcal{V}}(t+t_0)-\hat{R}^{\mathcal{V}}(t+t_0)) \quad (30)$$

The expression in (30) is similar to the bias correction for RF model found in (Efron et al., 2014) and presented in (12). □

4.3 Analysis of the IJ covariance estimate

Theorem 1 summarizes the expressions for the covariance estimate of the lifetime function. This section will explore and highlight some properties of the variance estimate. First, consequences of the bias correction is analyzed and the importance of the covariance estimate $\widehat{\operatorname{cov}}\left[\hat{R}^{\mathcal{V}}(t+t_0),\hat{R}^{\mathcal{V}}(t_0)\right]$ is demonstrated. Then, model selection based on confidence bands is discussed.

When $\widehat{\operatorname{cov}}\left[\hat{R}^{\mathcal{V}}(t+t_0),\hat{R}^{\mathcal{V}}(t_0)\right]$ and bias are estimated, it is possible to plot confidence bands for an estimate of the lifetime function $\mathcal{B}^{\mathcal{V}}(t,t_0)$. Fig. 3 shows a 95% confidence band for 4 vehicles from the validation set with a Gaussian assumption for the lifetime function estimate. The RSF model used for the figure had 1000 trees. To motivate the need in estimating the $\widehat{\operatorname{cov}}\left[\hat{R}^{\mathcal{V}}(t+t_0),\hat{R}^{\mathcal{V}}(t_0)\right]$ and the estimator bias, three types of confidence bands are plotted. Blue dashed curves are 95% confidence bands computed using the variance from (16) where biased IJ variance estimates are used, i.e., values $\operatorname{var}\left[\hat{R}^{\mathcal{V}}(t+t_0)\right]$, $\operatorname{var}\left[\hat{R}^{\mathcal{V}}(t_0)\right]$, and $\operatorname{cov}\left[\hat{R}^{\mathcal{V}}(t+t_0),\hat{R}^{\mathcal{V}}(t_0)\right]$ are biased. It can be seen that when biased estimates are used in (16), the confidence bands become conservative. The black curves are 95% confidence bands computed using the variance from

(16) with the unbiased IJ variance estimates $\text{var}\left[\hat{R}^{\mathcal{V}}(t+t_0)\right]$ and $\text{var}\left[\hat{R}^{\mathcal{V}}(t_0)\right]$ of reliabilities and assumption that values of $\hat{R}^{\mathcal{V}}(t)$ are independent at time point t and $t+t_0$, i.e., $\widehat{\text{cov}}\left[\hat{R}^{\mathcal{V}}(t+t_0), \hat{R}^{\mathcal{V}}(t_0)\right] = 0$. The red curves are 95% confidence bands computed using variance from (16) with the unbiased IJ variance estimates $\text{var}\left[\hat{R}^{\mathcal{V}}(t+t_0)\right]$ and $\text{var}\left[\hat{R}^{\mathcal{V}}(t_0)\right]$ of reliabilities and estimated $\text{cov}\left[\hat{R}^{\mathcal{V}}(t+t_0), \hat{R}^{\mathcal{V}}(t_0)\right]$. Fig. 3 shows that for three vehicles out of four black and red curves are close to each other, however, for a vehicle in top right corner they differ significantly. This indicates the importance in finding $\widehat{\text{cov}}\left[\hat{R}^{\mathcal{V}}(t+t_0), \hat{R}^{\mathcal{V}}(t_0)\right]$ and its bias.

Confidence bands can be used for the model selection. For example, an estimate for lifetime functions together with confidence bands for two RSF models, one with 100 trees and one with 1000 trees, are presented in Fig. 4. It is not surprising that more trees improves the variance of the predictor. However, let us consider model selection based on the available error metric. One of the metrics measuring error available in the RSF framework is the error rate (Ishwaran et al., 2008). It relies on the Concordance index which counts prediction as erroneous when for two randomly selected vehicles the shorter survival time has worse predicted value of survival function. The error rate curve for the given example starts to converge after about 100 trees, and then it is tempting to stop increasing the number of trees in the RSF model. However, it is evident from Fig. 4 that the quality of the prediction in the case of 1000 trees is significantly better than in the case of 100 trees, because confidence bands are narrower and estimates of the lifetime functions also differ. The experiment shows that adding confidence bands to the predictor helps to find better model than the one created by relying only on the error rate values.

It is evident from the results above that the unbiased covariance estimates give less conservative variance estimate of the lifetime function and, in addition, confidence bands can be used as a complimentary tool, for example, to the error rate for model selection.

5 Synthetic data set study

A main problem with the vehicle database is that actual battery degradation profiles are not known and therefore it is hard to validate lifetime estimates and confidence bands in, for example, Fig. 3. To answer the question how trustworthy the results received in Section 4 are, a synthetic data set is considered where the underlying degradation profiles are known, controllable, and with similar properties as the vehicle data set.

The generated synthetic data has 6 variables and 1000 vehicles. One variable is important for prognosis as it controls the degradation of the battery. The other five variables are noise in the sense that they do not influence the battery

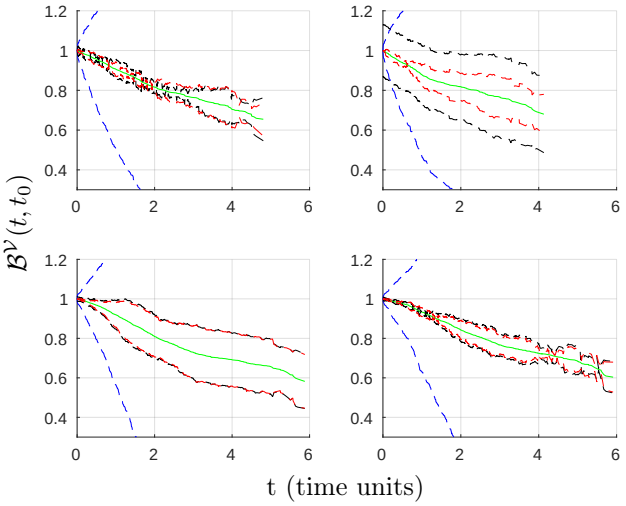


Figure 3: IJ variance estimates of lifetime function for 4 vehicles from validation set. Green curve is an estimate of lifetime function $\mathcal{B}^{\mathcal{V}}(t, t_0)$. Blue curves are 95% confidence bands computed using variance from (16) with biased IJ variance estimates of covariance of reliabilities. Black curves are 95% confidence bands computed using variance from (16) with unbiased IJ variance estimates of covariance of reliabilities and assumption that values of $\hat{R}^{\mathcal{V}}(t)$ are independent at time point t and $t + t_0$. Red curves are 95% confidence bands computed using variance from (16) with unbiased IJ variance estimates of covariance of reliabilities.

degradation. The battery degradation is controlled by varying the hazard rate (Cox and Oakes, 1984), i.e., the probability of instantaneous failure at time t, according to the one important variable. Expected lifetime of batteries with the selected nominal hazard rate is set to 10 years and it is assumed that the important variable v_1 has an impact on the battery hazard rate h as

$$h = \begin{cases} 1 \cdot h_0, & \text{if } v_1 = 1 \\ 1.5 \cdot h_0, & \text{if } v_1 = 2 \\ 2.5 \cdot h_0, & \text{if } v_1 = 3 \\ 2.9 \cdot h_0, & \text{if } v_1 = 4 \\ 3.4 \cdot h_0, & \text{if } v_1 = 5 \end{cases} \tag{31}$$

where $h_0 = \frac{1}{10}$ is the nominal hazard rate. Censoring is added to the database at a level comparable with the one in the real database. Vehicles are uniformly distributed among the five classes, meaning, that each class has about 200 vehicles.

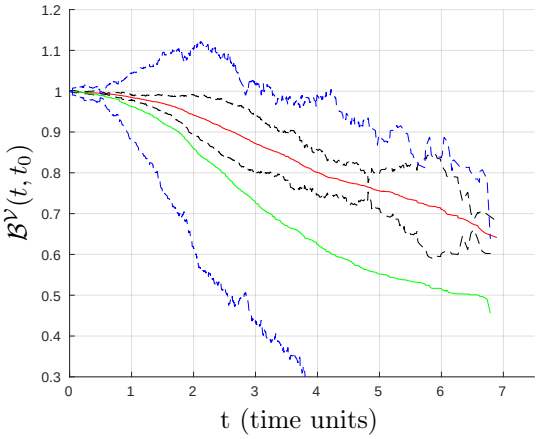


Figure 4: Estimate of the lifetime function $\mathcal{B}^{\mathcal{V}}(t, t_0)$, green curve corresponds to the model with 100 trees and red to the model with 1000 trees, with the 95% confidence bands, blue curves correspond to the model with 100 trees and black curves to the model with 1000 trees.

Here, in contrast to the vehicle data set, the class of each vehicle is known and then it is possible to compute the Kaplan-Meier estimate, $R(t)$, of the reliability function, which is the maximum likelihood estimator, for every class together with confidence bands computed using the standard Greenwood formula. This corresponds to the estimates based on an ideal vehicle classifier, i.e., that perfectly separates vehicles into the 5 defined classes. Aforementioned estimates for the third class of the vehicles are presented in Fig. 2. Now, let us compare them to estimates from the RSF model. There are two main parameters when building the model, minimal node size and number of trees in the forest. The parameters are selected or changed while constructing the forest and all other options are set to their default (Ishwaran and Kogalur, 2007). Minimal node size was selected to be 200 to allow a fair comparison to the maximum likelihood estimates. Several options are tried for number of trees B, namely, $B \in \{100, 500, 1000, 2000, 5000, 10000\}$.

First, let us make prediction for one of the vehicles in the validation set belonging to the third class. Fig. 5 shows the predictions from the forest of 1000 trees together with maximum-likelihood estimates. The magenta curve is the true reliability, green and blue curves are the Kaplan-Meier estimate and 95% confidence bands respectively, and the RSF reliability and 95% confidence bands based on IJ variance estimate are black and red curves respectively. It can be seen from the Fig 5 that the confidence bands based on IJ variance estimate is close to the the ones given by the maximum likelihood estimate, Greenwood formula. To show how variance estimate varies with different number of trees,

variance estimate is computed for a vehicle at time point $t = 0.2$ and $t = 0.8$ for the various numbers of the trees B. The result is presented in Fig. 6 showing that the variance estimates, red and blue curves, converge to some non-zero positive. Green and black lines are variances received using Greenwood formula, i.e., computed under an ideal classifier assumption. As it can be seen, the variance estimate at time point $t = 0.2$, blue curve, is very close to the Greenwood estimate. Variance estimate at time point $t = 0.8$, red curve, is biased with respect to the Greenwood estimate which is suspected due to censoring.

It should be noticed that when the number of the trees in the forest is small, around 100 trees, the IJ variance and bias estimates have significant variances which make it possible that the IJ variance estimate can be negative due to the additive bias and small value of variance of the predictor. For example, when computing the IJ variance estimate at time point $t = 0.8$ for the case of $B = 100$ trees, it is negative for a given model realization. A question may arise what to do in this situation. For now, absolute value of the variance is taken as an estimate of the true variance. It is possible to use some other value in this case, for instance variance not defined, to show that we are uncertain about the variance estimate. However, it is mentioned above that the negative IJ variance estimate can happen not only due to the small number of trees in the forest, but also when the true variance of a predictor is small. Therefore, taking the absolute value of the variance estimate could give an idea about the true variance and several experiments with the RSF model corroborates this.

As a conclusion, it is illustrated that IJ variance estimate is a good tool in finding the true variance of a predictor and variance estimate gives more relevant information about the model than the error rate.

6 Performance evaluation with several metrics

Every prognostic model should be evaluated such that their predictive performance is known. As mentioned, this is problematic since the output from the forest model is a survival or reliability function and there is no record of their true values in the data set. In a pure classification or regression problem there are established metrics to evaluate performance, however, this is not the case for survival analysis. A metric to use in the case of the RSF framework is an error rate based on the concordance index (Ishwaran et al., 2008). A question is if this error rate is descriptive enough. For example, the authors in (Moradian et al., 2016) conclude that it is possible that the error rate is not an appropriate performance measure. The example given below supports this observation and shows that with similar values of error rates models give significantly different survival curves.

The example relies on simulated data similar to the one used in Section 5. Degradation of the battery is controlled by the hazard rate h_0 which corresponds

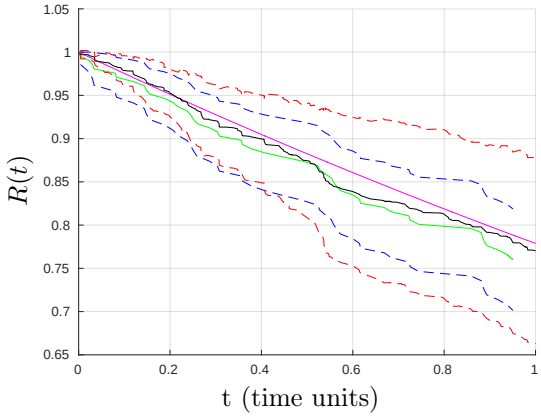


Figure 5: Reliability with confidence bands. Theoretical values vs estimates from the RSF. Magenta curve is the true reliability curve, green and blue curves are the Kaplan-Meier estimate and 95% confidence bands with Gaussian assumption computed using Greenwood formula, black and red curves are the RSF estimate of the reliability and 95% confidence bands with Gaussian assumption estimated using IJ technique.

to 10 years mean battery life. As in the previous example it is assumed that there is one important variable v_1 which influences hazard rate h_0 such that three classes of vehicles exist with different degradation profiles corresponding to the new hazard rate h

$$h = \begin{cases} 1 \cdot h_0, & \text{if } v_1 = 1 \\ 2 \cdot h_0, & \text{if } v_1 = 2 \\ 3 \cdot h_0, & \text{if } v_1 = 3 \end{cases} \tag{32}$$

Two models with 2 and 100 noisy variables are considered where the censoring rate is about 80% which is similar to the value from the example in Section 5. RSF models are trained on 1000 vehicles with minimal node size chosen to be 200 and number of trees in the forest 1000. Fig. 7 shows the comparison of the predicted survival curves from RSF model, dashed blue curves, with theoretical values, red curves, for three randomly chosen vehicles that were not included in the training sets. It is evident from the left plot in Fig. 7 that, as expected, predictions for the model with only 2 noisy variables are significantly better than for the model with 100 noisy variables, right plot in Fig. 7. However, values of the error rate for the both models are 0.4097 and 0.4270 for two models respectively. The error rates are close for both models, therefore, one would expect that forest outputs would be similar as well, but this is clearly not the case. Thus, new evaluation techniques are needed to be able to say more about predictive performance of the model.

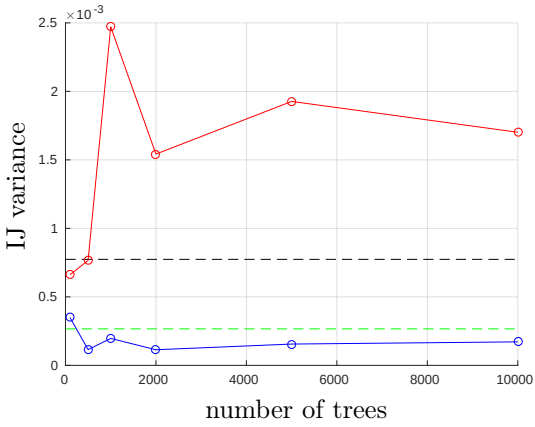
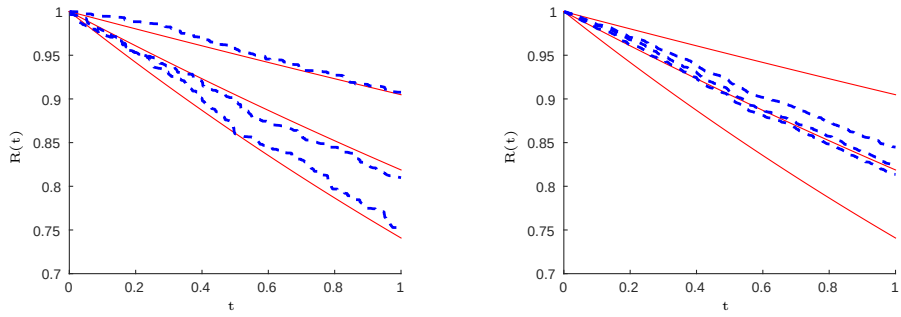


Figure 6: Reliability with confidence bands. Theoretical values vs estimates from the RSF. Green and black lines are the true variances, the Greenwood estimates, at time point $t = 0.2$ and $t = 0.8$ respectively. Red and blue curves are the IJ variance estimates at time point $t = 0.2$ and $t = 0.8$ respectively for different number of trees B.



(a) 2 noisy variables

(b) 100 noisy variables

Figure 7: RSF predictions for two models with different number of noisy variables. Red curves are the theoretical reliabilities for three classes and blue dashed curves are the outputs from the RSF model.

6.1 Performance analysis of predictive model for battery data

A vehicle used the same way should never leave its class of similar vehicles, however, with year or mileage variables present in the database, the model might be dominated by age effects which is not the intention and could possibly mask

the effects of different vehicle usage. The problem of using accumulative variables like age or mileage is addressed by the authors in (Frisk and Krysander, 2015). It was suggested that instead of using accumulative variables directly it is better to preprocess them first. For example, there are two accumulative variables in the current data set, namely, age and mileage. First, a new variable mileage per day is created and, then, two models are considered, namely, a model based on all variables except the accumulative ones and another model where the variable mileage per day has been added.

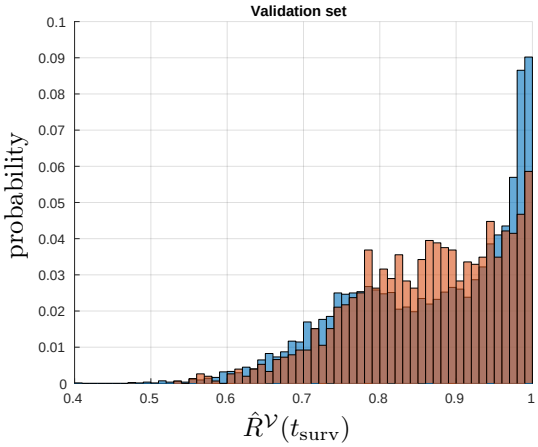


Figure 8: Histograms of $\hat{R}^{\mathcal{V}}(t_{\text{surv}})$ for the failed vehicles, red bins, and censored vehicles, blue bins. Data set without mileage per day.

One way to evaluate performance of a predictor is to look at values of $\hat{R}^{\mathcal{V}}(t_{\text{surv}})$ survival/reliability curves at the time of either failure or censoring, and see how predictions of the two classes of vehicles differ. First, the data set is split into two parts, 2/3 of 30,000 vehicles are used for training and 1/3 for validation. Only 30,000 vehicles are randomly chosen out of 56,163 for training due to the limitations of computational resources. Then, the RSF model is grown with 1,000 trees in the forest with minimal node size of 200 on the training set while validation set is used for performance evaluation. Fig. 8 and Fig. 9 show the histograms of reliabilities for failed, red color, and censored vehicles, blue color, for the two models with and without the variable mileage per day. It is seen that the histograms of the two classes of vehicles are different on the one hand, however, have a big overlap on the other, therefore not much can be said about the performance of the predictor.

Another approach for performance evaluation is to plot lifetime functions $\hat{\mathcal{B}}^{\mathcal{V}}(t;t_0)$ where $t_0 = t_{\text{surv}}$ and observe how they differ between the two classes of vehicles. Result of prediction for 100 randomly selected vehicles from failed and censored classes is depicted in Fig. 10 and Fig. 11 where red curves correspond

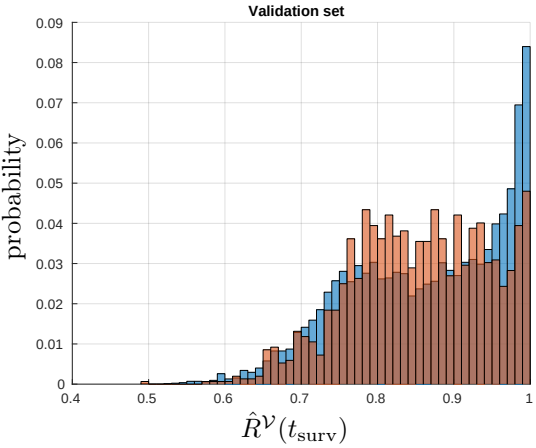


Figure 9: Histograms of $\hat{R}^{\mathcal{V}}(t_{\text{surv}})$ for the failed vehicles, red bins, and censored vehicles, blue bins. Mileage per day variable is included.

to the failed class and blue curves to the censored. What can be seen in the figures is that on average predictions for both classes of vehicles are different. However, overlap between two classes is big, therefore, it would be good to find more informative measure of performance. Instead of considering predictions of reliability and lifetime curves at time t_{surv} when vehicle is either censored or failed, let us consider the cross section of the respective curves at some fixed time point t which is similar for all vehicles.

Results of reliability histograms after 3 time units for the two classes of vehicles and two models are shown in Fig. 12 and Fig. 13. This particular time point was selected to allow the batteries to be in operation for some time, so their different usage patterns influence the degradation and it is expected that the predictions for the two classes should differ. It is seen that the difference between the histograms of the two classes is more clear than in Fig. 8 and Fig. 9. There is still an overlap between the two histograms and one would expect them to be completely separated in the ideal case, however, it is possible that some of the censored vehicles are really close to failure, but leave the study before failure and the problem of the battery is not recorded. Therefore, left tails of the censored histograms with small values of reliabilities could be not a mistake of the algorithm, but a correct indication that a vehicle belongs to the failed class. On the other hand, group of vehicles from the failed class with large values of reliabilities, right tails of failed histograms, experience problem with battery due to the reasons which are not present in the current data set. Thus, it is impossible for the algorithm to see that the vehicle has potential problems with a battery. One other thing to notice is that for the model that includes mileage per day variable there is a peak in the failed histogram coinciding with the peak

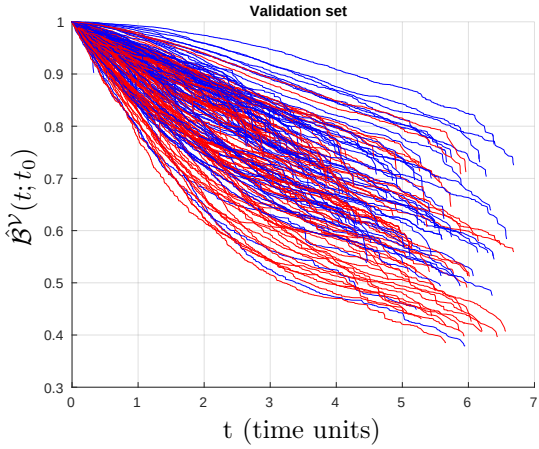


Figure 10: Lifetime functions $\hat{\mathcal{B}}^{\mathcal{V}}(t;t_0)$ computed for 100 randomly selected vehicles from censored and failed classes on validation set. Data set without mileage per day variable.

of censored histogram, see Fig. 13. For now, it is unclear what it represents, however the results are affected by including or excluding the variable.

Histograms of the lifetime functions $\hat{\mathcal{B}}^{\mathcal{V}}(t;t_0)$ for two models at time point $t = t_{\text{surv}} + 1$ time unit and $t_0 = t_{\text{surv}}$ are presented in Fig. 14 and Fig. 15. Our industrial partner Scania CV is, say, interested in predictions up to 1 time unit to be used in their maintenance planner, therefore, only a time point within 1 time unit in the future is selected. Separation between histograms for failed and censored classes is not so distinctive as in the case of the reliability curves, nevertheless, similar behavior is seen for the model with the mileage per day variable where the peak of the histogram of the failed batteries coincides with the peak of the censored one, see Fig. 15. In addition, the histogram of the failed batteries for the model with mileage per day are skewed more to the right than for the model without the variable.

6.2 Lifetime prognosis for vehicles with similar mileage

It is natural to do maintenance based on age and mileage where batteries which reached the predefined period of their life or vehicle operated predefined amount of miles considered as the ones to be replaced. To demonstrate that the RSF framework partition vehicles into classes based on usage profiles and not simply on age and mileage, vehicles with similar mileage are selected. The base value of mileage m is selected and the interval plus-minus 5% from the base value m is considered. From this set of vehicles with similar mileage, vehicles with similar age is selected. There are 84 vehicles satisfying the stated requirement

Table 1: Values of variables selected among 50 most important given by VIMP for vehicle V_1 with best prognosis, V_2 with the worst prognosis for the model excluding mileage per day and vehicle V_3 with the worst prognosis for the model including mileage per day

Variables	V_1	V_2	V_3
Country	0	1	2
Bed type	0	2	0
Ambient temperature bin 3	0	1.26	0.58
Atmospheric pressure bin 7	14.23	1	4.33
Atmospheric pressure bin 8	1	4.1	3.32
Battery SOC vs Poweroff 2d bin 17	40.84	18.43	1
Battery voltage bin 6	0	1.46	82.22
Battery voltage bin 7	0	1.88	0.37
Fuel consumption vs speed 2d bin 4	3.06	1.7	1
Fuel consumption vs speed 2d bin 5	4.2	1.82	1
Fuel consumption vs speed 2d bin 6	4.1	1.50	1
Fuel consumption vs speed 2d bin 7	4.41	1.46	1
Fuel consumption vs speed 2d bin 8	3.31	1.23	1
Fuel consumption vs speed 2d bin 15	1	5.79	15.17
Vehicle speed bin 0	4.14	4.07	1
Vehicle speed bin 1	1.96	1.05	1
Vehicle speed bin 2	3.05	1	1.76
Vehicle speed bin 6	1	6.45	8.06
Vehicle speed bin 7	1	12.15	38.75
Engine Load 2d bin 30	7.91	1	1.67
Engine Load 2d bin 31	15.8	1.4	1
Engine Load 2d bin 32	76	4	1
Engine Load 2d bin 41	4.71	6.58	1
Engine Load 2d bin 42	12.25	1	1.85

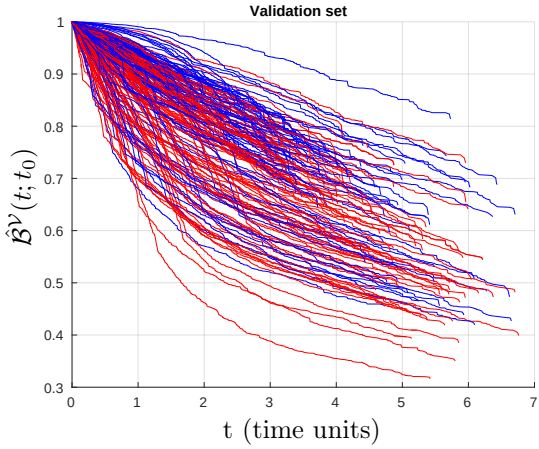


Figure 11: Lifetime functions $\hat{\mathcal{B}}^{\mathcal{V}}(t;t_0)$ computed for 100 randomly selected vehicles from censored and failed classes on validation set. Mileage per day variable is included.

on mileage in the validation set. The lifetime function estimates $\hat{\mathcal{B}}^{\mathcal{V}}(t;t_0)$ with $t_0 = t_{\text{surv}}$ for the selected vehicles are presented in Fig. 16 and Fig. 17 showing the prediction for model with and without mileage per day variable respectively. First notice that the difference between the best and the worst predictions is significant which shows that how vehicles are used is important. Next, three vehicles V_1, V_2 and V_3 are selected from the set of the vehicles with similar mileage. The vehicle V_1 corresponds the best prognosis and is the same vehicle for the both models, when vehicles V_2 and V_3 with the worst prognosis are different for the two models. Age of batteries for the vehicles V_1, V_2 and V_3 are 1.3, 0.83 and 0.98 time units respectively where the vehicle V_1 with the best prognosis lived the longest among three vehicles, therefore, vehicle usage pattern plays a significant role.

Table 1 shows selected variables for three vehicles V_1, V_2 and V_3 among 50 most important variables for the prediction obtained using VIMP (Ishwaran et al., 2008), Variable IMPortance, with the most important variables at the top. Only variables that have different values for three vehicles are left among 50 most important. VIMP is the framework integrated in the RSF method and assigns importance value for every variable. First, vehicles operated in the different countries that can explain the difference in the degradation profiles of the batteries as climate, quality of roads can vary. Bin 3 of the ambient temperature histogram appears important for the prediction. This bin corresponds to operation of a vehicle under the low temperatures. Vehicles V_2 and V_3 have operated more time under the low temperatures which corroborates the fact that the vehicles have worse degradation prediction than the vehicle V_1. Two bins of

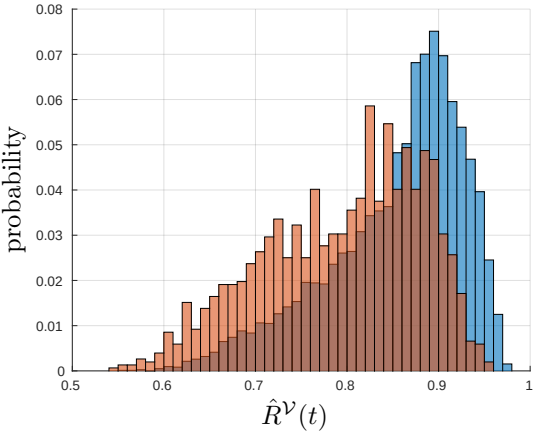


Figure 12: Histograms of $\hat{R}^{\mathcal{V}}(t)$ for the failed vehicles, red bins, and censored vehicles, blue bins, at time point $t = 3$ time units. Data set without mileage per day variable.

the atmospheric pressure histogram are important, namely, bins 7 and 8. Vehicle V_1 has much bigger value in the 7th bin compared to the values for the vehicles V_2 and V_3, at the same time has much lower value in the 8th bin. Two bins from the battery voltage histogram is also important. They correspond to the operation of the battery under high voltage. It can be seen that vehicles with worse prediction operated more under high voltage which can be considered as counterintuitive at first. However, it is possible that the generator that charges the battery has malfunctions that lead to overcharging and faster degradation of the battery. Overall, there is a significant amount of the variables in Table 1 that indicate different usage of the vehicles. This fact gives positive signs for using the RSF method as a predictive tool.

It can also be seen that predictions for two models are different, namely, lifetime function estimates for the model with mileage per day variable are comprised of two types of curves. One is convex and another is concave with a joint point for both curves between 1 and 2 time units. Taking into account that batteries by themselves are of age 1 to 2 time units, the joint point for lifetime function estimates lie near the peak of distribution for the failed vehicles from Fig. 1.

Now, consider the lifetime function estimates which correspond to the best, worst and intermediate prediction for two models. They can be found in Fig. 18 and Fig. 19. The lifetime function estimates correspond to the solid lines in the figures, and dashed curves are 95% confidence bands with Gaussian assumption and IJ variance estimate from Section 4. It can be seen that confidence bands for the model with mileage per day variable are wider than for the model without

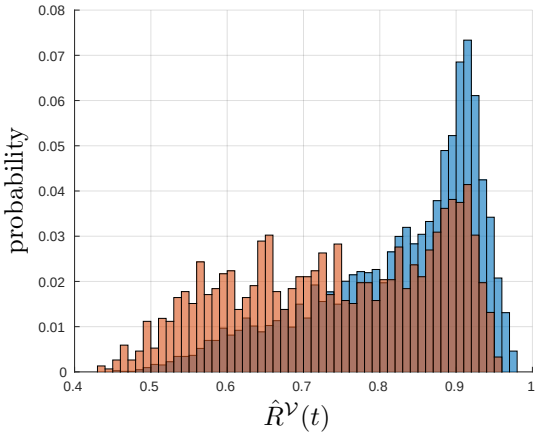


Figure 13: Histograms of $\hat{R}^{\mathcal{V}}(t)$ for the failed vehicles, red bins, and censored vehicles, blue bins, at time point $t = 3$ time units. Mileage per day variable is included.

which is a surprising result. Intuitively the more variables the better predictions, however, the result shows opposite. It means that relying on usage profile rather than on time related variables would give more accurate predictor for the given data. More studies should be carried out to see if incorporation of time related variables can give better performance.

As a conclusion, RSF model applied to the given data set gives on average different predictions for failed and censored class of vehicles, results show that vehicle usage profile is important for predicting the degradation of a battery and that there are indication not to include accumulative variables into the training RSF model as it increases uncertainty of the predictor. It is impossible to determine and validate a failure time for a battery with the given data set, because only one snapshot of data is available for every vehicle. However, it will be done in the future when data set is augmented with several snapshots per vehicle.

7 Conclusion

It is shown in the paper that the RSF model can be applied to the static data, i.e., one snapshot only per vehicle, in the data set. It is different from the data sets being used in (Daigle and Goebel, 2011; Medjaher et al., 2012; Zhao et al., 2015), and therefore a new approach is needed. The RSF model output is the estimate of the reliability function which can be used to compute the lifetime function estimate (2). The confidence bands of the lifetime function estimate (2) are computed using the Infinitesimal Jackknife (IJ) variance estimate approach

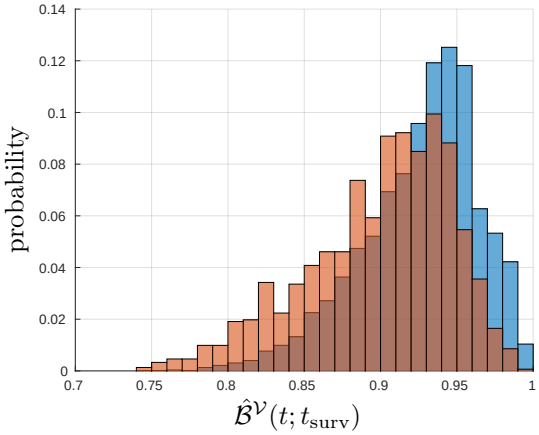


Figure 14: Histograms of lifetime function estimates $\hat{\mathcal{B}}^{\mathcal{V}}(t; t_{\text{surv}})$ for the failed vehicles, red bins, and censored vehicles, blue bins, at $t = 1$ time unit point in future from t_{surv}. Data set without mileage per day variable.

and its properties are analyzed. First, confidence bands can be used for the model selection, for example, it is shown that the prediction for the forest model with 1000 trees is significantly better than for the model with 100 trees in terms of confidence bands, however, in terms of the standard error rate, the two models are similar. Second, IJ variance estimate starts to converge for forest with 1000 trees or larger which means that the variance estimate of the predictor with 1000 trees is appropriate. Models with and without accumulative variables give different results and currently it seems that excluding accumulative variables gives better results based on the fact that the confidence bands become narrower. Performance evaluation is done and it has been shown that prediction for a censored and failed vehicle is different. In general, the battery lifetime function can be used to schedule and optimize the cost of the battery replacement which leads to more flexible maintenance.

Acknowledgment

The authors acknowledge Scania and FFI (Vehicle Strategic Research and Innovation) for sponsorship of this work.

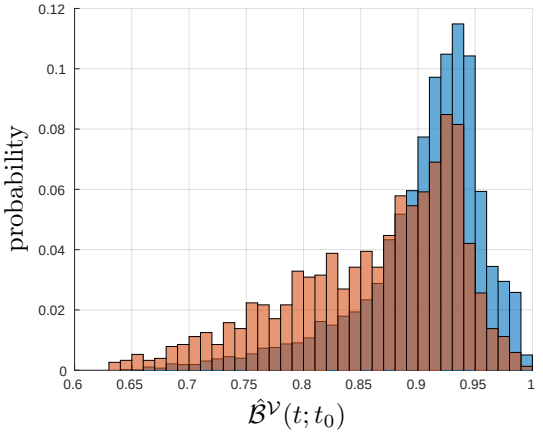


Figure 15: Histograms of lifetime function estimates $\hat{\mathcal{B}}^{\mathcal{V}}(t;t_{\text{surv}})$ for the failed vehicles, red bins, and censored vehicles, blue bins, at $t = 1$ time unit point in future from t_{surv}. Mileage per day variable is included.

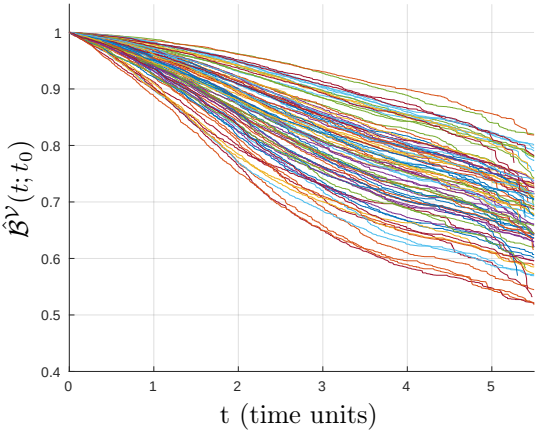


Figure 16: Lifetime functions estimates $\hat{\mathcal{B}}^{\mathcal{V}}(t;t_0)$ for 84 vehicles which mileage values are in plus-minus 5% interval around base mileage value m and age of batteries are within 1 to 2 time unit interval. Model does not contain mileage per day variable.

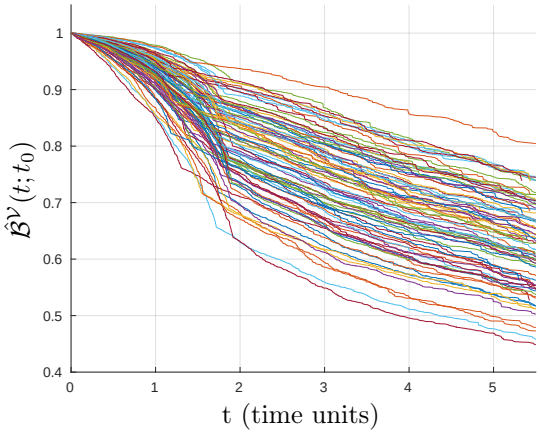


Figure 17: Lifetime functions estimates $\hat{\mathcal{B}}^{\mathcal{V}}(t;t_0)$ for 84 vehicles which mileage values are in plus-minus 5% interval around base mileage value m and age of batteries are within 1 to 2 time unit interval. Model contains mileage per day variable.

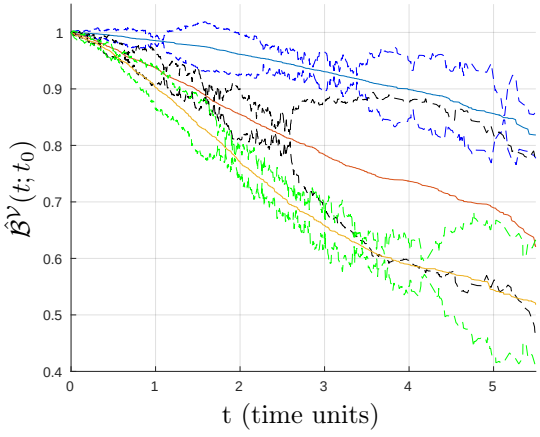


Figure 18: Lifetime function estimates $\hat{\mathcal{B}}^{\mathcal{V}}(t;t_0)$ for the best, worst and intermediate predictions from Fig. 16, solid lines, together with 95% confidence bands with Gaussian assumption and IJ variance estimate, dashed curves. Without mileage per day model.

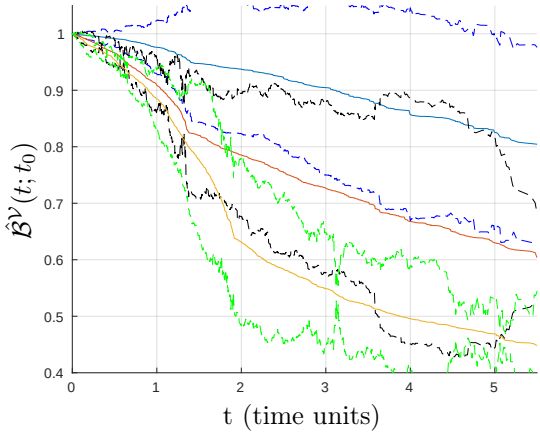


Figure 19: Lifetime function estimates $\hat{\mathcal{B}}^{\mathcal{V}}(t;t_0)$ for the best, worst and intermediate predictions from Fig. 17, solid lines, together with 95% confidence bands with Gaussian assumption and IJ variance estimate, dashed curves. With mileage per day model.

References

L. Breiman. Random forests. *Machine learning*, 45(1):5–32, 2001.

L. Breiman, J. Friedman, R. Olshen, and Stone C. *Classification and regression trees*. Taylor and Francis, 1984.

A. Ciampi, J. Thiffault, J.-P. Nakache, and B. Asselain. Stratification by stepwise regression, correspondence analysis and recursive partition: A comparison of three methods of analysis for survival data with covariates. *Computation Statistics and Data Analysis*, 4:185–205, 1986.

D. R. Cox and D. Oakes. *Analysis of survival data*, volume 21. CRC Press, 1984.

D.R. Cox. Regression model and life-table. *Journal of the Royal Statistical Society*, 34(2):187–220, 1972.

M. Daigle and K. Goebel. A model-based prognostics approach applied to pneumatic valves. *International Journal of Prognostics and Health Management*, 2(2):1–16, 2011.

T.G. Dietterich. An experimental comparison of three methods for constructing ensembles of decision trees: bagging, boosting, and randomization. *Machine Learning*, 40(2):139–157, 2000.

B. Efron. Estimation and accuracy after model selection. *Journal of the American Statistical Assosiation*, 109:991–1007, 2014.

B. Efron, T. Hastie, and S. Wager. Confidence intervals for random forests: The jackknife and the infinitesimal jackknife. *Journal of Machine Learning Research*, 15:1625–1651, 2014.

E. Frisk and M. Krysander. Treatment of accumulative variables in data-driven prognostics of lead-acid batteries. In *Proceedings of IFAC Safeprocess'15*, Paris, France, 2015.

H. Hanachi, J. Liu, A. Banerjee, Y. Chen, and A. Koul. A physics-based modeling approach for performance monitoring in gas turbine engines. *IEEE Transactions on Reliability*, 64(1), 2015.

T.K. Ho. The random subspace method for constructing decision forests. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 20(8):832–844, 1998.

H. Ishwaran and U. Kogalur. Random survival forests for r. *Rnews*, 7/2:25–31, 2007.

H. Ishwaran, U. Kogalur, E. Blackstone, and M. Lauer. Random survival forests. *The Annals of Applied Statistics*, pages 841–860, 2008.

L. Liao and F. Kottig. Review of hybrid prognostics approaches for remaining useful life prediction of engineered systems, and an application to battery life prediction. *IEEE Transactions on Reliability*, 63(1), 2014.

K. Medjaher, D. A. Tobon-Mejia, and N. Zerhouni. Remaining useful life estimation of critical components with application to bearings. *IEEE Transactions on Reliability*, 61(2), 2012.

H. Moradian, D. Larocque, and F. Bellavance. L1 splitting rules in survival forests. *Lifetime Data Analysis*, 21(1), 2016.

M. Roemer, C. Byington, G. Kacprxynski, and G. Vachtsevanos. An overview of selected prognostic technologies with reference to an integrated phm architecture. In *Proceedings of the First International Forum on Integrated System Health Engineering and Management in Aerospace*, Napa, CA, USA, 2005.

B Saha and K. Goebel. Modeling li-ion battery capacity depletion in a particle filtering framework. In *Proceedings of the Annual Conference of the Prognostics and Health Managment Society*, San Diego, CA, USA, 2009.

F. Zhao, Z. Tian, E. Bechhoefer, and Y. Zeng. An integrated prognostics method under time-varying operating conditionds. *IEEE Transactions on Reliability*, 64(2), 2015.

Paper D

Battery failure prognostics using multilayer perceptron*

D

*Technical report.

Battery failure prognostics using multilayer perceptron

Sergii Voronov

Vehicular Systems, Department of Electrical Engineering, Linköping University, SE-581 83 Linköping, Sweden.

Abstract

An interest for predictive maintenance of different systems raises due to the possibility to reduce costs for maintenance and to eliminate unexpected failures. In this work lead-acid battery failure prognostics for heavy-duty trucks is considered with a predictive model being a multilayer perceptron (MLP). Data available for the study contains information about how vehicles are operated from the delivery date to the client till the date when it comes to the workshop. The model estimates the reliability function for a vehicle with data $\mathcal{V}$ that comes to a workshop. First, this work demonstrates how heterogenous data is handled, then the architecture of the MLP model is discussed. A two stage predictive model, i.e. a combination of the classification and survival stages, is suggested for estimating the reliability function of a particular vehicle. A problem of imbalance in the given data set is identified and different modifications to the neural network architecture are discussed. Finally, the result of the prediction from the MLP model is compared with prediction from the Random Survival Forest approach.

1 Introduction

The amount of information obtained from various systems increases every year. Data is logged in during the lifetime and may include operational characteristics useful for predictive maintenance which allows to anticipate a probable problem in a system and act in advance and thus avoiding failures and unplanned stops. A prognostic model development is a key part of the predictive maintenance approach as the model is used to predict the failure time.

Lead-acid starter batteries is an important part of the electrical power system of a heavy-duty truck. The primary aim of the battery is to power the starter motor to get the diesel engine running, however, it can be used for some auxiliary purposes such as cabin heating and powering kitchen equipment. Battery failure happens due to the its degradation caused by, for example sulfation, corrosion or internal short. The various deviations from normal operation of the battery may lead to the aforementioned battery problems and it is difficult to estimate, for example a technician at a workshop, the state of the battery health without proper battery diagnostics which is sometimes not possible to carry out due to time restrictions or cost of the operation. Therefore, having a predictive system, that can anticipate the failures in the battery, is a useful tool for techinicians and truck manufactures.

Different approaches can be used for lifetime prognostics of systems components. Two main directions are model-based and data-driven methods where one of the directions is chosen based on the data and type of measurements which are available to an engineer. Cornerstones of the model based methods are physical laws and equations that describe degradation of the components. Data sets, that are used in model based methods, contain time series of measurements from the system or component directly related to its health. Examples of model-based prognostics are given in (Daigle and Goebel, 2011; Hanachi et al., 2015; Saha and Goebel, 2009). It is sometimes hard to develop an accurate degradation models for a particular system, and then data-driven methods can be an alternative if reliability data is available. Authors in (Frisk et al., 2014) and (Prytz et al., 2015) propose data-driven non-parametric methods for predicting failures in components of heavy-duty trucks. Another example of a non-parametric data-driven models is an artificial neural network (Cheng and Titerington, 1994).

The aim of this work is to apply multilayer perceptron to the problem of battery failure prognostics. Different models are considered and compared giving an idea how neural networks can be useful within the prognostics domain.

2 Problem formulation

This work is done in cooperation with an industrial parter Scania CV and they would like to design a flexible maintenance planner which takes into account the state of the vehicle's components. A candidate model in such a maintenance planner is the conditional probability function $\mathcal{B}^{\mathcal{V}}(t; t_0)$ defined as

$$\mathcal{B}^{\mathcal{V}}(t;t_0) = P(T > t + t_0 \mid T \geq t_0,\ \mathcal{V}) \tag{1}$$

where T is a random variable of battery failure, t_0 is the time when a vehicle comes to the workshop and $\mathcal{V}$ is the data retrieved from the vehicles at the workshop.

The function $\mathcal{B}^{\mathcal{V}}(t;t_0)$ is the probability of the survival for a battery given data $\mathcal{V}$ from the vehicle t time units into the future from the time point t_0 when the vehicle visits a workshop. The function (1) is used as a target predictive function in the previous works by Frisk et al. (2014) and Frisk and Krysander (2015) where a Random Survival Forest method is incorporated to estimate $\mathcal{B}^{\mathcal{V}}(t;t_0)$. The conditional probability function can be expressed in terms of the reliability or survivor functions $S_T(t)$, (Cox and Oakes, 1984), as

$$\mathcal{B}^{\mathcal{V}}(t;t_0) = \frac{S_T^{\mathcal{V}}(t+t_0)}{S_T^{\mathcal{V}}(t_0)} \tag{2}$$

where reliability function $S_T^{\mathcal{V}}(t) = P(T \geq t \mid \mathcal{V})$ is defined as probability of the battery to survive more than t time units. Although the final aim is to estimate the conditional probability function $\mathcal{B}^{\mathcal{V}}(t;t_0)$, this work focuses on the estimation of the reliability function $S_T^{\mathcal{V}}(t)$ with the help of multilayer perceptron (MLP).

3 Data description

A vehicle fleet database is provided by an industrial partner Scania CV and includes variables that describe the specification of each vehicle and how it is operated during its lifetime till the date when a workshop is visited. A record in the database logged from a vehicle at the workshop is called a snapshot. Each snapshot contains static and varying variables that correspond to the variables which do not change and vary with time respectively. For example, static data includes variables such as an engine type, battery position, if there is kitchen equipment or not in the truck etc. The aforementioned variables are discrete valued, for instance, the battery position variable can take three possible values such as left, right mounting positions and rear frame position. Non-static variables are numeric valued and mostly represent data in the form of histograms. For example, there exists a temperature histogram with 10 bins in the data set which shows what percentage of time a vehicle has been operated in a particular temperature range and that corresponds to one bin of the histogram.

There is a variable in the data set that indicates if a vehicle has had problems with a battery. If a vehicle leaves the study without experiencing a battery problem, it is called censored. More than 90 per cent of the vehicles are censored in the data set, meaning that a battery failed only for a small fraction of them.

Missing data is an essential attribute of many real life data sources and the missing data rate for the data set under study is between 30 and 40 percent

which means that some variables has no record in a specific snapshot. A main reason for missing data in the data set is due to the fact that variables introduced for one type of vehicles are no longer present for another type of a vehicle.

Main characteristics of the data set are outlined bellow:

- 33603 vehicles from 5 EU markets
- 284 variables stored for each vehicle snapshot
- A single snapshot per vehicle
- Heterogeneous data, i.e., it is a mixture of categorical and numerical data
- Availability of histogram variables
- Censoring rate more than 90 percent
- Significant missing rate
- No measurements directely related to battery health

It should be noticed that there are no time series for a vehicle and it means that it is impossible to track the health of the battery during the vehicle's lifetime. Therefore, what is needed from the model is to find groups of vehicles based on similar usage profiles and, then use the group of vehicles to estimate the reliability function $S_T^{\mathcal{V}}(t)$.

4 MLP models

Neural network with several hidden layers and hidden nodes is called a multilayer perceptron (MLP). All models analyzed in the paper, as well as their configurations, are presented in this section. There are three different models presented in the paper: 1) a single survival MLP, 2) a two-staged classification-survival MLP and 3) categorical/numerical data MLP. The network models are implemented in Python with the help of the *Keras* library (Chollet, 2015). Before going to the description of the neural network architectures, adaptation of the data set to the MLP model is presented.

4.1 Arranging data for MLP model

This subsection demonstrates how variables from the data set are preprocessed before submission into the neural network. Discrete or categorical data is transformed into one-hot vectors. For example, if possible values of variable v_1 are $\{A, B, C\}$, then new values to be submitted into the MLP model look like

$$\begin{array}{lcl} A & \rightarrow & (0 \quad 0 \quad 1) \\ B & \rightarrow & (0 \quad 1 \quad 0) \\ C & \rightarrow & (1 \quad 0 \quad 0). \end{array} \tag{3}$$

Numerical valued variables are scaled in such way that they take values within interval $[0, 1]$. Every histogram bin is treated as one variable. Therefore, if a histogram consists of 10 bins, then the histogram contributes with 10 variables to the data set.

The reliability function $S_T^{\mathcal{V}}(t)$ is estimated by the output layer in an MLP model. A procedure similar to (Chi et al., 2007) is used to assign a reliability function to a particular vehicle. The prediction for lifetime of the vehicle's battery is done for 5 time units with a time step of 0.25 time unit. This means that the output layer of an MLP model should have 20 nodes, each one showing probability of battery to survive another 0.25 time units. For the failed batteries the following approach is used. If battery lived 3 time units, all 12 first nodes in the output layer are 1s and 0s afterwards. The result will look like

$$3 \quad \text{time units} \quad \rightarrow \quad (\underbrace{1 \quad 1 \quad \ldots \quad 1}_{\text{12 nodes}} \underbrace{0 \quad 0 \quad \ldots \quad 0}_{\text{rest of the nodes}}). \tag{4}$$

In the case of a censored vehicle/battery the procedure is a bit different. Nodes in the output layer that correspond to the time when a vehicle is under study are assigned with 1 values. Remaining part of the reliability function is estimated with a Kaplan-Meier estimate, see (Cox and Oakes, 1984), where two approaches are applied. First, only vehicles after censoring time are used in the Kaplan-Meier estimator. The second approach is to use all vehicles to compute an estimate and then only the part of the curve which corresponds to the interval of interest is assigned to the remaining nodes of the reliability function. Reliability function for a vehicle censored after 3.75 time units could be as

$$\begin{array}{l} \qquad\qquad\qquad\qquad 3.75 \quad \text{time units} \qquad\qquad\qquad\qquad \rightarrow \\ (\underbrace{1 \quad 1 \quad \ldots \quad 1}_{\text{15 nodes}} \quad 0.86 \quad 0.76 \quad 0.623 \quad 0.256 \quad 0). \end{array} \tag{5}$$

Missing values in the variables are treated in the following way. The mean of each column which corresponds to the values of a variable is computed and used instead of missing values.

4.2 MLP ARCHITECTURES

The first model, called a survival MLP, that is tested has the following layout. The input layer is the union of the one-hot vectors, which represent the discrete or categorical data, with scaled numerical valued data, as described in Subsection 4.1. The output layer consists of 20 nodes where each node has a sigmoid activation function and estimates the reliability function for a given vehicle. The sigmoid activation function is selected due to the fact that all nodes should have values in the interval $[0, 1]$ which should correspond to the values of the reliability function. There are two hidden layers in the model where each node has a rectified linear activation function. The first hidden layer has 120 nodes and the second consists of 60 nodes. The dropout value is chosen to be 0.5 and

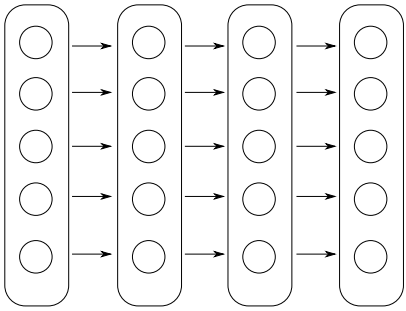


Figure 1: Model of MLP with two hidden layers with rectified linear activation function. Output layer has 20 nodes each with sigmoid activation function.

the objective function is chosen to be a mean square error (MSE). The stochastic gradient descent (SGD) method is selected as an optimizer for training the MLP. An illustration of the survival MLP can be found in Fig. 1.

The idea behind the choice of the MSE objective function is to allow the network to learn/estimate the mean survival function of each class, failed or censored, vehicles. However, the results, Section 5, show that the initial idea gave bad performance which led to the introduction of a two staged MLP model.

Another model used is a classification-survival MLP, see Fig. 2. It consists of the two stages or two MLPs, namely, a classification MLP and a survival MLP. The first step is to build a classification MLP where each vehicle should be classified either to a failed or a censored class of vehicles. This step assigns the vehicle either to the risk class, failed vehicles, or healthy class, censored vehicles. The structure of the network is similar to the survival MLP except of the last layer which now has two nodes with a softmax activation function and categorical crossentropy as an objective. The second step is to train a survival MLP model as described above on two classes of the vehicles, failed and censored, separately. To make a prediction for a unseen vehicle that comes to a workshop the following procedure is applied: a) classify the vehicle as a failed or censored, b) estimate the reliability function of the vehicle based on a survival MLP.

A third architecture of the neural network is illustrated in Fig. 3. The two types of variables, discrete and numerical, are separated from each other into two branches. The interaction between the two branches is organized through the *Merge* layer in *Keras*, see (Chollet, 2015). An idea of this approach is to see how different merging of heterogenous data influences the prediction. The results of applying the models described in the current section can be found in Section 5.

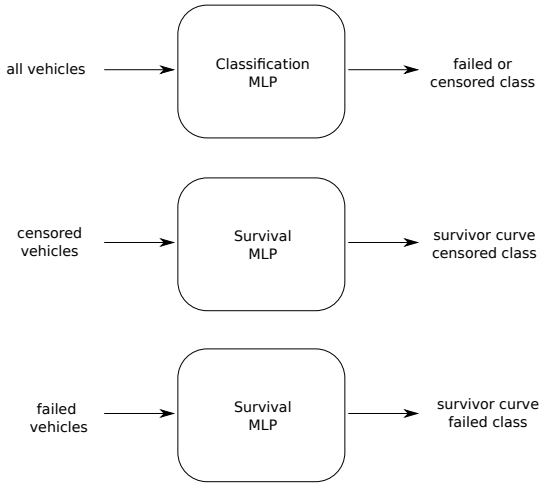


Figure 2: Classification-survival system. Classification step is trained on all vehicles. Survival step is performed separately for each class of vehicles.

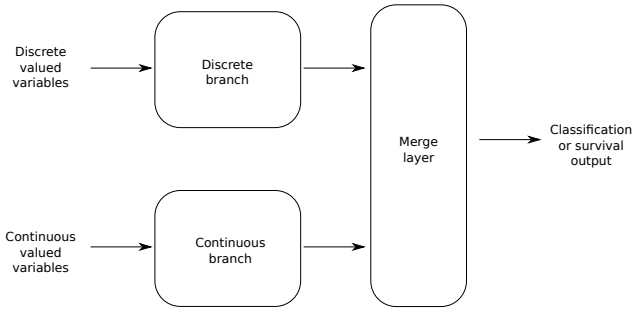


Figure 3: MLP model with separate branches for discrete and continuous valued variables.

5 Analysis and results

Now the three models introduced in Section 4 are analyzed together with results of their predictions. The prognostics domain does not have one established metric that would allow us to evaluate the performance of the model, therefore, it is not straightforward which approach to use here. The following two approaches were selected as candidates for validation of the results. First, predicted mean reliability functions for the failed and censored vehicles are plotted together with their reference values, i.e., reference functions from the training set of vehicles. Predictions from the model for every vehicle in the validation set are received, the reliability functions, then predictions for the failed and censored vehicles are grouped together and mean values of the predicted reliability functions for each class are computed. It is not known in advance how average reliability functions should look like due to the absence of knowledge of the true degradation profiles in the data set, but it is expected that the predictions should differ on average for the two classes of vehicles. Second plot chosen for the validation is the histogram of reliability values for both classes of vehicles at a particular time point, in this case time point is $t = 3$. The given time point corresponds somewhere near the half of the study time. This type of plot shows how the values of the reliability function are scattered around the mean value at the given time point.

5.1 Survival MLP with imbalanced data

First, the survival MLP is applied to the data set. The setup of the survival MLP is described in Section 4. The objective function is the MSE and a stochastic gradient descent is used as optimizer. The data set is split into the training set, corresponding to 2/3 of the whole data, and the validation set, corresponding to 1/3 of the whole data set. The idea of the survival MLP is to allow the network to learn the differences between two classes of vehicles through the training procedure. It was expected to receive two different average reliability funtions for failed and censored vehicles. However, the results do not support the expectation.

The results, a plot with the average realibity functions and a histogram of reliabilities for a particular time point, for the survival MLP model are presented in Fig. 4 and Fig. 5. In Fig. 4 the red dotted curve corresponds to the average reliability function for the censored vehicles and the blue dotted curve to the average reliability function for the failed vehicles. The latest is not visible in Fig. 4, because the average predictions for the failed and censored vehicles are almost identical. This fact can be confirmed by looking at the histogram of reliability values in Fig. 5 which is quite similar for both classes of vehicles. These results indicate that the model does not make any distinction when making prediction for the censored and failed vehicles. Another thing to notice is that the prediction from the model follow reference reliability function for the class of censored vehicles. Thinking about the structure of the data, one can recall that our data is imbalanced, only less than 10% of the vehicles have problems with

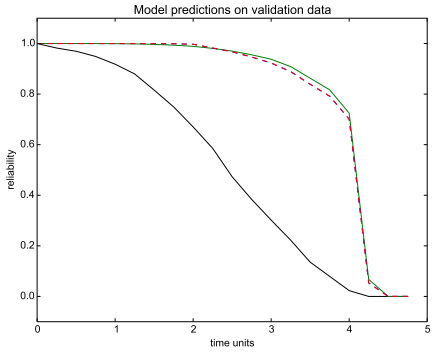


Figure 4: Average reliability functions for the initial model with imbalanced data on validation set. Blue and red dotted curves are the predictions for failed and censored vehicles respectively (blue curve is not seen because it is almost identical to red one). Green and black solid curves are actual survival curves based on validation data for censored and failed vehicles respectively.

the battery. Therefore, the next step was to see how balancing of data would influence the results.

5.2 Survival MLP with balanced data

There are different approaches available when it comes to the question how to balance the classes in the data set. Some of them can be found in (He and Ma, 2013) where the undersampling technique is chosen here. The amount of vehicles in the failed class is left unchanged, however, the amount of censored vehicles is reduced to the same amount of failed ones by sampling uniformly from the censored class. This procedure substantially reduce number of training samples (about 1500 vehicles compared to 21000 in the case of the imbalanced data set), but increasing number of epochs and decreasing batch size when compiling model in *Keras* helps to mitigate the problem.

Results for the given set up can be seen in Fig. 6 and Fig. 7. It is still the case that average predictions for failed and censored classes are very similar, Fig. 6. However, predictions are not biased to any of the reference reliability functions. Here, the results achieved in Subsections 5.1 and 5.2 are surprising. It was expected that the average predictions for the two classes of vehicles will be different and the idea is that the hidden units should learn the difference in degradation depending on class of the vehicle. The results show that it is not the case which requires additional thinking on the reasons for this behavior.

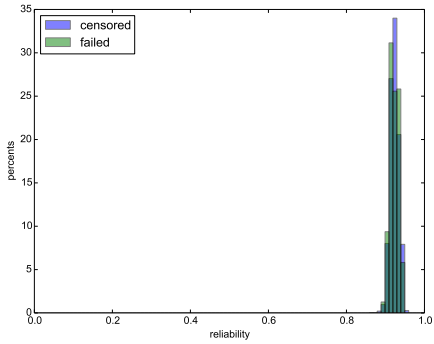


Figure 5: Histogram of reliabilities at time unit 3 for the initial model with imbalanced data. Green bars correspond to failed vehicles and purple to censored ones.

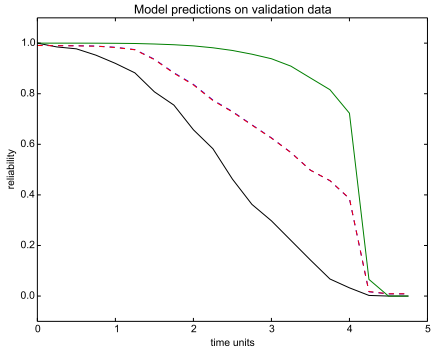


Figure 6: Average reliability functions for the initial model with balanced data on validation set. Blue and red dotted curves are the predictions for failed and censored vehicles respectively (blue curve is not seen because it is almost identical to red one). Green and black solid curves are actual survival curves based on validation data for censored and failed vehicles respectively.

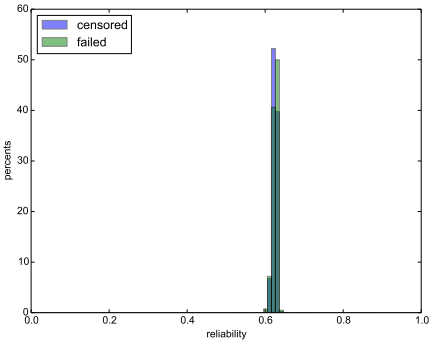


Figure 7: Histogram of reliabilities at time unit 3 for the initial model with balanced data. Green bars correspond to failed vehicles and purple to censored ones.

5.3 Classification-survival MLP

One conclusion from the experiments in Subsections 5.1 and 5.2 is that the models fail to capture different degradation profiles for failed and censored vehicles. However, it seems that the model with the MSE objective function can follow the referene average survival curve of a particular class of the vehicles if it is trained only on the representatives from that class. Therefore, a two staged classification-survival MLP model is developed as described in Section 4. First, a model learns how to classify the vehicles as failed or censored, i.e., a classification problem is considered. Then, two survival MLP models are trained separately for failed and censored vehicles with network output as the reliability functions.

The results of this approach on the validation data set are presented in Fig. 8 and Fig. 9. As evident from Fig. 8 the situation is different compared to the previous cases. The average reliability functions differ from each other and the histogram of reliabilities is separated for the two classes of vehicles. Here, the conclusion that the reliability functions are different is based on a visual inspection. Predicted reliability functions, dotted lines in Fig. 8, do not follow precisely the reference reliability functions, shown as solid lines, due to the missclassification which happens in the classification MLP. This can also be confirmed by looking at Fig. 9 where some censored vehicles falls into the category of failed ones, receiving a lower prediction of reliability than they should have and the opposite is true for the failed vehicles.

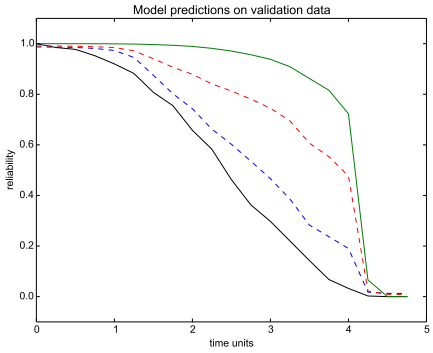


Figure 8: Average reliability functions for two staged neural network model on validation data. Blue and red dotted curves are the predictions for failed and censored vehicles respectively. Green and black solid curves are actual survival curves based on validation data for censored and failed vehicles respectively.

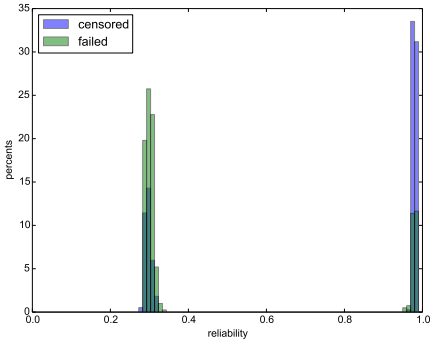


Figure 9: Histogram of reliabilities at time unit 3 for two staged neural network model. Green bars correspond to failed vehicles and purple to censored ones.

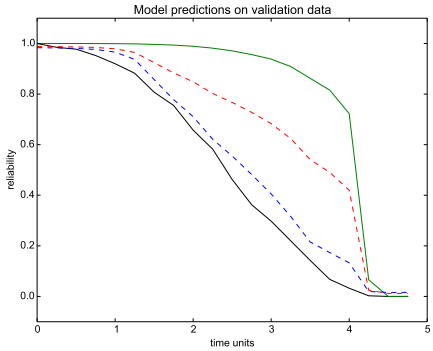


Figure 10: Average reliability functions for the two staged neural network model applied to validation data. Columns with a missing data more than 40% are neglected. Blue and red dotted curves are the predictions for failed and censored vehicles respectively. Green and black solid curves are actual survival curves based on validation data for censored and failed vehicles respectively.

5.4 Modifying survival prediction within classification-survival MLP

This subsection contains the results of modifications applied to the survival prediction MLP within classification-survival model from Subsection 5.3. The results are collected into one subsection because they are quite similar.

The first modification is to remove variables with a rate of missing values larger than 40% from the study and training process. The idea is to see how the results change when variables with a high degree of distorted/lost information are removed.

The results of this approach can be seen in Fig. 10 and Fig. 11. One can see that the results are similar to ones from Fig. 8 and Fig. 9. One conclusion is that adding columns with a lot of missing data do not make the average prediction of the model worse. If only these two plots are used for judging the performance of the method, it is reasonable to choose the model with reduced number of variables, because it will take less time to build the model and make a prediction for an unseen vehicle.

Fig. 12 and Fig. 13 show the results for the case when the part of the reliability function after the censoring time for the vehicles without battery problems is approximated with the corresponding part of the Kaplan-Meier estimate computed using all vehicles. For example, to estimate the reliability function for a battery that is censored after 3 time units, first, the Kaplan-Meier estimate is computed which is based on all vehicles, then its part corresponding to time interval from 3 to 5 time units is assigned to the respective part of reliability function of the vehicle. In this case it can be seen that the predicted

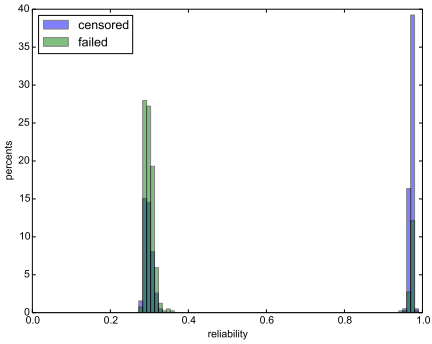


Figure 11: Histogram of reliabilities at time unit 3 for the two staged neural network model. Columns with a missing data more than 40% are neglected. Green bars correspond to failed vehicles and purple to censored ones.

average reliability function for the censored vehicles in some parts of the curve follows the reference function better than the same curve in Subsection 5.3, see Fig. 12. However, the prediction for the failed vehicles follows the reference function worse than for the previous model. Therefore, it is not possible to say that the results with this setup differ significantly from the model in Subsection 5.3.

The last modification to the classification-survival MLP model architecture is that the multilayer perceptron with two branches, a separate branch for numeric and a separate branch for categorical data, is considered. The two branches are concatenated with the help of the *Merge* layer in *Keras*, Fig. 3. The idea is to see how the results will change if one-hot vectors that represent categorical data are separated from the numerical data by means of the two branches. It seems that the average predictions are not affected, see Fig. 14. However, the histogram of the reliability functions at time point $t = 3$ is different from the previous models, see Fig 15. It can be seen that the bars of the failed vehicles are higher than the bars of the censored which is not the case for the model from Subsection 5.3, shown in Fig 9. Therefore, one conclusion is that different merging of data can lead to different predictions from the model. It is left for future research to answer the question if this change in the prediction improves or degrades performance.

5.5 Comparing MLP and RSF models

This subsection compares the MLP model with the Random Survival Forest (RSF) model, which is used in (Frisk et al., 2014) and (Frisk and Krysander, 2015) for estimating conditional probability function. Similar plots as for the MLP models are shown for the RSF model. A fair comparison is difficult to

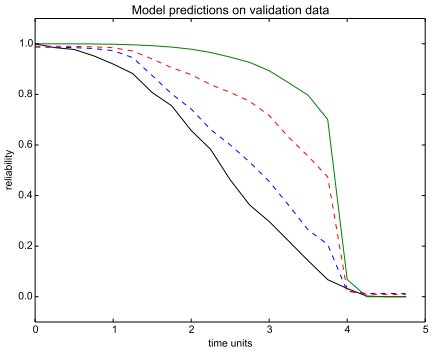


Figure 12: Average reliability functions for two staged neural network model on validation data where missing parts of survival curves for censored vehicles are approximated with Kaplan-Meier estimator common for the whole data. Blue and red dotted curves are the predictions for failed and censored vehicles respectively. Green and black solid curves are actual survival curves based on validation data for censored and failed vehicles respectively.

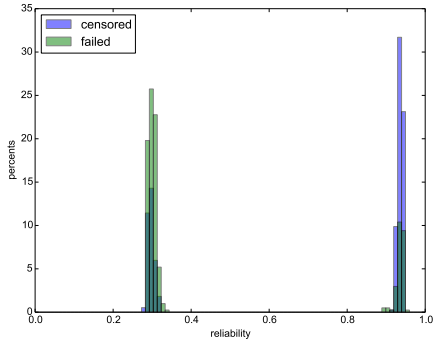


Figure 13: Histogram of reliabilities at time unit 3 for two staged neural network model where missing parts of survival curves for censored vehicles are approximated with Kaplan-Meier estimator common for the whole data. Green bars correspond to failed vehicles and purple to censored ones.

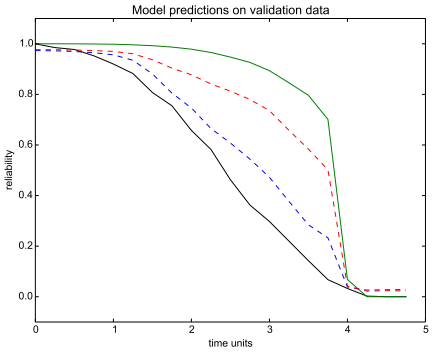


Figure 14: Average reliability functions for two staged neural network model on validation data for the case when discrete and continuous data are split into separate branches in neural network. Blue and red dotted curves are the predictions for failed and censored vehicles respectively. Green and black solid curves are actual survival curves based on validation data for censored and failed vehicles respectively.

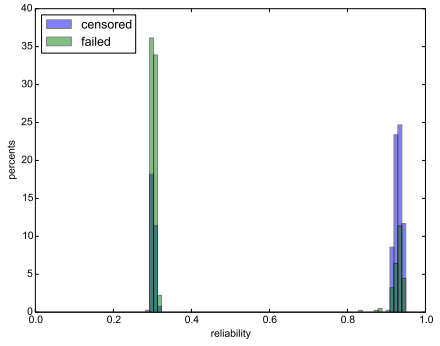


Figure 15: Histogram of reliabilities at time unit 3 for two staged neural network model for the case when discrete and continuous data are split into separate branches in neural network. Green bars correspond to failed vehicles and purple to censored ones.

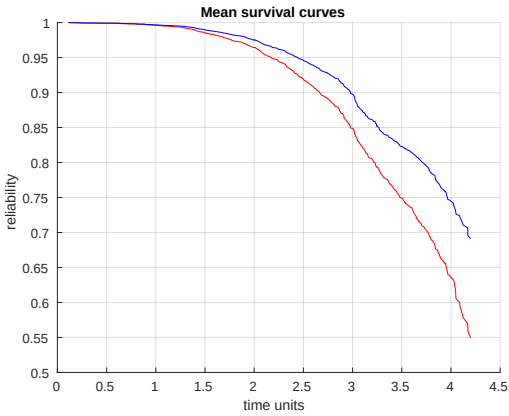


Figure 16: Average reliability functions for censored vehicles, blue solid line, and failed vehicles, red solid line, computed on validation set. Random Survival Forest (RSF) model is used with 1000 trees and minimal node size 200.

perform at this moment, because methods for the reliability function prediction differ. Nonetheless, some trends can be noticed. The same dataset and the same split between validation/training sets are used for building the RSF model as for MLP one. The forest has 1000 trees with minimal node size of 200, which is used in (Frisk et al., 2014) and (Frisk and Krysander, 2015). The average reliability functions for both classes of vehicles are shown in Fig. 16. Up to time unit 1.5 both curves are quite similar after that they start to diverge. In the case of the MLP model, curves are similar up to 1 time unit and then start to diverge. As was discussed above we can not compare results directly, but one thing that can be said is that both methods give on average a prediction that is different for the failed and censored vehicles which is something that one expects to happen. The histogram of the reliabilities for vehicles at the same time point as for MLP, namely, 3 time units is presented in Fig. 17. One can see that the figure is significantly different from the one in Fig. 9. This happens due to the fact that in the MLP model classification step is performed first. Therefore, we expect a large separation between the histograms for the failed and censored vehicles. However, when forest is being built, data is grouped into classes where the vehicles are used in a similar manner, therfore, one can say that a classification problem is hidden within construction of the forest, but the number of those classes are significantly bigger than for the MLP case and the difference between them in terms of vehicle usage profile is not as distinctive as for the case of only failed and censored classes. This explains a larger overlapping area between the two histograms in Fig. 17.

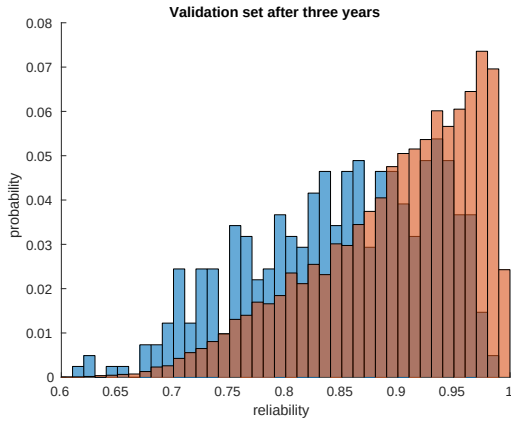


Figure 17: Histogram of reliabilities at time unit 3 for RSF model. Blue bars correspond to failed vehicles and orange to censored ones.

6 Conclusions

The problem of battery lifetime prediction with a multilayer perceptron as the predictive model is considered in the given work. The model estimates the reliability function for a vehicle with data $\mathcal{V}$ that comes to a workshop. A first contribution is an adaptation of the multilayer perceptron model to the data set under study, i.e., to the mixture of categorical and numerical data. The problem of imbalance in the data set is identified and addressed with undersampling technique which results in an improved prediction. Training a multilayer perceptron with a mean square error objective function gives unacceptable results, since the average prediction for failed and censored vehicles is almost identical, and it is expected that they are different. Therefore, a two stage approach with the first step as a classification problem and, then, a survival prediction at the second stage leads to improved results. Comparing the result of the prediction with the random survival forest and the multilayer perceptron models indicates that the model which is based on a neural network approach is an alternative direction for the battery lifetime prediction.

References

J. Cheng and D.M. Titerington. Neural networks: A review from a statistical perspective. *Statistical Science*, 9(1):2–54, 1994.

C. Chi, N. Street, and W. Wolberg. Application of artificial neural network-based survival analysis on two breast cancer datasets. In *AMIA Symposium Proceedings*, 2007.

François Chollet. Keras. `https://github.com/fchollet/keras`, 2015.

D. R. Cox and D. Oakes. *Analysis of survival data*, volume 21. CRC Press, 1984.

M. Daigle and K. Goebel. A model-based prognostics approach applied to pneumatic valves. *International Journal of Prognostics and Health Management*, 2(2):1–16, 2011.

E. Frisk and M. Krysander. Treatment of accumulative variables in data-driven prognostics of lead-acid batteries. In *Proceedings of IFAC Safeprocess'15*, Paris, France, 2015.

E. Frisk, M. Krysander, and E. Larsson. Data-driven lead-acide battery prognostics using random survival forests. In *Proceedings of the Annual Conference of The Prognostics and Health Management Society*, Fort Worth, Texas, USA, 2014.

H. Hanachi, J. Liu, A. Banerjee, Y. Chen, and A. Koul. A physics-based modeling approach for performance monitoring in gas turbine engines. *IEEE Transactions on Reliability*, 64(1), 2015.

H. He and Y. Ma. *Imbalanced learning: Foundations, Algorithms, and Applications*. IEEE Press, 2013.

R. Prytz, S. Nowaczyk, T. Rögnvaldsson, and S. Byttner. Predicting the need for vehicle compressor repairs using maintenance records and logged vehicle data. *Engineering applications of artificial intelligence*, 41:139–150, 2015. ISSN 0952-1976.

B Saha and K. Goebel. Modeling li-ion battery capacity depletion in a particle filtering framework. In *Proceedings of the Annual Conference of the Prognostics and Health Managment Society*, San Diego, CA, USA, 2009.