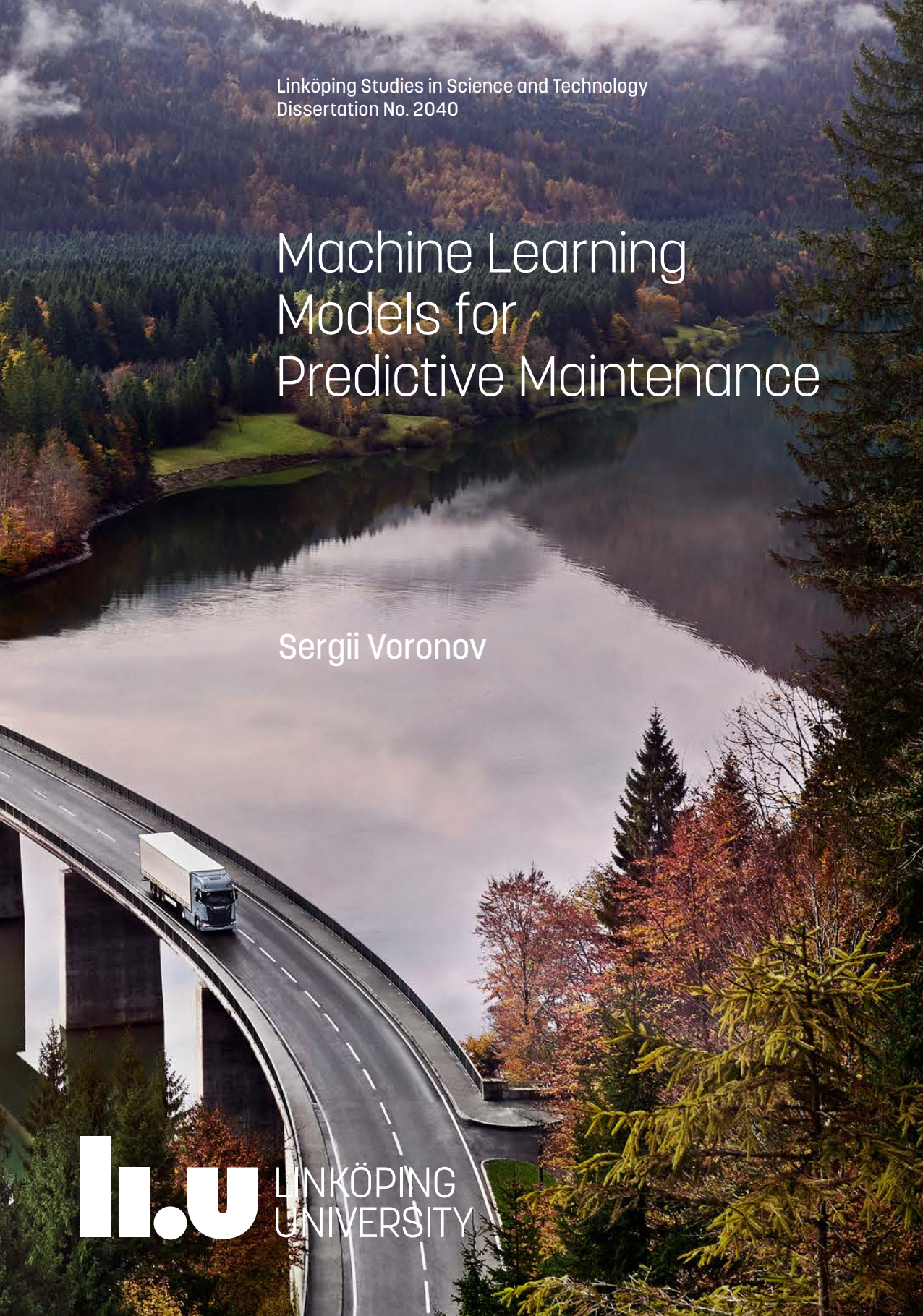
Linköping Studies in Science and Technology
Dissertation No. 2040

Machine Learning Models for Predictive Maintenance

Sergii Voronov



LINKÖPING UNIVERSITY

Linköping Studies in Science and Technology
Dissertations, No. 2040

Machine Learning Models for Predictive Maintenance

Sergii Voronov

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

Linköping 2020

Cover picture: The cover page picture is an edited version of the picture taken by Gustav Lindh near Sylvenstein, Germany in 2018. Scania CV AB, copyright holder, has given a consent for the author to use the picture.

ISBN 978-91-7929-923-1
ISSN 0345-7524
URL http://urn.kb.se/resolve?urn=urn:nbn:se:liu:diva-162649

Typeset using XeTeX

Printed by LiU-Tryck, Linköping 2020

POPULÄRVETENSKAPLIG SAMMANFATTNING

Mängden gods som produceras och transporteras världen runt ökar och tunga fordon är en viktig del i logistikkedjan. För att garantera pålitliga leveranser krävs hög tillgänglighet hos fordonen genom att bland annat undvika oplanerade stopp längs vägen. Tid då fordonet ej är tillgängligt kan reduceras genom att byta ut komponenter baserat på statistik fråntidigare fel. En sådan ansats kan dock vara dyr på grund av för täta besök på verkstäder samt att många komponenter fungerar avsevärt längre beroende på hur hårt komponenten använts. En prognostikmetod för individualiserade underhållsplaner har därför en stor potential i fordonsfältet. Prognostikmetoden uppskattar komponenters degradation och tillgänglig livstid baserat på registrerade data och hur fordonet har använts.

Blysyrabatterier är en del av det elektriska kraftsystemet i en lastbil, primärt ansvariga för att kraftsätta startmotor, men också för att ge kraft åt hjälpsystem som kabinvärme och köksutrustning, vilket betyder att batteriet är en viktig komponent för fordonets tillgänglighet. Att utveckla fysikaliska modeller för batteridegradation är svårt och kräver tillgång till mätdata direkt kopplat till batteriets hälsa, något som inte är tillgängligt i det här arbetet. En alternativ ansats, som utforskas här, är datadrivna metoder baserade på stora mängder inspelade data som beskriver hur fordonet använts. I studien är insamlad data ej direkt relaterad till batterihälsa vilket gör prognostikproblemet utmanande.

Ett huvudbidrag är utveckling av maskininlärningsmodeller för prediktivt underhåll baserad på Random Survival Forests (RSF) och Recurrent Neural Networks (RNN). En viktig egenskap hos insamlade data är att för specifika fordon så kan det finnas flera, eller endast enstaka, datautläsningar vilket också gör prediktiv modellering svårt. Metoder för att hantera detta för modeller baserade på RSF och neuronnät behandlas. Datakvalitet är viktigt vid utveckling av datadrivna modeller. Insamlade data är obalanserade eftersom det är få batterier som felar i relation till antalet fordon. Vidare, insamlade data inkluderar många oinformativa variabler och bland de informativa så finns komplexa beroenden och korrelationer. Metoder för att välja väl valda variabler att bygga modeller på för den här situationen är utmanande och ett huvudbidrag i arbetet. En central fråga är hur säker en punktskattning är och hur den osäkerheten kan vägas in när prediktiva underhållsplaner bestäms, speciellt när modellen baseras på så osäkra data och så ostrukturerade modeller som här. Ett viktigt bidrag är metodik för att estimera prediktionsvarians för RSF-modeller. Slutligen, ett huvudresultat för användarfallet är att LSTM-nät, ett typ av RNN, är den modellstruktur som ger bäst prestanda för prognostik av blysyrabatterier med det data som använts i avhandlingen.

ABSTRACT

The amount of goods produced and transported around the world each year increases and heavy-duty trucks are an important link in the logistic chain. To guarantee reliable delivery a high degree of availability is required, i.e., avoid standing by the road unable to continue the transport mission. Vehicle downtime can be reduced by replacing components based on statistics of previous failures. However, such an approach is both expensive due to the required frequent visits to a workshop and inefficient as many components from the vehicles in the fleet are still operational. A prognostic method, allowing for vehicle individualized maintenance plans, therefore poses a significant potential in the automotive field. The prognostic method estimates component degradation and remaining useful life based on recorded data and how the vehicle has been operated.

Lead-acid batteries is a part of the electrical power system in a heavy-duty truck, primarily responsible for powering the starter motor, but also powering auxiliary units, e.g., cabin heating and kitchen equipment, which makes the battery a vital component for vehicle availability. Developing physical models of battery degradation is a difficult process which requires access to battery health sensing that is not available in the given study as well a detailed knowledge of battery chemistry. An alternative approach, considered in this work, is data-driven methods based on large amounts of logged data describing vehicle operation conditions. In the use-case studied, recorded data is not closely related to battery health which makes battery prognostic challenging.

A main contribution of the thesis is development of machine learning models for predictive maintenance, estimating conditional reliability functions, using Random Survival Forests (RSF) and recurrent neural networks (RNN). An important property of the data is that for a specific vehicle there may be multiple data readouts, but also one single data readout which makes predictive modeling challenging and dealing with this situation is discussed for both RSF and neural networks models. Data quality is important when building data-driven models, and here the data is imbalanced since there are few battery failures relative to the number of vehicles. Further, the data includes many uninformative variables and among those that are informative, there are complex dependencies and correlation. A method for selecting which data features to use in the model in this situation is also a key contribution. When a point estimation of the conditional reliability functions is available, it is of interest to know how uncertain the estimate is as it allows to take quality of the prediction into account when deciding on maintenance actions. A theory for estimating the variance of the RSF predictor is another contribution in the thesis. To conclude, the results show that Long Short-Term Memory networks, which is a type of RNN, is the most suitable for the vehicle operational data and give the best performance among methods evaluated in the thesis.

ACKNOWLEDGMENTS

First of all, I would like to thank Scania CV and Vehicular Systems group at Linköping University for an opportunity to participate in the lead-acid battery prognostic project. All work and results would be impossible without a chance that had been given to me.

I would like to express my gratitude to the main supervisor Dr. Erik Frisk and co-supervisor Dr. Mattias Krysander for all the help and advises, 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 data science. I would also like to thank Dr. Jonas Biteus, Anders Vesterberg and Olof Steinert who work at Scania CV for the great discussions and insights into the problem and indispensable help with data extraction. I am grateful to all colleagues from Vehicular Systems group for a nice and pleasant working environment.

The very special gratitude goes to my family and, in particular, to my parents. Thank you for always supporting me. Only you can cheer me up in the situations when hope is lost. Thank you for teaching me to be a man who I am now. The acknowledgment section would be impossible without mentioning my friends. Thank you all a lot for the great memories and fun. Philosophical discussions or doing sports, traveling to famous attractions or trying to survive in the middle of nowhere, we can always find what to do 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, 2020

Sergii Voronov

Contents

1 Introduction

Predictive maintenance aims at estimating or predicting failure time of a system or its components based on experience, physical laws, or machine learning techniques and replacing the faulty components before failure, and as the result minimizing downtime of the systems. People were interested in predicting lifetime of various systems from the ancient times. For example, carts are used for transporting goods starting from the time when a wheel was invented. It is a very simple mechanical system, nevertheless, there are several components that can break such as an axle and a wheel which will lead to a failure or a stop. At that time, people used visual inspection and personal experience to decide if a particular part has to be replaced before a journey starts. Nowadays, technology has evolved, systems have become more complex, a lot of new systems have appeared and it is not possible anymore to use only visual assessment to reliably decide when one or another component will fail.

Intelligent, condition-based, maintenance, i.e., predictive maintenance, is an efficient and cost-effective tool to improve avail-

ability and uptime in many industrial applications. Prognostics, given the current state of health of the system/component, predicts its state in the future and estimates the possible time of failure. The predictive model is a cornerstone in predictive maintenance.

Application of predictive maintenance and prognostic models to modern industrial systems range from electronics, aeronautical, automotive to industrial machinery and more applications are implemented every day. For example, authors in (Batzel and Swanson, 2009) presented a framework for predicting electrical failures in an aircraft power generator that allows to avoid unexpected failures and reduce expenses for system maintenance. Articles (Miao, Xie, Cui, Liang, and Pecht, 2013) and (Nuhic, Terzimehic, Soczka-Guth, Buchholz, and Dietmayer, 2013) introduce machine learning models to predict failures in Lithium-ion batteries where the investigations show that time of failures can be estimated reliably. Bearings are important components in mechanical systems and authors in (Ali, Chebel-Morello, Saidi, Malinowski, and Fnaiech, 2015) suggest a data-driven prognostic approach to accurately predict breakdown of the bearings to reduce maintenance cost of the mechanical systems. Predictive models for degradation in industrial applications can be integrated into new technologies, such as Internet of Things (IoT), that can improve the manufacturing process by increasing uptime of the plant infrastructure. For instance, opportunities and suggestions on how predictive models can be implemented together with IoT are described in (Kwon, Hodkiewicz, Fan, Shibutani, and Pecht, 2016).

Areas of significant current interest are autonomous systems and autonomous vehicles where the main aim is systems that can perform required tasks without any human supervision. For example, self-driving vehicles should deliver goods to customers or run autonomously at construction sites. Autonomous operation of dynamic systems with high uptime and safety requirements and no direct user feedback is a challenging task where predictive maintenance is important.

1.1 Predictive maintenance in transportation

Fast delivery of various goods is important to the world economy as the success of many businesses depend on it. According to Eurostat, a statistical office of the European Union (EU) situated in Luxembourg, the aggregated inland transport performance over EU is 2,438 billion tonne-kilometers in 2017 (Eurostat, 2017). A tonne-kilometer is a unit of measure which represents the transport of one tonne of goods over one kilometer. As can be seen, it is an enormous amount of goods being transported over one year only in the EU. Moreover, the share of road transport, long-haulage and distribution trucks, in the total amount of freight being transported in the EU is 51.5% if air and maritime transports are taken into account, and 76.7% when only inland freight transport is considered (Eurostat, 2017). Apart of delivering goods, heavy-duty vehicles are used, for example in mines and construction cites where they are a vital link in the production process.

Therefore, it is important that vehicles have a high degree of availability and in particular avoid vehicle failures causing stops along the road and aborted transport missions. 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 parts involved. Thus, 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 instead sell transports in tonne-kilometers.

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 lead-acid batteries is to power the electrical starter motor that helps the diesel engine during engine starts. At the same time, the battery can also be used to power auxiliary units such heating system during cold periods of year and kitchen equipment, coffee maker and microwave, in long-haulage vehicles. In general, predicting degradation of a

component in a truck is a challenging task due to the fact that vehicles are operated in diverse operating conditions. This is also true for a lead-acid battery where degradation depends on temperature of the ambient environment, extreme values of temperature has negative impact on battery lifetime, number of vehicle starts or stops, distribution trucks performs many stops in a city, or on degradation of other components, for example if the generator overcharges the battery its lifetime decreases, etc.

Predictive 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.

1.2 Lead-acid batteries

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

Considering modern batteries, there are many 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. Fig. 1.1 shows one example of a lead-acid starter battery used in Scania heavy-duty trucks.

A schematic illustration of a lead-acid battery is given in Fig. 1.3. It consists of positive terminal, lead dioxide PbO_2, negative terminal, lead Pb, and sulfuric acid H_2SO_4 as electrolyte. Notice that the chemical representation given above is for a charged battery. During the battery discharge chemical processes change

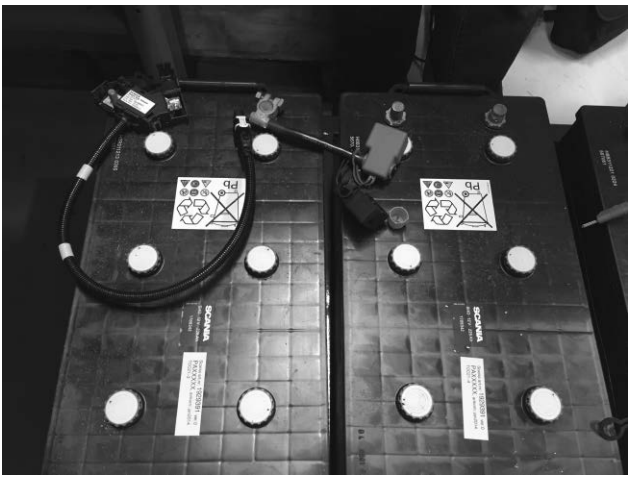


Fig. 1.1: Two 12 V super heavy-duty lead-acid starter batteries used at Scania CV.

(a) healthy positive plate (b) degraded positive plate

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

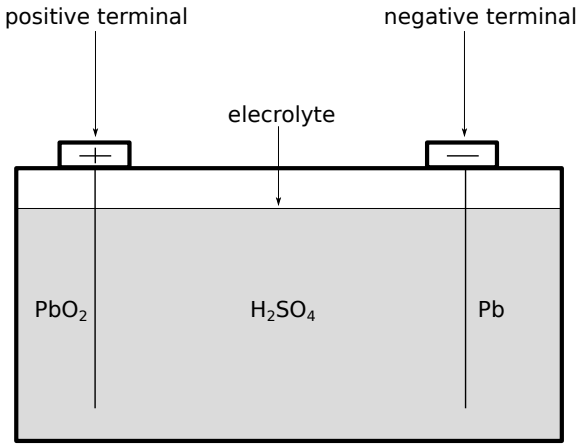


Fig. 1.3: Schematic representation of fully charged lead-acid battery.

components of the battery. For instance, in a fully discharged state, both positive and negative terminals of the battery become sulfate of lead $PbSO_4$ where the electrolyte transforms into primarily a water solution. In Fig. 1.3, the positive and negative terminals are represented with a single lead plate as in Fig. 1.2, however in practice the terminals are a pack of plates. The number of lead plates in the package depends on battery usage, for instance, terminals of batteries for deep discharge applications have a low number of thick plates, but for high peak current applications such as the batteries in heavy-duty trucks, the terminals consists of large number of thin plates.

This thesis concerns estimating the time of battery failure and main reasons for the lead-acid battery degradation, which can eventually cause a failure, are sulfation, corrosion, and internal short. Sulfation happens when the battery is not charged to its full charge value. For example, it is a common problem for trucks operated within cities as the battery is only charged by the generator while the vehicle is moving. In cities, there are many starts and stops, as a result the traveling distances and charging times

are short. Corrosion appears due to high charge voltage which can happen if a vehicle is operated in a broad temperature range and if the charging procedure is not designed properly. Internal short is a type of a degradation that progresses during the lifetime when small parts of the lead from the plates accumulate at the bottom of the battery container which can lead to a conducting layer formation that connects two plates. Fig. 1.2 illustrates a healthy and a severely degraded positive plate from a lead-acid battery that are used in the trucks produced by Scania CV.

1.3 Prognostics

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 intended function as defined by engineers or component owners. 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 (SOH) related parameters during the lifetime of the component.

A procedure for using estimated and predicted SOH together with the RUL definition is shown in Fig. 1.4. The SOH can be a sensor measurement of a parameter that directly relates to the health of a system or it can be an entity that is derived using sensor measurements. For instance, an example of a SOH definition is given in (Kim, Lee, and Cho, 2011) where SOH of the lithium-ion batteries is a ratio of the battery resistances

$$\mathrm{SOH} = \left\| \frac{R_{\mathrm{current}} - R_{\mathrm{aged}}}{R_{\mathrm{fresh}} - R_{\mathrm{aged}}} \right\| \tag{1.1}$$

where R_{fresh} and R_{aged} are the specified battery resistances of a new and worn out battery respectively, and R_{current} is a current battery resistance which is evaluated using impedance spectroscopy technique. The SOH takes values in range from 1 to 0 for this

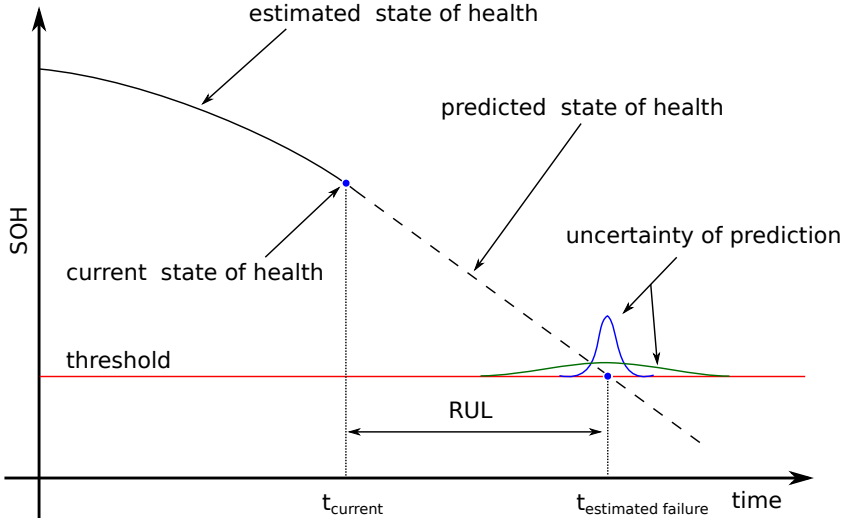


Fig. 1.4: Demonstration of the RUL concept. Estimated time of failure is when predicted state of health, dashed, curve intersects with the threshold, red line. RUL is a time difference between the estimated time of failure and the current time value. Uncertainties of the SOH prediction are denoted with two example probability densities one in blue and one in green.

example where the value 1 indicates a new battery and the value 0 - totally worn out one.

The SOH is estimated, for instance as in (1.1), and measurements up to the current time value are collected, see a black solid curve in Fig. 1.4. Then, a predictive model takes a sequence of estimated SOH values and makes a prediction for a particular number of time steps into future, denoted as the dashed curve in Fig. 1.4. An estimated time of failure is defined by intersection of the predicted SOH curve and the threshold as shown in Fig. 1.4, where the predictive model and the threshold value is determined using experimental data. The threshold value is not necessarily selected to predict failure time, but rather the time of component replacement which means that the threshold is set higher, deter-

mined by the requirement for false alarm for instance, than the threshold selected for time of failure estimation.

The RUL is the time difference between the estimated time of failure or time of replacement and the current value of time as shown in Fig. 1.4. Main challenges in estimating RUL are lack of experimental data for building the predictive models, measurement noise, and varying operating conditions that are often hard to model. Therefore, when we talk about predicting the RUL of a component, it is necessary to estimate the uncertainty of the prediction. In Fig. 1.4 the uncertainties of the prediction at a probable time of failure are illustrated with the blue and green failure time probability density functions (PDF:s). Here, the uncertainties are shown only at time of failure, but they are generally estimated at every prediction time step. In the case of the blue PDF, a model is more certain about the prediction than in the case of the green PDF, i.e., it has heavier tails of the distribution. As a result, one should probably select a more conservative threshold, i.e., earlier component replacement, when the model is uncertain.

There are, coarsely, two main categories of approaches to lifetime prognostics, namely, model-based and data-driven methods (Roemer, Byington, Kacprzynski, and Vachtsevanos, 2005). Cornerstones of model-based methods are physical laws and equations that describe degradation of the components. However, accurate predictions by model-based methods rely on detailed degradation models. It is sometimes, and this is certainly true for the lead-acid batteries, hard to develop an accurate degradation model for a particular system, and then the data-driven methods can be an alternative if reliability data is available, for example vehicle operational data in the case of lead-acid battery lifetime prognostics.

Examples of model-based prognostics are given in (Daigle and Goebel, 2011; Hanachi, Liu, Banerjee, Chen, and Koul, 2015; Saha and Goebel, 2009). The authors in (Daigle and Goebel, 2011) developed a detailed physics-based model of a pneumatic valve in a cryogenic refueling system and predicted 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, Liu, Banerjee, Chen, and Koul, 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 mentioning that these works have the possibility to either measure key parameters of the models during operation 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 here.

Data-driven models use methods from statistics, machine learning, artificial intelligence, genetic algorithms and other fields to either estimate RUL, the health of the component, or prognostic related information, for example whether a component will fail in a specified period of time. Examples of different approaches can be found, for instance, in (Cheng and Titerington, 1994; Fan, Nowaczyk, and Rögnvaldsson, 2015; Ishwaran, Kogalur, Blackstone, and Lauer, 2008; Medjaher, Tobon-Mejia, and Zerhouni, 2012; Prytz, Nowaczyk, Rögnvaldsson, and Byttner, 2015). For example, an approach based on Hidden Markov Models is proposed in (Medjaher, Tobon-Mejia, and Zerhouni, 2012) to predict the degradation of bearings. There, it is assumed that the observation probability densities are a mixture of Gaussian densities 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. In (Ishwaran, Kogalur, Blackstone, and Lauer, 2008) an ensemble or a collection of tree-based models, Random Survival Forest, method is presented that can be used for prognostic purposes. The method predicts probability of a component being operational at a particular time into the future and partitions data into groups where components have similar degradation profile. For the battery case, this corresponds to partitioning the set of vehicles into the subsets where the vehicles within a subset have

similar battery degradation properties. Another non-linear ensemble tree-based method for regression and classification problems, called Random Forest, is used in a data-driven approach proposed by the authors in (Prytz, Nowaczyk, Rögnvaldsson, and Byttner, 2015) to plan maintenance workshop visits. The type of data set, which is available for the study in (Prytz, Nowaczyk, Rögnvaldsson, and Byttner, 2015), is similar to the data set used in this thesis, however, the vehicles from the study in (Prytz, Nowaczyk, Rögnvaldsson, and Byttner, 2015) have more frequent data readouts.

Nowadays, hybrid methods, a fusion of model-based and data-driven, are proposed. For example, in (Zhao, Tian, Bechhoefer, and Zeng, 2015) an integrated prognostic method is demonstrated that uses Paris' law Paris and Erdogan, 1963 as a degradation model (model-based approach) of a gear and Bayesian inference (data-driven approach) as the tool to address the changes in the physical model depending on operating conditions. To be effective, model-based methods require a deep understanding of the degradation process which leads to complex prognostic models. In turn, data-driven methods do not rely on physical laws, but require a large amount of experimental or sensor measurement data. Hybrid methods allow to find a compromise between the two approaches, i.e., to build less complex model-based methods and incorporate uncertainties with the help of data-driven approaches requiring less data than in a standalone data-driven model.

1.4 Scope and aim

The data under study in this thesis is mainly a collection of sensor measurements that are retrieved when a vehicle comes to a workshop for a maintenance visit. Data is irregular and non-equidistant, i.e., the number of maintenance visits vary from vehicle to vehicle together with the time between the workshop visits. For example, the maximum number of readouts per truck in the database is three. This is due to the fact that maintenance visits

are scheduled by the fleet owners. The visits are more frequent if there is a contract with Scania AB that reduces maintenance costs during a particular period of time, and the number of workshop visits can be less frequent if no such contract is offered. Thus, it is not possible to get a long and equidistant in time sequence of measurements per vehicle during battery lifetime and, in addition, there is no access to information directly related to the battery health.

Therefore, the main aim of the work is to develop a data-driven prognostic framework that can be applied primarily to the lifetime prognostics of lead-acid batteries in heavy-duty trucks. At the same time, the framework should be generic, i.e., be applicable to other components of a vehicle with similar data structure as in the given study. There are uncertainties in the data, for example a significant missing data rate and the fact that the failure dates of the batteries are known to happen within a time window, whereas the exact failure date is not always available.

Taking into account the uncertainties in the data, a conditional probability function is selected as the target function to estimate for the prognostic framework instead of RUL. The conditional probability function being estimated is the probability for a random variable T, the time of a battery failure, to 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. The conditional probability function is formally written as

$$\mathcal{B}^{\mathcal{V}}(t; t_0) = P(T > t + t_0 \mid T \geq t_0,\ \mathcal{V}) \tag{1.2}$$

where $\mathcal{V}$ are the measurements from the vehicle retrieved at the workshop at time point t_0.

Examples of the lifetime functions $\mathcal{B}^{\mathcal{V}}(t; t_0)$ for batteries with different age are shown in Fig. 1.5. Similar to the RUL and SOH curve from Fig. 1.4 it is possible to set up a threshold, which has value 0.9 in Fig. 1.5, such that a crossing of the lifetime function with the threshold corresponds to the estimated time of battery

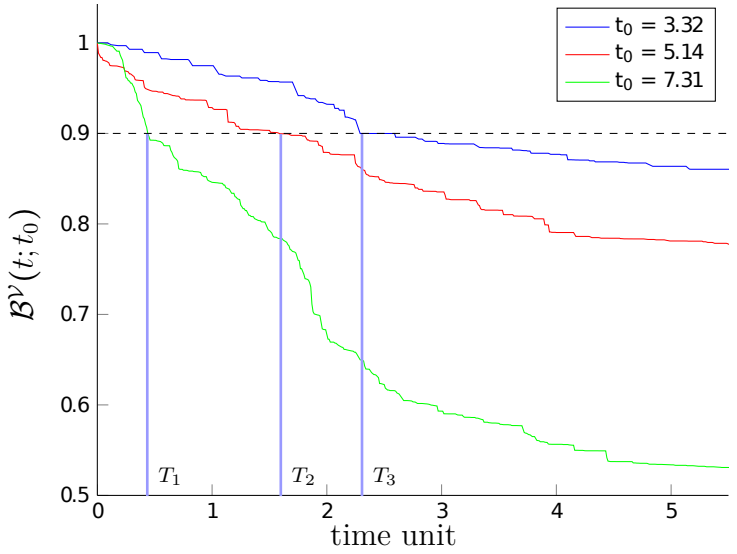


Fig. 1.5: Estimation of the battery lifetime functions $\mathcal{B}^{\mathcal{V}}(t;t_0)$ for three vehicles with different age. Threshold at the value 0.9 indicates that if the lifetime function $\mathcal{B}^{\mathcal{V}}(t;t_0)$ reaches the given value, the battery should be replaced.

failure, denoted as time T_1, T_2 and T_3 in the figure. Here, the battery that corresponds to the green lifetime function is replaced first and the battery associated with the blue curve is changed last. The replacement of the batteries in the given example is connected with the age of the vehicle, i.e., the battery is replaced first in the most driven vehicle. However, there are many other aspects that affect battery degradation. How to estimate the lifetime function with different machine learning methods and set up the threshold is described in the research papers. The scope of the work is defined by the research questions which are presented below.

Correspondence between the RUL and the lifetime function $\mathcal{B}^{\mathcal{V}}(t;t_0)$ with some assumptions is given in (1.3) and the derivation of the relation is given in Chapter 3 in the form of Proposition 1. The lifetime function $\mathcal{B}^{\mathcal{V}}(t;t_0)$ is related to the expected RUL as

$$E(\text{RUL}) = \int_0^\infty \mathcal{B}(\tau;t_0)\,d\tau. \tag{1.3}$$

1.5 Research questions

As mentioned in Section 1.4, the vehicle operational data is a collection of sparse and non-equidistant data readouts without access to the true degradation profiles of the batteries. This poses serious challenges for the predictive maintenance of the component, and how to develop and validate prognostic models is a major part of the thesis.

Different machine learning methods are used in the thesis to estimate the lifetime function $\mathcal{B}^{\mathcal{V}}(t; t_0)$, for example, one of them is Random Survival Forest. As mentioned above, it is an ensemble tree-based method with main advantages such as: can be applied to the data without preprocessing and can be applied to the type of data where failures are only available for a small number of batteries, a reduced variance (property of the ensemble) compared to the single tree model.

The number of data readouts per vehicle is low, however the number of vehicles in the database is relatively large, here about 50,000. In this case, it is worth to investigate how neural and/or recurrent networks, that is capable of finding complex dependencies in data (model contains a large number of parameters), can be applied to the vehicle operational dataset.

It is common for the data-driven predictive models that the number of inputs to the model, i.e., variables, are large. Therefore, it is of interest to find and use only informative variables for the predictive model as this increases interpretability of the model and makes it simpler.

Sensor measurements that are retrieved from the vehicles are stored in the database in the form of histograms that accumulate measurements from the beginning of vehicle usage until the time of the maintenance visit to a workshop. This nature of the data is investigated in the thesis giving suggestions regarding the usage of this type of data in the predictive models.

Aforementioned directions in the thesis are summarized in the form of research questions and presented below.

1. How should a framework for predictive maintenance be constructed that is generic and handles infrequent and non-equidistant data readouts?
2. How informative variables for predictive models should be selected?
3. How to estimate a variance of a predictive model?
4. How to validate the models for predictive maintenance when the ground truth of the component degradation profile is not available?
5. How the accumulative nature of the data influence predictive models?

1.6 Contributions

The main contributions are summarized for each included paper. The first author in each paper contributed the majority of the research work and writing.

Heavy-duty truck battery failure prognostics using random survival forests

Authors: Sergii Voronov, Daniel Jung, and Erik Frisk

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

The paper introduces a new method for identifying important variables for battery failure prognosis in the case of highly correlated feature space using Random Survival Forest model. The method is applied to the vehicle operational data which is given by an industrial partner Scania CV. There is large amount of data from trucks in operation, however, data is not closely related to battery health

which makes battery prognostic challenging. Identifying important variables and significantly reducing feature space helps data scientist to understand and interpret the process behind battery degradation in a better way compared to the case when all features are used. It is shown in the paper why the proposed method is applicable in the case of highly correlated feature space and the results of applying the method to the given dataset are compared to the existing methods such VIMP and minimal depth. Validity of the results from the proposed variable selection method is confirmed by generating prognostic model and analyzing its predictions.

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

Authors: Sergii Voronov, Daniel Jung, and Erik Frisk

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

The paper is a continuation of the first paper on variable selection. The previous paper focuses more on the automotive application case study when the given work on method development. The decision space of the proposed method is augmented which leads to the reduced set of the important variables if compared to the one from the previous paper. At the same time reducing the set of important variables does not deteriorate prognostic performance. The advantages of the proposed method are highlighted and compared to VIMP and minimal depth method on the synthetic dataset.

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

Authors: Sergii Voronov, Erik Frisk, and Mattias Krysander

Published in *IEEE Transactions on Reliability 67.2 (2018), pp. 623–639*

A data-driven method based on Random Survival Forest that estimates lifetime function instead of remaining usefull life is proposed for predicting the reliability of the batteries given that only one set of measurements per vehicle is available. Another contribution is a theory development of confidence bands for the Random Survival Forest model. The proposed approach for confidence estimation is an extension of an existing technique for variance estimation in the Random Forest method. Adding confidence bands to the RSF method gives an opportunity for an engineer to evaluate the confidence of the model prediction. Some aspects of the confidence bands are considered: their asymptotic behavior and usefulness in model selection. A problem of including time-related variables is addressed with the argument that why it is a good choice not to add them into the model. The approach is illustrated extensively using the real-life truck data case study.

Lead-acid battery maintenance using multilayer perceptron models

Authors: Sergii Voronov, Erik Frisk, and Mattias Krysander

Published in Proceedings of *IEEE International Conference on Prognostics and Health Management, Seattle, USA, 2018*

Predictive maintenance of lead-acid batteries using fully connected neural networks, i.e. multilayer perceptron models, is a topic of the paper. A battery maintenance planning method and predictive

performance evaluation based on reliability and lifetime functions is one of the conributions of the paper. Models of the reliability functions in a case when information about their true shapes is not available are introduced with a reasoning why particular models give better performance than others. An improved loss function is introduced for a neural network that predicts reliability function. It is possible to satisfy monotonically decreasing condition of a reliability function with a help of the new loss function. A neural network architecture that handles imbalanced heterogenous data, i.e. a mixture of numerical and categorical variables, is suggested and performance of different neural networks is evaluated on the vehicle operational data.

Predictive maintenance of lead-acid batteries in the presence of sparse vehicle operational data

Authors: Sergii Voronov, Mattias Krysander, and Erik Frisk

Submitted to a journal

The paper presents a development of two methods which are based on Long Short-Term Memory (LSTM) network and Random Survival Forest (RSF) to estimate time of battery failures for sparse and non-equidistant vehicle operational data. The vehicle operational dataset is a collection of readouts of vehicle sensors during their visits to a workshop. The dataset has three characteristics: 1) there are no sensor measurements directly related to battery health, 2) the number of data readouts vary from one vehicle to another and 3) readouts are collected at different time periods. First, missing data is addressed by comparing different imputation approaches together with performance of multilayer perceptron and RSF models for the case of one data readout per vehicle. Second, RSF- and LSTM-based models are suggested for the case of sparse vehicle operational data with comparison of their per-

formances. Third, model performance dependency on amount of vehicle information is discussed.

Forest-based algorithm for selecting informative variables, an automotive use-case

Authors: Sergii Voronov, Daniel Jung, and Erik Frisk

Submitted to a journal

This paper extends and summarizes the first two papers. Dependence between number of suggested informative variables and model performance is demonstrated. This is done for Random survival forest and neural network models, demonstrating similar dependency for both of them which indicates applicability of the method for other models apart of Random Survival Forest. The results confirm that a decision space augmentation which is done in the second paper leads to the models with better performance. An algorithm for automatic variable selection is formulated and advantages before other variable selection methods are demonstrated.

1.7 Thesis outline

The thesis has the following outline. Taking into account that the given work is inspired by the data that is available from the industrial partner, a reader is familiarized with the data in Chapter 2. Machine learning models are used to estimate the lifetime function of a battery and the theoretical basis required for understanding the material in the attached papers is given in Chapter 3 followed by the six research publications outlined in Section 1.6.

2 Vehicle operational data

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

Throughout the project we have had access to three versions of datasets that share common characteristics, but also have some differences such as the number of vehicles, dimensionality of the feature space, the amount of available information per vehicle, etc. All differences and similarities among the datasets are presented below. However, we begin the description with presenting the main common characteristics.

2.1 Data description

First, the data under study is either static, i.e., there is only one data readout of measurements for each vehicle, or it is sparse, i.e.,

with only a few non-equidistant in time data readouts per vehicle. Two out of the three datasets have static data, and only the last dataset contained several readouts for every vehicle. This means that it is not possible to track battery lifetime related measurements during a vehicle's lifetime and this greatly influences the method development process.

Another distinctive feature of the datasets 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. This issue is addressed in the papers by carefully selecting and developing model validation techniques.

Another important characteristic of the datasets is the fact that a majority of the batteries did not experience a failure during the observation period. The situation is common in survival analysis (Cox and Oakes, 1984) where the entities that have time of event of interest outside observational period are called *censored*. It is possible to ignore censored batteries and build predictive models based on only failed cases. However, this introduces a bias to the failed batteries, i.e., an estimation of the battery life will be pessimistic which results in many batteries being replaced when they are operational, and results in wasting a lot of useful information from the censored batteries. How to deal with censored batteries is an important part of the method development in the thesis. The censoring rate, i.e., the fraction of censored batteries, varies between the datasets and respective values can be found in Table 2.1.

Every data readout, i.e., a single record in a dataset, is logged into a database when a vehicle visits a workshop and the truck is connected to a remote diagnostic system. There is static and time varying information in every readout. Static data is given in the form of discrete/categorical variables and represent specification of a vehicle, i.e., how it is assembled, for example engine model, battery mount position, if kitchen equipment is present and so on. As an example, a battery position variable takes one out of three possible values: right, left mount point and rear frame position.

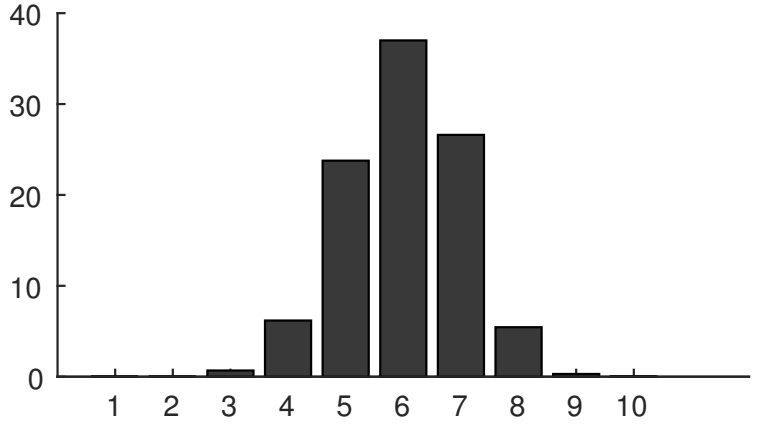


Fig. 2.1: Battery voltage 10 bins histogram.

Time-varying information from a readout is sensor measurements that are aggregated into histograms, Fig. 2.1, and retrieved at a workshop. Here, every bin in a histogram shows how a vehicle has been operated during the period up to the visit to the workshop when the data is collected. For example, for a battery voltage histogram in Fig. 2.1, each bin shows what fraction of time, given in per cents, out of the total operational time a vehicle has been operated in voltage range that corresponds to a particular bin. Notice that the bin values do not show the immediate value of the battery voltage, but rather an accumulated value during the lifetime of a component. Every bin of the histogram is treated as a separate variable and therefore the voltage histogram contributes with 10 variables to the dataset. Examples of other types of histograms that are encountered in the datasets are ambient temperature, fuel consumption, distance traveled between stops. There are also two dimensional histograms, for example a battery voltage vs temperature and engine speed vs load.

It should be noted that measurements directly related to battery health are not available. Consider the case of the voltage histogram where the logged voltage is the value right before the vehicle is switched on. The measured value of the voltage is not the open circuit voltage for many of the vehicles, because the bat-

tery requires hours of relaxation time to get to the open circuit voltage and this does not happen for distribution trucks, for example, as they make many short stops in a city. In addition, the values of battery voltage are accumulated during the battery lifetime as mentioned above. As a result, the measurements that are available do not carry enough information to build a health index for a battery.

The rate of missing data in the data readouts in all datasets is significant, totaling to about 40%. Missing data is not uniformly distributed among the variables, i.e., particular variables can have much higher missing rate than others. Missing data is a common problem for databases in automotive industry. A main reason is that initially the databases were not developed for prognostic purposes. Therefore, the requirements for data collection are not well defined and it is common that there is no primary or foreign keys in the data tables and records are missing for several reasons.

There are two main reasons for missing data in the Scania CV databases. First, when a vehicle is connected to a remote diagnostic system at a workshop, errors can occur during the process of sending data from Electronic Control Unit (ECU), which is responsible for storing particular sensor measurements, to the database and data is not logged in properly. However, missing data due to the aforementioned reason is a minority of cases. The majority of missing data cases are caused by varying sensor installations on the vehicles. Different families of vehicles may have different sensor setups and it is possible that some sensor is not installed any more or vice versa, a new sensor is introduced. The problem of dealing with missing data is addressed in the research publications.

Three datasets that represent three versions of similar information are extracted from Scania CVs databases. The datasets have information from vehicles operating in 5 European markets: Sweden, Germany, Belgium, Netherlands and France. For example, the first data set contains 33,603 vehicles each with 284 variables in a data readout, while the latest dataset has data from 46,974 vehicles with 417 variables in a data readout. The first

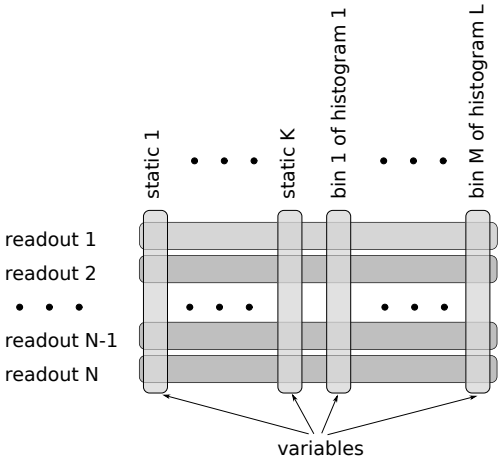


Fig. 2.2: Structure summary of a dataset that is prepared for being used in a predictive model.

dataset contains only one data readout per vehicle while the third includes several data readouts per vehicle. Battery failures are observed for only a fraction of the vehicles and this corresponds to a high censoring rate, see Table 2.1 for more details and Fig. 2.2 for summary of a dataset structure.

2.2 Building the vehicle dataset

A first step before starting to use a dataset for building the predictive models, there are steps to be performed to extract information from either structured or unstructured databases and prepare data for the analysis. This is the case with the datasets summarized in Table 2.1 where three different sources were used. These three sources are three databases, here denoted: *VehicleConfiguration* database, *VehicleOperation* database and *WorkshopInformation* database. The *VehicleConfiguration* database provides information regarding assembling of a vehicle and this is a source of static information in each data readout. Histogram variables such as battery voltage or fuel consumption mentioned above are contained in

Table 2.1: Summary of the 3 datasets that are considered in the papers of the thesis. Abbreviation EU stands for European union and 5 EU means five European markets: Sweden, Germany, Belgium, Netherlands and France. The sign # stands for "number".

Characteristics	Dataset #1	Dataset #2	Dataset #3
markets	5 EU	5 EU	5 EU
# of vehicles	33,603	54,163	46,974
# of variables	284	536	417
# of data readouts per vehicle	single	single	multiple
total # of data readouts	33,603	56,163	115,342
# of vehicles with 1 data readout	33,603	56,163	5,192
# of vehicles with 2 data readouts	-	-	15,196
# of vehicles with 3 data readouts	-	-	26,586
test set availability	no	yes	yes
# of censored batteries in test set	-	1,000	1,000
# of failed batteries in test set	-	1,000	1,000
heterogeneous data	yes	yes	yes
missing data	yes	yes	yes
censoring rate	90	80	80

VehicleOperation database. Information regarding battery repairs is logged into the *WorkshopInformation* database. Every database consists of set of tables and many of them contain information not relevant for the predictive maintenance. Therefore, the first stage is an exploration phase where key tables for the battery problem are identified in each database.

Information from the three databases is joined using a unique vehicle identifier that can be accessed in some of the tables in the aforementioned databases. Thus, the next step in the data preparation process is to set a data observation period that provides vehicle identifiers of interest and limit the amount of data to be

joined while merging information from several tables. A database for prognostic purposes of this kind can be rather big, e.g., gigabytes or even terabytes of data, and restricting the size of the population, at least during the initial exploration phase, is an important step. For example, the tables in the *VehicleConfiguration* database have tens of millions rows which is not a problem to work with in terms of consumed computational resources, however, the tables in the *VehicleOperation* database already contain tens of billions rows. Therefore, first extracting variables of interest and then limiting the result that corresponds to the target vehicles is inefficient in terms of computational time and memory.

When the vehicle population is defined by selecting identifiers that fall into a particular time period, sensor measurements from the *VehicleOperation* database can be extracted. Here, the records of values for every histogram and its bins are logged into the tables as rows, i.e., one column in a table has information about all histograms and their bins. However, one row in a dataset to be used in the machine learning models should contain measurements from all the histogram bins. Therefore, a process called pivoting is performed for the tables in the *VehicleOperation* database. Then, information from the tables in the *VehicleConfiguration* and the *VehicleOperation* databases is joined using vehicle identifiers that make a table with data readouts for every vehicle in the population.

The last step in data preparation process is to annotate the vehicles with failed and censored batteries using information in the *WorkshopInformation* database. If a component has failed and been replaced at a Scania CV workshop, a record in the *WorkshopInformation* database is made by a technician. Every component that is produced by Scania CV, or by a third party for Scania CV, has a unique identifier that can be used to find date and a vehicle number for which the component is replaced. In principle, this component identifier could be used to find dates of battery failures and the corresponding trucks. However, if a component is not produced by Scania CV, its identifier is not known and it is hard to find its failure. For the case of batteries, there are many

battery manufactures on the market and many clients choose to install a battery that is not produced by Scania CV. This makes the use of battery identifies for finding failures insufficient as the significant amount of failures are missed. For instance, if the Scania CV battery identifier is used to find failures, this gives 20 times higher battery failure rate in the Swedish market compared to the French market. This indicates that a vast majority of clients chose to install non-Scania battery.

When a failed component is replaced at a workshop, a technician has to leave a short description of a problem or cause for replacement. This information is present in the *WorkshopInformation* database and is used instead of component identifier to find a battery failure. First, text description related to battery problems is aggregated using the component identifier. This shows how in five markets technicians record battery problems. Patterns for the regular expressions are constructed using key words from the component description. This gives new battery identifiers that are used to find broken battery cases. With such an approach, the battery failure rate in all markets is about 20% and do not vary as much as when only the Scania CV battery identifiers are used. This result adds credibility to the proposed approach, because failure rates may vary slightly from country to country but not as much as an order of magnitude. The text description for every failed case is examined and it is confirmed that identified cases are lead-acid battery failures. At the same time, there could be some missed batteries failures due to the fact that it was changed at the non-Scania workshop or the constructed regular expression rejects a record when it describes a battery failure.

Results of battery failure identification method is shared with Scania CV experts and they confirm credebility of the approach at this time. Finding broken batteries in the *WorkshopInformation* database requires more investigation and it could be a fruitful field for a research. However, the given work stops at this stage due to the resource and time limitation.

3 Theory

The chapter introduces necessary theoretical background and methods that are used. Machine learning models that are used to estimate probability of battery failure in the thesis rely on techniques from survival analysis. Therefore, it is introduced first followed by the machinery behind Random Forest (RF), Random Survival Forests (RSF), and neural networks. These are the machine learning models that are involved in estimation of probability of battery failure. One of the contributions in the thesis is uncertainty estimation of the RSF predictors. Therefore, a full derivation of the Infinitesimal Jackknife variance estimate for bagged predictors is given in the given chapter to prepare a reader for the further development in the one of the papers.

3.1 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 Fig. 3.1 where the time of censoring is depicted with circles and time of failures with squares. Fig. 3.1 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 that 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 input data for a model.

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

$$R(t) = P(T \geq t). \tag{3.1}$$

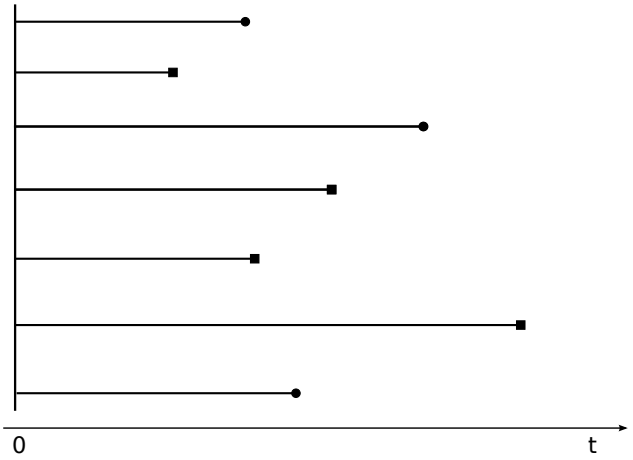


Fig. 3.1: Lifetime for failed, represented by squares, and censored, represented by circles, individuals as data for survival analysis.

The function gives the probability that the failure time T does not occur before 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

$$f(t) = -\frac{d}{dt}R(t). \tag{3.2}$$

The functions in (3.1) and (3.2) are two different ways to characterize the distribution of the failure time T. Another special notion in survival analysis is the hazard function that defines probability of instantaneous failure as

$$h(t) = \lim_{\delta \to 0+} \frac{P(t \le T < t + \delta | T \ge t)}{\delta}. \tag{3.3}$$

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 func-

tion for the random variable T with $F(t)$ and expanding (3.3) as

$$\begin{aligned} h(t) &= \lim_{\delta \to 0+} \frac{P(t \le T < t+\delta | T \ge t)}{\delta} = \\ &= \lim_{\delta \to 0+} \frac{1}{P(T \ge t)} \frac{P(t \le T < t+\delta)}{\delta} = \\ &= \frac{1}{R(t)} \lim_{\delta \to 0+} \frac{F(t+\delta) - F(t)}{\delta} = \frac{f(t)}{R(t)} = \\ &= -\frac{\frac{d}{dt}(1 - F(t))}{R(t)} = -\frac{\frac{d}{dt} R(t)}{R(t)} = -\frac{d}{dt} \log R(t) \end{aligned}$$

Then the relation between the hazard and reliability functions is

$$R(t) = \exp\left(-\int_0^t h(u)du\right) = \exp\left(-H(t)\right) \tag{3.4}$$

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 and requires less data in the model estimation. 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

$$\hat{R}(t) = \prod_{t_j<t}(1-\hat{h}_j) \tag{3.5}$$

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

$$\hat{h}_j = \frac{d_j}{r_j} \tag{3.6}$$

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 survival analysis. It estimates the variance of the reliability estimate $\hat{R}(t)$ under the assumption that (3.5) is efficient, i.e., reaches its Cramer-Rao bound and using a linearization approach, as follows

$$\operatorname{var}\left[\hat{R}(t)\right] = \left(\hat{R}(t)\right)^2 \sum_{t_j<t} \frac{d_j}{r_j(r_j-d_j)}. \tag{3.7}$$

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

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

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

$$\hat{H}(t) = \sum_{t_j<t} \frac{d_j}{r_j} \tag{3.8}$$

and according to (3.4) 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.

Relation between RUL and lifetime function

A definition of the RUL and lifetime function $\mathcal{B}^{\mathcal{V}}(t;t_0)$ is given in Chapter 1. The relationship between this two concepts is demonstrated with (1.3) and formally summarized and derived in the result below.

Proposition 1. *Let T be a continuous random variable that records time of a component failure with a density function $f(t)$. Assume also that a hazard function $h(t)$ satisfies*

$$\forall t \geq t_0, \;\; h(t) \geq h_0 > 0$$

for some positive constant h_0. Then,

$$E(RUL) = \int_0^\infty \mathcal{B}(\tau;t_0)\, d\tau.$$

Proof. First, consider the product $tP(T \geq t)$ and express it using the hazard function $h(t)$.

$$\begin{aligned} tP(T \geq t) = tR(t) = te^{-H(t)} = te^{-\int_0^t h(u)\,du} = \\ = te^{-\int_0^{t_0} h(u)\,du} e^{-\int_{t_0}^t h(u)\,du} = Cte^{-\int_0^t h(u)\,du} \leq \\ \leq Cte^{-h_0(t-t_0)} \to 0, \;\; \text{when} \;\; t \to \infty \end{aligned}$$

where C is a constant.

Thus, the hazard function assumption guarantees that the product $tP(T \geq t) \to 0$ when $t \to \infty$. Then, it holds that

$$
\begin{aligned}
E(\text{RUL}) = E(T \mid T \geq t_0) - t_0 &= \int_{t_0}^{\infty} t f(t \mid T \geq t_0)\, dt - t_0 = \\
&= \int_{t_0}^{\infty} t\, dF(t \mid T \geq t_0) - t_0 = \int_{t_0}^{\infty} t\, dP(T \leq t \mid T \geq t_0) - t_0 = \\
&= -\int_{t_0}^{\infty} t\, dP(T > t \mid T \geq t_0) - t_0 = -\int_{t_0}^{\infty} t\, d\frac{P(T > t \text{ and } T \geq t_0)}{P(T \geq t_0)} - t_0 = \\
&= -\int_{t_0}^{\infty} t\, d\frac{P(T \geq t)}{P(T \geq t_0)} - t_0 = \\
&= -\frac{1}{P(T \geq t_0)} \left(\left[tP(T \geq t) \right]_{t_0}^{\infty} - \int_{t_0}^{\infty} P(T \geq t)\, dt \right) - t_0 = \\
&= \frac{1}{P(T \geq t_0)} \int_{t_0}^{\infty} P(T \geq t)\, dt = \frac{1}{P(T \geq t_0)} \int_{0}^{\infty} P(T \geq \tau + t_0)\, d\tau = \\
&= \int_{0}^{\infty} \frac{P(T \geq \tau + t_0)}{P(T \geq t_0)}\, d\tau = \int_{0}^{\infty} \frac{P(T \geq \tau + t_0 \text{ and } T \geq t_0)}{P(T \geq t_0)}\, d\tau = \\
&= \int_{0}^{\infty} P(T \geq \tau + t_0 \mid T \geq t_0)\, d\tau = \int_{0}^{\infty} \mathcal{B}(\tau; t_0)\, d\tau
\end{aligned}
$$

which ends the proof. □

3.2 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 and, in addition, it is not clear how to estimate the parameters in the case of a parametric model such as Cox regression, (Cox, 1972). Therefore, a non-parametric approach is chosen. Random Survival Forest (RSF), (Ishwaran, Kogalur, Blackstone, and Lauer, 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 for further analysis in 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, Friedman, Olshen, and C., 1984) where the features and the 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

$$\hat{\theta}(x_i) = \hat{y}_i \tag{3.9}$$

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 the 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 data points in the given region. An example of the aforementioned process is illustrated in Fig. 3.2 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 partitioning process happens at every node of the tree, see Fig. 3.3 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 Fig. 3.3.

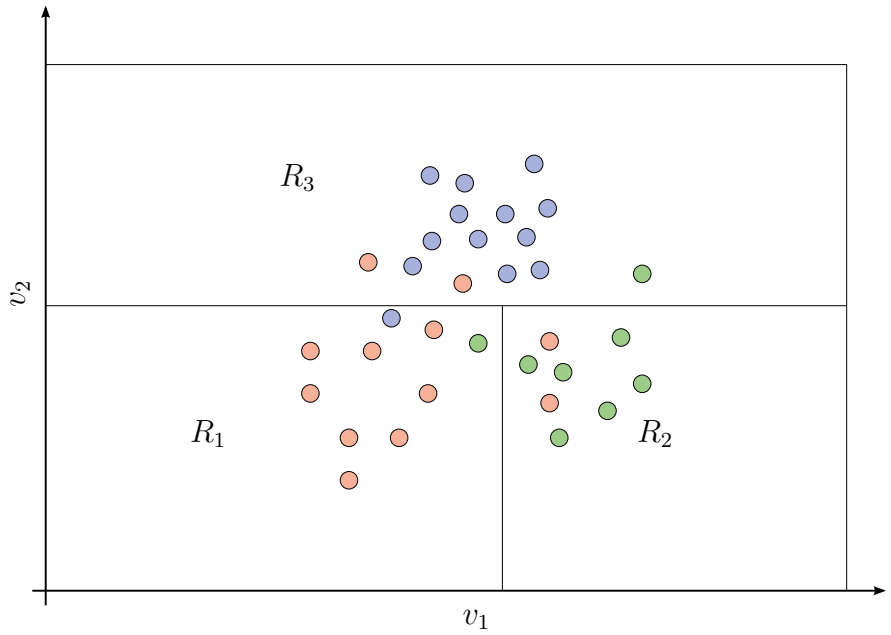


Fig. 3.2: Example of the partitioning of the space X with two variables (v_1, v_2) into disjoint regions R_i.

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, however, a weak classifier, (Hastie, Tibshirani, and Friedman, 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,

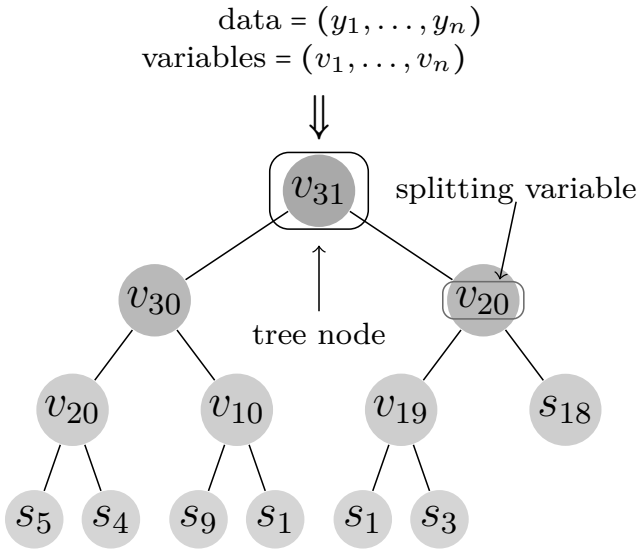


Fig. 3.3: Structure of the ordinary binary classification and regression tree.

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 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 with replacement. For instance, one realization of 3 bootstrap samples is presented below as*

$$\begin{aligned}(y_1^1, \ldots, y_5^1) &= (5.8425, 4.6918, 4.6918, -1.8254, -1.8254)\\ (y_1^2, \ldots, y_5^2) &= (3.0924, 3.0924, 5.8425, 3.9905, 4.6918)\\ (y_1^3, \ldots, y_5^3) &= (4.6918, -1.8254, 3.9905, 3.0924, 3.0924).\end{aligned}$$

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 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*

$$var\ \hat{\mu} = \frac{1}{N}\sigma_y^2 \tag{3.10}$$

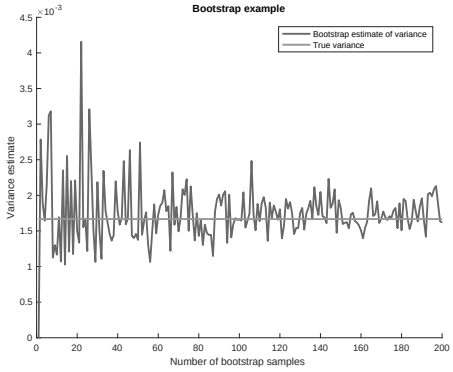


Fig. 3.4: Estimate of the estimator variance compared to the true variance as a function of number of bootstrap samples.

where σ_y^2 is the variance of the data.

The bootstrap approach, applicable to any non-linear estimator, is 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 $N = 50$. Fig. 3.4 shows how the bootstrap variance estimate compares to the theoretical value from (3.10) 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 the 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 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

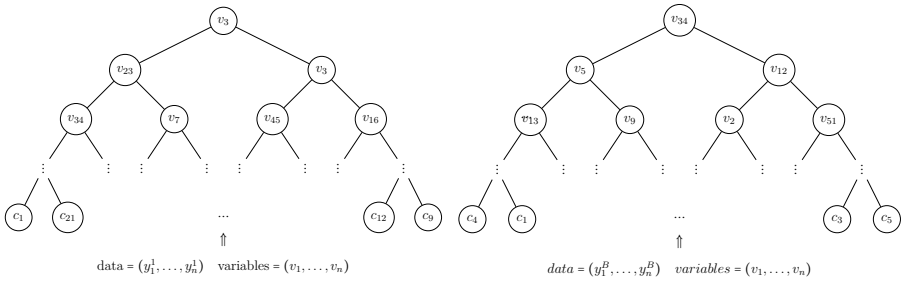


Fig. 3.5: Structure of the RF model.

aggregation model is the mean of outputs of all trees

$$\hat{\theta}_{\text{BAGG}}(x) = \frac{1}{B}\sum_{i=1}^{B}\hat{\theta}_i(x) \tag{3.11}$$

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 Fig. 3.5.

A Random Survival Forest (RSF) model is an RF model modified for the purpose of survival analysis (Ishwaran, Kogalur, Blackstone, and Lauer, 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, Thiffault, Nakache, and Asselain, 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 (3.8) 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

$$\hat{R}(t) = e^{-\hat{H}(t)} \tag{3.12}$$

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

3.3 Confidence estimate of a bagged predictor

This section describes the process behind finding Infinitesimal Jackknife (IJ) estimates of predictor variance for a 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 and are therefore included here.

Assume a bagged predictor (3.11) that is complex, nonlinear, and deriving an exact 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, Hastie, and Wager, 2014). The theoretical fundamentals are described, based on the works by Efron, Hastie, and Wager, (2014) and then extended to RSF models.

To summarize the results from (Efron, Hastie, and Wager, 2014), consider the i^{th} bootstrap sample $\boldsymbol{Y}_{\boldsymbol{i}}^{*} = (y_{i1}^{*}, y_{i2}^{*}, \ldots, y_{in}^{*})$

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 vehicle operational data introduced in Chapter 2, is included in bootstrap sample $\boldsymbol{Y}^*_{\boldsymbol{i}}$. Introduce a resampling vector as

$$\boldsymbol{P} = (p_1, p_2, \ldots, p_n) \tag{3.13}$$

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

$$\mathcal{L}_n = \left\{ \boldsymbol{P} : P_i \geq 0, \sum_{i=1}^{n} P_i = 1 \right\} \tag{3.14}$$

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 $\hat{\theta}_{\text{BAGG}}$ is

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

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

$$\hat{\theta}_{\text{BAGG}} = \hat{\theta}_{\text{BAGG}}(\boldsymbol{P}) = \hat{\theta}_{\text{BAGG}}(\boldsymbol{P}^0) + (\boldsymbol{P} - \boldsymbol{P}^0) \cdot \boldsymbol{U} + \\ + \mathcal{O}((\boldsymbol{P} - \boldsymbol{P}^0) \cdot (\boldsymbol{P} - \boldsymbol{P}^0)'). \tag{3.15}$$

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

$$U_i = \lim_{\epsilon \to 0} \frac{\hat{\theta}_{\mathrm{BAGG}}(\boldsymbol{P}^0 + \epsilon(\boldsymbol{\delta}_i - \boldsymbol{P}^0)) - \hat{\theta}_{\mathrm{BAGG}}(\boldsymbol{P}^0)}{\epsilon}, \quad i = 1, \ldots, n \tag{3.16}$$

with $\boldsymbol{\delta}_i$ being the ith coordinate vector. Taking the expectation of (3.15), ignoring higher order terms, gives

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

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

$$var\left[\hat{\theta}_{\mathrm{BAGG}}\right] \approx E\left[\hat{\theta}_{\mathrm{BAGG}}(\boldsymbol{P}^0) + (\boldsymbol{P} - \boldsymbol{P}^0) \cdot \boldsymbol{U} - \hat{\theta}_{\mathrm{BAGG}}(\boldsymbol{P}^0)\right]^2 =$$
$$= E\left[(\boldsymbol{P} - \boldsymbol{P}^0) \cdot \boldsymbol{U}\right]^2 = E\left[\left(p_1 - \frac{1}{n}, \ldots, p_n - \frac{1}{n}\right) \cdot \begin{pmatrix} U_1 \\ \vdots \\ U_n \end{pmatrix}\right]^2 =$$
$$= E\left[\sum_{i=1}^{n}\left(p_i - \frac{1}{n}\right)U_i\right]^2 = E\left[\sum_{i=1}^{n}\left(p_i - \frac{1}{n}\right)^2 U_i^2\right] +$$
$$+ 2E\left[\sum_{i \neq j}\left(p_i - \frac{1}{n}\right)U_i\left(p_j - \frac{1}{n}\right)U_j\right] = \sum_{i=1}^{n} E\left[\left(p_i - \frac{1}{n}\right)^2\right]U_i^2 +$$
$$+ 2\sum_{i \neq j} E\left[\left(p_i - \frac{1}{n}\right)\left(p_j - \frac{1}{n}\right)\right]U_i U_j = \sum_{i=1}^{n}\frac{1}{n^2}\left(1 - \frac{1}{n}\right)U_i^2 +$$
$$+ 2\sum_{i \neq j}\left(-\frac{1}{n^3}\right)U_i U_j = \frac{1}{n^2}\sum_{i=1}^{n} U_i^2 - \frac{1}{n^3}\left(\sum_{i=1}^{n} U_i\right)^2 \tag{3.17}$$

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

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

Thus, the directional derivative U_i can be expressed as

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

From this follows that

$$U_i = \left(\underbrace{-\frac{1}{n}, \ldots, -\frac{1}{n}}_{\text{i - 1}}, 1 - \frac{1}{n}, -\frac{1}{n}, \ldots, -\frac{1}{n} \right) \cdot \begin{pmatrix} D_1 \\ \vdots \\ D_n \end{pmatrix} =$$
$$= \sum_{j \neq i} \left(-\frac{1}{n} \right) \cdot \frac{\partial}{\partial p_j} \hat{\theta}_{\text{BAGG}}(\boldsymbol{P}) \Big|_{\boldsymbol{P}=\boldsymbol{P}^0} + \left(1 - \frac{1}{n} \right) \cdot \frac{\partial}{\partial p_i} \hat{\theta}_{\text{BAGG}}(\boldsymbol{P}) \Big|_{\boldsymbol{P}=\boldsymbol{P}^0} \tag{3.18}$$

It is evident from (3.18) that the sum of U_i is 0. Taking this result into account and using (3.17), the infinitesimal jackknife (IJ) variance estimate $\hat{V}_{\text{IJ}}$ of the true variance $\text{var}\left[\hat{\theta}_{\text{BAGG}}\right]$ of the bagged predictor is

$$\hat{V}_{\text{IJ}} = \frac{1}{n^2} \sum_{i=1}^{n} U_i^2. \tag{3.19}$$

To compute the variance estimator, we then need the directional derivatives. For a bagged estimator $\hat{\theta}_{\text{BAGG}}$, 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

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

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

$$\hat{\theta}_{\text{RF}}(\boldsymbol{P}) = \sum_{i=1}^{B} P(\boldsymbol{Y}_i^*) t_i^* \tag{3.20}$$

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

$$P(\boldsymbol{Y}_i^*) = \frac{P(\boldsymbol{Y}_i^*)}{P_0(\boldsymbol{Y}_i^*)} P_0(\boldsymbol{Y}_i^*) = \frac{\frac{n!}{y_{i1}^*!\ldots y_{in}^*!} p_1^{y_{i1}^*} \cdot \ldots \cdot p_n^{y_{in}^*}}{\frac{n!}{y_{i1}^*!\ldots y_{in}^*!} p_0^{y_{i1}^*} \cdot \ldots \cdot p_0^{y_{in}^*}} \cdot \frac{1}{B} = \frac{1}{B} \prod_{k=1}^{n} (np_k)^{y_{ik}^*} \tag{3.21}$$

Combining results in (3.20) with (3.21) 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

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

$$A(\epsilon) = (1 + \epsilon(n-1))^{y_{ij}^*} (1-\epsilon)^{\sum_{k \neq j} y_{ik}^*} = (1 + \epsilon(n-1))^{y_{ij}^*} (1-\epsilon)^{n - y_{ij}^*}. \tag{3.22}$$

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 (3.22) gives us

$$\hat{\theta}_{\mathrm{RF}}(\boldsymbol{P}_j) = \frac{1}{B} \sum_{i=1}^{B} \left(1 + n(y_{ij}^* - 1)\epsilon + \mathcal{O}(\epsilon^2)\right) t_i^*. \tag{3.23}$$

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

$$\hat{\theta}_{\mathrm{RF}}(\boldsymbol{P}^0) = \sum_{i=1}^{B} \frac{1}{B} \prod_{k=1}^{n} (np_0)^{y_{ik}^*} t_i^* = \frac{1}{B} \sum_{i=1}^{B} t_i^*, \tag{3.24}$$

gives the result for computing the directional derivatives U_j, (3.16), as

$$U_j = \lim_{\epsilon \to 0} \frac{\hat{\theta}_{\mathrm{RF}}(\boldsymbol{P}^0 + \epsilon(\boldsymbol{\delta}_j - \boldsymbol{P}^0)) - \hat{\theta}_{\mathrm{RF}}(\boldsymbol{P}^0)}{\epsilon} =$$

$$= \lim_{\epsilon \to 0} \left(\frac{\frac{1}{B}\sum_{i=1}^{B} t_i^* + \frac{1}{B}\sum_{i=1}^{B} n(y_{ij}^* - 1)\epsilon}{\epsilon} + \frac{\frac{\mathcal{O}(\epsilon^2)}{B}\sum_{i=1}^{B} t_i^* - \frac{1}{B}\sum_{i=1}^{B} t_i^*}{\epsilon} \right) =$$

$$= \frac{n}{B}\sum_{i=1}^{B} \left(y_{ij}^* - 1\right) t_i^*. \quad (3.25)$$

Let us show that the result for U_j in (3.25) can be expressed as

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

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

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

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

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

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

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

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

$$\hat{V}_{\mathrm{IJ}} = \sum_{i=1}^{n} \widehat{\mathrm{Cov}}_i^2 \tag{3.26}$$

where

$$\widehat{\mathrm{Cov}}_i = \frac{1}{B} \sum_{b=1}^{B} (y_{bi}^* - 1)(t_b^* - \bar{t}). \tag{3.27}$$

In general, estimators are biased and it would be great to take that 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 $B \to \infty$,

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

where Cov_j means perfect covariance estimate of (3.27) when $B \to \infty$. Thus, the bias of the estimator can be written as follows

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

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

$$\begin{aligned}\mathrm{Bias} &= \sum_{j=1}^{n} \mathrm{var}\left[\widehat{\mathrm{Cov}}_j\right] = n \cdot \mathrm{var}\left[cov_1\right] = \\ &= n \cdot \mathrm{cov}\left[\frac{1}{B} \sum_{i=1}^{B} \left(y_{i1}^* - 1\right)\left(t_i^* - \bar{t}\right) ; \frac{1}{B} \sum_{j=1}^{B} \left(y_{j1}^* - 1\right)\left(t_j^* - \bar{t}\right) \right] = \\ &= \frac{n}{B^2} \sum_{i=1}^{B} \sum_{j=1}^{B} \Big(E\left[(y_{i1}^* - 1)(y_{j1}^* - 1)(t_i^* - \bar{t})(t_j^* - \bar{t})\right] - \\ &\qquad - E\left[(y_{i1}^* - 1)(t_i^* - \bar{t})\right] E\left[(y_{j1}^* - 1)(t_j^* - \bar{t})\right] \Big). \end{aligned} \tag{3.28}$$

Assuming that original sample $\boldsymbol{Y}$ is large enough and that t_i^* and y_{ij}^* are independent, (3.28) simplifies as

$$\text{Bias} = \frac{n}{B^2}\sum_{i=1}^{B}\sum_{j=1}^{B}\Big(E\left[(y_{i1}^*-1)(y_{j1}^*-1)\right]E\left[(t_i^*-\bar{t})(t_j^*-\bar{t})\right]- \\ -E\left[(y_{i1}^*-1)\right]E\left[t_i^*-\bar{t}\right]E\left[(y_{j1}^*-1)\right]E\left[t_j^*-\bar{t}\right]\Big). \quad (3.29)$$

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

$$\text{Bias} = \frac{n}{B^2}\sum_{i=1}^{B}\Big(E\left[(y_{i1}^*-1)^2\right]E\left[(t_i^*-\bar{t})^2\right]\Big) = \frac{n}{B^2}\sum_{j=1}^{B}(t_j^*-\bar{t})^2 \quad (3.30)$$

An improved unbiased estimator of $\operatorname{var}\left[\hat{\theta}_{\text{RF}}\right]$ using the results in (3.26) and (3.30) can then be written as

$$\hat{V}_{\text{IJ-U}} = \hat{V}_{\text{IJ}} - \frac{n}{B^2}\sum_{b=1}^{B}(t_b^*-\bar{t})^2. \quad (3.31)$$

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 (1.2). The lifetime function could be expressed through the ratio of the reliability function estimates as

$$\hat{\mathcal{B}}^{\mathcal{V}}(t,t_0) = \frac{\hat{R}^{\mathcal{V}}(t+t_0)}{\hat{R}^{\mathcal{V}}(t)} \quad (3.32)$$

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 $\text{var}\left[\hat{\theta}_{\text{RSF}}\right]$ becomes

$$\hat{V}_{\text{IJ}}^{\text{RSF}}(t) = \sum_{i=1}^{n} \widehat{\text{Cov}}_i^2(t) \tag{3.33}$$

where

$$\widehat{\text{Cov}}_i(t) = \frac{1}{B}\sum_{b=1}^{B}(y_{bi}^* - 1)(\hat{R}_b^{\mathcal{V}}(t) - \hat{R}^{\mathcal{V}}(t)) \tag{3.34}$$

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 (3.26) respectively. An unbiased IJ variance estimate $\hat{V}_{\text{IJ-U}}^{\text{RSF}}$ in analogy with Efron's estimate is then

$$\hat{V}_{\text{IJ-U}}^{\text{RSF}}(t) = \hat{V}_{\text{IJ}}^{\text{RSF}}(t) - \frac{n}{B^2}\sum_{b=1}^{B}(\hat{R}_b^{\mathcal{V}}(t) - \hat{R}^{\mathcal{V}}(t))^2 \tag{3.35}$$

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

$$var\left[\hat{\mathcal{B}}^{\mathcal{V}}(t, t_0)\right] \approx \left(\frac{\mu_X}{\mu_Y}\right)^2 \cdot \left(\frac{var[X]}{\mu_X^2} + \frac{var[Y]}{\mu_Y^2} - 2\frac{cov[X, Y]}{\mu_X \mu_Y}\right) \tag{3.36}$$

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 (3.32) 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 (3.36) 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 $\mathrm{var}[X]$ and $\mathrm{var}[Y]$ correspond to IJ variance estimates $\hat{V}_{\text{IJ-U}}^{\text{RSF}}(t)$ computed at time $t+t_0$ and t_0 respectively. Covariance $\mathrm{cov}[X,Y] = \widehat{\mathrm{cov}}\left[\hat{R}^{\mathcal{V}}(t+t_0), \hat{R}^{\mathcal{V}}(t_0)\right]$ 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 $\mathrm{cov}[X,Y] = \widehat{\mathrm{cov}}\left[\hat{R}^{\mathcal{V}}(t+t_0), \hat{R}^{\mathcal{V}}(t_0)\right]$ 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*

$$cov[X,Y] = cov_{Bias}[X,Y] - Bias \tag{3.37}$$

where

$$cov_{Bias}[X,Y] = \widehat{cov}\left[\hat{R}^{\mathcal{V}}(t+t_0), \hat{R}^{\mathcal{V}}(t_0)\right] = \sum_{i=1}^{n} \widehat{Cov}_i(t_0)\widehat{Cov}_i(t+t_0) \tag{3.38}$$

$$Bias = \frac{n}{B^2}\sum_{i=1}^{B}(\hat{R}_i^{\mathcal{V}}(t_0) - \hat{R}^{\mathcal{V}}(t_0))(\hat{R}_i^{\mathcal{V}}(t+t_0) - \hat{R}^{\mathcal{V}}(t+t_0)) \tag{3.39}$$

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 (3.38) and (3.39) 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. □

3.4 Neural networks

Neural networks, similarly to Random Forests, are non-linear models that estimate relationships between input and target variables. When the amount of data is large and dependencies between variables are complicated, it is possible that neural networks perform better than a tree-based model. In a feedforward/fully connected neural network, the input layer is connected with hidden layers and an output layer by the means of weights (Hopfield, 1982).

An example of a fully connected network is presented in Fig. 3.6. A fully connected network consists of layers that are connected by the weights W_i. The layers in turn are comprised of nodes, i.e., circles in Fig. 3.6. The first layer in the network is called an *input* layer and it accepts a vector of variables x, i.e. $x = (x_1, x_2, x_3)$ in Fig. 3.6. The rest of the layers in the network are called *hidden* where the last layer is an *output* layer, because it estimates the target variable y.

It can be seen in Fig. 3.6 that every hidden node in a hidden layer performs two data transformation operation with the input

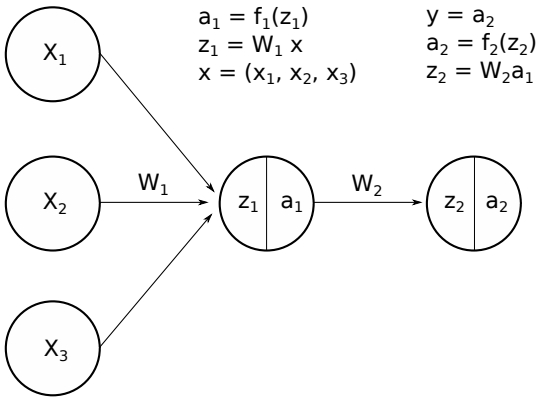


Fig. 3.6: Illustration of feedforward or fully connected neural network.

data. The first transformation is a linear weighting of the input variables, i.e., represented by two vectors $z_1 = W_1 \cdot x$ and $z_2 = W_2 \cdot a_1$ in Fig. 3.6. Then, the result of linear transformation is fed into functions f_i that are called *activation* functions (Goodfellow, Bengio, and Courville, 2016). In general, the activation functions are non-linear, except a case of a regression problem where the nodes in the last layer have a linear activation function. A non-linear nature of the activation function allows a neural network to capture complex interdependencies between the variables, and if all activation functions are linear a neural network becomes a linear model.

Three main choices of the activation functions that have been used in the past and are used today are: sigmoid, tangent hyperbolic and rectified linear unit (RELU). A sigmoid function is defined by

$$f(x) = \frac{1}{1 + e^{-x}}, \tag{3.40}$$

a tangent hyperbolic function as

$$f(x) = \frac{e^{x} - e^{-x}}{e^{x} + e^{-x}}, \tag{3.41}$$

and a rectified linear unit function

$$f(x) = \begin{cases} x, & x \geq 0 \\ 0, & x < 0. \end{cases} \tag{3.42}$$

RELU activation function is commonly used by neural network practitioners and that is due to two reasons. One is that it is easy to compute gradient of the function, i.e., it is 1 or 0, which is used during the training phase. The second reason is that the RELU function increase gradient propagation which is also needed during training.

Given the data and set of weights θ, for example (W_1, W_2) in Fig. 3.6, we want to find such values of θ thus describe the observed data as well as possible and at the same time generalize well for unseen data. This is accomplished by minimizing a loss function $L(\theta; \text{data})$ with respect to the weights θ following a backpropagation procedure (Rumelhart, Hinton, and Williams, 1986). For a regression problem, a mean squared error function, presented below, can in the simplest case be used

$$L(\theta; \text{data}) = \frac{1}{N} \sum_{i=1}^{N} (y_i - \hat{y}_i)^2 \tag{3.43}$$

where N is a size of data, y_i and $\hat{y}_i$ are a target and estimated by a neural network values respectively. In case a binary classification problem is considered, a loss function which is often used is the cross entropy loss

$$L(\theta; \text{data}) = -\frac{1}{N} \sum_{i=1}^{N} \left[y_i \log \hat{y}_i + (1 - y_n) \log(1 - \hat{y}_i) \right]. \tag{3.44}$$

There are other loss functions suggested in the scientific community for the different problems. It is not uncommon that a problem under study requires to develop a new loss function which is the case for the problem considered here.

Procedure for training a neural network, i.e., finding the values of the weights that generalize well on the unseen data, is called

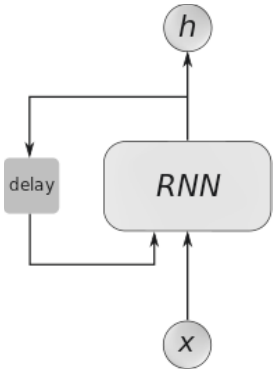


Fig. 3.7: Illustration of an RNN network where x is a sequence of input data and h is a sequence of output from an RNN layer.

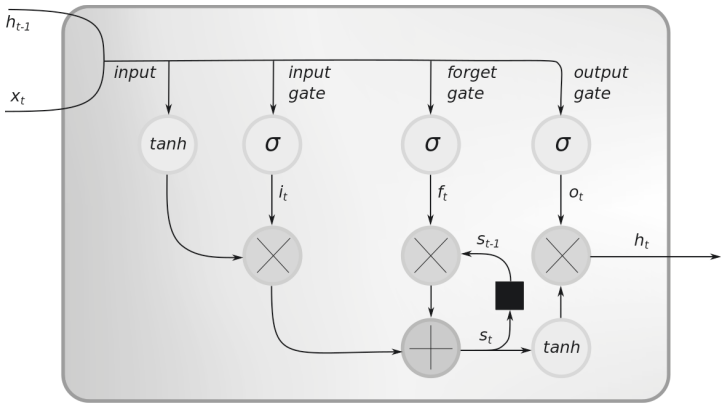


Fig. 3.8: Illustration of an LSTM cell. The sign ⊗ stands for element wise multiplication, σ is a sigmoid activation function, tanh is a tangent hyperbolic activation function.

backpropagation. It is an algorithm of computing gradients of the loss function $L(\theta; \text{data})$ with respect to the weights, W_1 and W_2 in Fig. 3.6, using a chain differentiation rule. When gradients of the loss function are computed with respect to the weights, iterative optimization algorithms are used to find updates of the weights, i.e., find a new value of the weights that explain the observed data better than with old values. A widely used algorithm is *stochastic gradient descent* (SGD) (Goodfellow, Bengio, and Courville, 2016), name comes from the fact that the gradient calculated using all

dataset is approximated by the gradient that is found using only a small part of the data which is called *mini-batch*. A new value of the weights after processing one mini-batch in SGD method is as follows

$$\theta^{(t+1)} = \theta^{(t)} - \eta \frac{\partial L}{\partial \theta^{(t)}} \tag{3.45}$$

where η is a learning rate or a step size. As with loss functions, there are extensions and variants for SGD algorithm, for example, adaptive moment estimation (Adam) (Kingma and Ba, 2014). In this algorithm the weights updates depend on the values of the gradients and the second moments of the gradients from the previous iterations, where an influence of the past gradients on the weight updates is controlled by forget factors β_1 and β_2. A procedure for the weight updates is formalized as

$$\begin{aligned} m^{(t+1)} &= \beta_1 m^{(t)} + (1-\beta_1)\frac{\partial L}{\partial \theta^{(t)}} \\ v^{(t+1)} &= \beta_2 v^{(t)} + (1-\beta_2)\left(\frac{\partial L}{\partial \theta^{(t)}}\right)^2 \\ \hat{m} &= \frac{m^{(t+1)}}{1-(\beta_1)^{t+1}} \\ \hat{v} &= \frac{v^{(t+1)}}{1-(\beta_2)^{t+1}} \\ \theta^{(t+1)} &= \theta^{(t)} - \eta \frac{\hat{m}}{\sqrt{\hat{v}} + \epsilon} \end{aligned} \tag{3.46}$$

where ϵ is a small number that is introduced to avoid computational issues, i.e., division by zero.

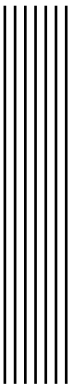
Information in the fully connected neural networks is flowing only in the forward direction. This means that there are no connections between nodes from the current layer with themselves or to the previous layers. Long Short-Term Memory networks (Hochreiter and Schmidhuber, 1997) is a class of recurrent neural networks (RNN) (Goodfellow, Bengio, and Courville, 2016) that allow feedback connections. In the recurrent neural networks an output from a layer is fed back through a delay unit as shown in

Fig. 3.7. RNNs are designed to work with sequential data such as translating text from one language to another, speech recognition, or analyzing frames in a video stream. Thus, this type of neural networks can be a good choice for a sparse and non-equidistant sequence of data readouts from a vehicle and the development and application of the RNN networks for the given data is a part of the thesis.

Training RNNs is in general difficult, in particular due to the problems of vanishing or exploding gradients when sequences of data are long or networks have many layers. LSTM networks mitigate these problems and the core building block is an LSTM cell with a schematic illustration shown in Fig. 3.8. As can be seen, the data from the previous layer x_t and the previous time step h_{t-1} are inputs to the LSTM cell. Flow of information is controlled by three gates, input, forget, and output. Each of the gates has a sigmoid activation function layer, a regular layer as in feedforward networks, which take values in the interval between 0 and 1, and multiplied elementwise with the input data vector. Values that are close to zero indicate which parts of the data that is ignored during the current time step. Conversely, values close to 1 signify active parts of data for making prediction. The most important part of the LSTM cell is the forget gate in combination with internal state s_t. The current internal state is memorized and is multiplied with the forget gate at the next time step. The interesting step happens when the result of filtering an internal state through the forget gate is combined with the result of filtering input data through the input gate. It can be seen that the results are combined by summation which is an important part in tackling vanishing gradients problem. The formula

$$\begin{aligned} f_t &= \sigma\left(W_f x_t + U_f h_{t-1} + b_f\right) \\ i_t &= \sigma\left(W_i x_t + U_i h_{t-1} + b_i\right) \\ o_t &= \sigma\left(W_o x_t + U_o h_{t-1} + b_o\right) \\ s_t &= f_t \otimes s_{t-1} + i_t \otimes \tanh\left(W_v x_t + U_v h_{t-1} + b_v\right) \\ h_t &= o_t \otimes \tanh\left(s_t\right) \end{aligned} \tag{3.47}$$

represents the LSTM-cell. Here, $\otimes$ denotes element wise multiplication of two vectors. Every input to the activation function layer has its own set of weight matrices. Dimensionality of the vectors after passing through the activation function layer is usually different from the dimensions of the vector which combines x_t and h_{t-1} and it is controlled by a network designer.



Bibliography

Ali, Jaouher Ben, Brigitte Chebel-Morello, Lotfi Saidi, Simon Malinowski, and Farhat Fnaiech (2015). "Accurate bearing remaining useful life prediction based on Weibull distribution and artificial neural network". In: *Mechanical Systems and Signal Processing* 56, pp. 150–172.

Batzel, Todd D and David C Swanson (2009). "Prognostic health management of aircraft power generators". In: *IEEE Transactions on Aerospace and Electronic Systems* 45.2, pp. 473–482.

Breiman, L. (2001). "Random forests". In: *Machine learning* 45.1, pp. 5–32.

Breiman, L., J. Friedman, R. Olshen, and Stone C. (1984). *Classification and regression trees*. Taylor and Francis. 368 pp.

Cheng, J. and D.M. Titerington (1994). "Neural Networks: A review from a statistical perspective". In: *Statistical Science* 9.1, pp. 2–54.

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

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

Cox, D.R. (1972). "Regression Model and Life-Table". In: *Journal of the Royal Statistical Society* 34.2, pp. 187–220.

Daigle, M. and K. Goebel (2011). "A model-based prognostics approach applied to pneumatic valves". In: *International Journal of Prognostics and Health Management* 2.2, pp. 1–16.

Dietterich, T.G. (2000). "An experimental comparison of three methods for constructing ensembles of decision trees: bagging, boosting, and randomization". In: *Machine Learning* 40.2, pp. 139–157.

Efron, B. (2014). "Estimation and Accuracy after Model Selection". In: *Journal of the American Statistical Assosiation* 109, pp. 991–1007.

Efron, B., T. Hastie, and S. Wager (2014). "Confidence Intervals for Random Forests: The Jackknife and the Infinitesimal Jackknife". In: *Journal of Machine Learning Research* 15, pp. 1625–1651.

Eurostat (2017). *Freight transport statistics - modal split*. https://ec.europa.eu/eurostat/statistics-explained/index.php?title=Freight_transport_statistics_-_modal_split [Accessed: January, 2020].

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

Goodfellow, Ian, Yoshua Bengio, and Aaron Courville (2016). *Deep learning*. MIT press.

Hanachi, H., J. Liu, A. Banerjee, Y. Chen, and A. Koul (2015). "A Physics-Based Modeling Approach for Performance Monitoring in Gas Turbine Engines". In: *IEEE Transactions on Reliability* 64.1.

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

Ho, T.K. (1998). "The random subspace method for constructing decision forests". In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 20.8, pp. 832–844.

Hochreiter, Sepp and Jürgen Schmidhuber (1997). "Long short-term memory". In: *Neural computation* 9.8, pp. 1735–1780.

Hopfield, John J (1982). "Neural networks and physical systems with emergent collective computational abilities". In: *Proceedings of the national academy of sciences* 79.8, pp. 2554–2558.

Ishwaran, H., U. Kogalur, E. Blackstone, and M. Lauer (2008). "Random survival forests". In: *The Annals of Applied Statistics*, pp. 841–860.

Kaplan, E. L. and P. Meier (1958). "Nonparametric Estimation from Incomplete Observations". In: *Journal of the American Statistical Association* 53.282, pp. 457–481.

Kim, Jonghoon, Seongjun Lee, and BH Cho (2011). "Complementary cooperation algorithm based on DEKF combined with pattern recognition for SOC/capacity estimation and SOH prediction". In: *IEEE Transactions on Power Electronics* 27.1, pp. 436–451.

Kingma, Diederik P and Jimmy Ba (2014). "Adam: A method for stochastic optimization". In: *arXiv preprint arXiv:1412.6980*.

Kwon, Daeil, Melinda R Hodkiewicz, Jiajie Fan, Tadahiro Shibutani, and Michael G Pecht (2016). "IoT-based prognostics and systems health management for industrial applications". In: *IEEE Access* 4, pp. 3659–3670.

Medjaher, K., D. A. Tobon-Mejia, and N. Zerhouni (2012). "Remaining Useful Life Estimation of Critical Components With Application to Bearings". In: *IEEE Transactions on Reliability* 61.2.

Miao, Qiang, Lei Xie, Hengjuan Cui, Wei Liang, and Michael Pecht (2013). "Remaining useful life prediction of lithium-ion battery with unscented particle filter technique". In: *Microelectronics Reliability* 53.6, pp. 805–810.

Nuhic, Adnan, Tarik Terzimehic, Thomas Soczka-Guth, Michael Buchholz, and Klaus Dietmayer (2013). "Health diagnosis and remaining useful life prognostics of lithium-ion batteries us-

ing data-driven methods". In: *Journal of power sources* 239, pp. 680–688.

Paris, P. and F. Erdogan (1963). "A Critical Analysis of Crack Propagation Laws". In: *Journal of Basic Engineering* 85.4, pp. 528–533. ISSN: 0021-9223.

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

Roemer, M., C. Byington, G. Kacprzynski, and G. Vachtsevanos (2005). "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.

Rumelhart, David E, Geoffrey E Hinton, and Ronald J Williams (1986). "Learning representations by back-propagating errors". In: *nature* 323.6088, pp. 533–536.

Saha, B and K. Goebel (2009). "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.

Voronov, Sergii, Daniel Jung, and Erik Frisk. "Forest-based algorithm for selecting informative variables, an automotive use-case". In: *Submitted to journal*.

Voronov, Sergii, Erik Frisk, and Mattias Krysander (2018a). "Data-driven battery lifetime prediction and confidence estimation for heavy-duty trucks". In: *IEEE Transactions on Reliability* 67.2, pp. 623–639.

— (2018b). "Lead-acid battery maintenance using multilayer perceptron models". In: *2018 IEEE International Conference on Prognostics and Health Management (ICPHM)*. IEEE, pp. 1–8.

Voronov, Sergii, Daniel Jung, and Erik Frisk (2016a). "Heavy-duty truck battery failure prognostics using random survival forests". In: *IFAC-PapersOnLine* 49.11, pp. 562–569.

— (2016b). “Variable selection for heavy-duty vehicle battery failure prognostics using random survival forests”. In: *Third European Conference of the Prognostics and Health Management Society 2016, Bilbao, July 5-8, 2016*, pp. 649–659.

Voronov, Sergii, Mattias Krysander, and Erik Frisk. “Predictive maintenance of lead-acid batteries with sparse vehicle operational data”. In: *Submitted to journal*.

Zhao, F., Z. Tian, E. Bechhoefer, and Y. Zeng (2015). “An Integrated Prognostics Method Under Time-Varying Operating Conditionds”. In: *IEEE Transactions on Reliability* 64.2.

Papers

The papers associated with this thesis have been removed for copyright reasons. For more details about these see:

http://urn.kb.se/resolve?urn=urn:nbn:se:liu:diva-162649

FACULTY OF SCIENCE AND ENGINEERING

Linköping Studies in Science and Technology, Dissertation No. 2040, 2020
Department of Electrical Engineering

Linköping University
SE-581 83 Linköping, Sweden

www.liu.se

LINKÖPING UNIVERSITY