Data-Driven Battery Lifetime Prediction and Confidence Estimation for Heavy-Duty Trucks

Sergii Voronov, Erik Frisk and Mattias Krysander

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

Sergii Voronov, Erik Frisk, and Mattias Krysander

***Abstract*—Maintenance planning is important in the automotive industry as it will allow fleet owners or regular customers to avoid unexpected failures of the components. One cause of unplanned stops of heavy-duty trucks is failure in the lead-acid starter battery. High availability of the vehicles can be achieved by changing the battery frequently, but such an approach is expensive both due to the frequent visits to a workshop and also due to the component cost. Here, a data-driven method based on Random Survival Forest (RSF) is proposed for predicting the reliability of the batteries. The data set available for the study, covering more than 50,000 trucks, has two important properties. First, it does not contain measurements related directly to the battery health, secondly there are no time series of measurements for every vehicle. In this paper, the RSF method is used to predict the reliability function for a particular vehicle using data from the fleet of vehicles given that only one set of measurements per vehicle is available. A theory for confidence bands for the RSF method is developed that 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: a) their asymptotic behavior and b) usefulness in model selection. A problem of including time related variables is addressed in the paper with arguments why it is a good choice not to add them into the model. Metrics for performance evaluation are suggested which show that the model can be used to schedule and optimize the cost of the battery replacement. The approach is illustrated extensively using the real-life truck data case study.**

***Index Terms*—Battery lifetime prognostics, flexible maintenance, reliability, infinitesimal jackknife confidence bands, data-driven prediction.**

I. INTRODUCTION

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

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

Methods for lifetime prognostics of system components can coarsely be split into two categories: model-based and data-driven methods [1]. Model-based methods rely on physical laws and equations that describe degradation of the components and for accurate predictions, accurate degradation models are required. However, it is sometimes hard to develop an accurate degradation model for a particular system, and then data-driven methods can be an alternative. It is common for both approaches to estimate the Remaining Useful Life (RUL), which is the remaining time until component failure, i.e., the point where it can no longer fulfill its function. In general, RUL is estimated using sensors that give health related information of the component, meaning, there is a possibility to track and predict the state of the health related parameters during the lifetime of the component. Examples of model-based prognostics are given in [2, 3, 4] where detailed physics-based degradation models are developed and used. Data-driven methods use machine learning algorithms to either estimate RUL, or health of the component, and can be categorized into parametric and non-parametric methods. A parametric approach assumes that the underlying degradation can be well described by a parametric distribution where the parameters of interest are estimated through the observations, see for example [5, 6]. In turn, non-parametric data-driven models use machine learning methods that do not have any basic assumption regarding underlying degradation distribution [7]. Nowadays, hybrid methods that fuse predictions both from the model-based and data-driven approaches are proposed, see [8, 9]. Unlike the aforementioned methods, where in most cases time series of sensor data is available, the data set under study only contains information retrieved from a vehicle during one of its workshop visits. Vehicle usage and environmental conditions are summarized

S. Voronov is with the Department of Electrical Engineering, Linköping university, Linköping, S-581 83 Sweden e-mail: sergii.voronov@liu.se.

E. Frisk is with the Department of Electrical Engineering, Linköping university, Linköping, S-581 83 Sweden e-mail: erik.frisk@liu.se.

M. Krysander is with the Department of Electrical Engineering, Linköping university, Linköping, S-581 83 Sweden e-mail: mattias.krysander@liu.se.

in a number of accumulative variables and histograms. For this reason, an alternative approach is adopted here where a conditional probability distribution of the battery lifetime, referred to as the battery lifetime function, is estimated instead of the RUL.

In recent decades, many works have been published regarding battery health diagnostics or prognostics. Authors in [10] and [11] give an overview of the existing methods. The majority of methods aim at establishing a battery model and estimating or measuring important battery properties such as open circuit voltage, state of charge (SoC), state of health (SoH), impedance, etc. Works cited in [12] use electrochemical impedance spectroscopy (EIS) that measures the impedance of batteries to estimate SoC and SoH. The review [12] suggests that there is a potential to use EIS in real-time systems. Examples of more data-driven methods for battery SoC estimation and prognostics are found in [4] and [13]. A particle filter approach is used in [4] to predict the RUL for any given discharge cycle of the battery and a SoC observer together with model parameter estimation and tuning is performed with the help of recurrent neural network in [13]. This work is also data-driven but a main difference compared to [4, 13] is that here no physical models or current measurements are available in this study. This changes the character of the problem significantly and is a main motivation for the work.

Given snapshots from a fleet of the vehicles coming into a workshop, the problem of estimating the lifetime function of the lead-acid battery, using a non-parametric approach, for the vehicles is considered in order to decide when to replace its battery. A lack of information directly related to the battery health is a distinctive feature of the data set. Therefore, battery health must be estimated using available information in variables correlated with battery usage. Taking this into account and considering the fact that models of battery degradation profiles are not available, a non-parametric method, Random Survival Forest (RSF) [7], is selected. The model is then used to estimate the reliability function of a particular battery and subsequently the lifetime function of the battery. Contributions in this paper are the following: a) the lifetime function is used instead of the RUL and the RSF model is proposed to estimate the lifetime function, b) a variance estimate of the predictor is suggested which uses the structure of the RSF model allowing to judge the quality of the prediction and c) an analysis of the predictive capabilities of the RSF model with different sets of input variables.

II. Problem formulation

Prognostics for flexible maintenance of batteries in heavy-duty vehicles is the topic of this study. A distinctive characteristic of the data set is that many vehicles are not observed for the full time to failure and this is referred to as censoring of failure times. The definition of censoring used here is equivalent to the one introduced in [14]. To illustrate the potential for flexible maintenance in the case under study, consider the distribution of failure and censoring times in Fig. 1 (time is scaled to avoid revealing sensitive information). The shape of the distribution of failed vehicles, the red curve in the figure, is such that it is impossible to set up an efficient maintenance point to replace the battery. If the maintenance point is scheduled, for instance, around 0.5 time units, then the majority of the batteries of the failed vehicles are replaced before failure. However, the batteries are then not used efficiently because the majority of the batteries of the censored vehicles are replaced as well as indicated by the blue curve in the figure. In addition, customers will not be satisfied with the quality of the batteries if they are changed too soon and may shift to another battery manufacturer who can deliver a better service. On the other hand, if the maintenance point is scheduled around 5 time units, a majority of the batteries are used until the end of their lives as shown by the blue curve in Fig. 1 but at the same time, this means significant numbers of battery failures with decreased reliability and uptime as a result. Therefore, the figure motivates the need for a vehicle specific prognostic model described in the paper. Before the studied problems are explicitly stated, the vehicle fleet data is introduced which is used to build the model.

A. *Vehicle fleet data*

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

Main characteristics of the database are:

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

The data set includes both categorical and numerical variables where categorical variables have a limited number of possible values. For example, the battery position variable has three possible values (right, left hand side and rear frame end). Numerical data is mostly organized in the form of histograms but there are, so called, accumulative variables such as mileage and age which increase with time. As an example, one of the histogram data is a voltage histogram that has ten bins, each showing what fraction of time the battery of the vehicle has been operating in a particular voltage range. Ranges of

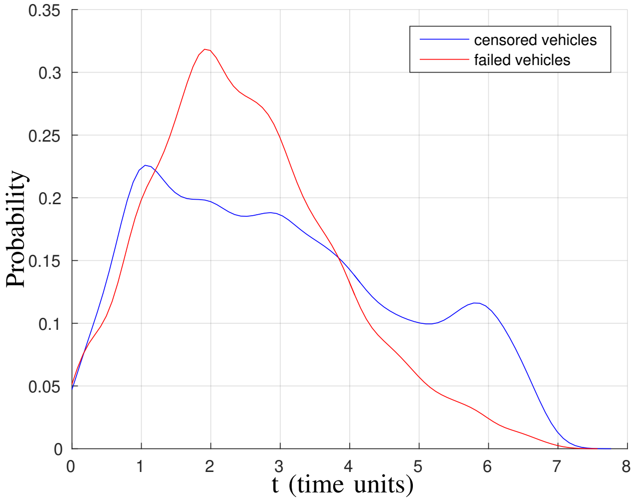


Fig. 1. Distribution of censoring time for censored, blue curve, and failed times for failed vehicles, red curve.

histogram bins are previously defined by engineers at Scania CV. Here, every bin of the histograms is treated as a separate variable and then the voltage histogram contributes with 10 variables to the study. Other examples of histograms present in the database are atmospheric pressure, ambient temperature, vehicle speed and fuel consumption vs speed that is a two dimensional histogram. In the data set under study there are 26 histograms with different number of bins. As mentioned above, the number of variables per vehicle is 536 which is the total number formed by all categorical variables and histogram bins. Percentage distribution of the categorical and histogram variables are 1.5 percent for categorical and 98.5 percent for histogram variables. The censoring rate is another distinctive property. Only a fraction of the vehicles has problems with batteries while all others do not, meaning that the failure times are censored. Missing data is also an essential characteristic of many real life data sources and the main reason in our case is the fact that variables introduced for one type of a vehicle are not relevant for another type. The missing data rate is about 40% and it should be noted that missing values are not uniformly distributed among variables. Specific variables can have significantly higher missing rate than others. Thus, systematic handling of missing data is important in the proposed approach.

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

B. Battery lifetime function

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

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

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

The function states the probability that a failure time T for a battery of interest is greater than $t + t_0$ time units given that it has survived t_0 time units conditioning on snapshot data $\mathcal{V}$. Prediction of battery lifetime can be made, for example, in the workshop when data is retrieved from the vehicle. The established reliability function $R^{\mathcal{V}}(t) = P(T \geq t \mid \mathcal{V})$, [14], is defined as a probability for a battery to survive t time units. The relationship between the lifetime function $\mathcal{B}^{\mathcal{V}}(t; t_0)$ and the reliability function $R^{\mathcal{V}}(t)$ is given directly by the definition of conditional probabilities as

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

and is used throughout the paper.

C. Estimate confidence of a predictor model

As mentioned in Section II-B, the main objective is to estimate the battery lifetime prediction function (2). To evaluate if an estimate is reliable or not, some measure of confidence is needed. A common approach is to use the confidence bands of the estimator. Here, the true estimator distribution is not known and one simple way to estimate the variance of the estimate is to make a Gaussian assumption of estimator distribution. This approach is used throughout the paper, but it is certainly possible to make other distribution assumptions, or simply form confidence bands as, e.g., $\pm$ one standard deviation.

A synthetic data set is used to show how confidence bands to an estimator can be computed in a simpler case than studied here. Assume that there are 5 classes of the vehicles with different degradation profiles of the batteries. Fig. 2 demonstrates estimation of the true reliability for one of the classes, see the magenta curve in the figure. Information about the true reliabilities is not available in the real data set, and, therefore, the synthetic data set is used to show statistical properties of the estimator. When all vehicles in a class have the same degradation profile, it is possible to compute a Kaplan-Meier estimate, a maximum likelihood estimate of the reliability function [15]. This is shown by the green curve in Fig. 2, and 95% confidence bands, based on a Gaussian assumption and a standard deviation estimated by the Greenwood formula [14], is shown by dashed blue curves. A main problem studied in the paper is how to estimate standard errors and confidence intervals for a battery lifetime function estimator. In contrast to the example where basic survival analysis is directly applicable, the data set under study has not a set of distinct degradation

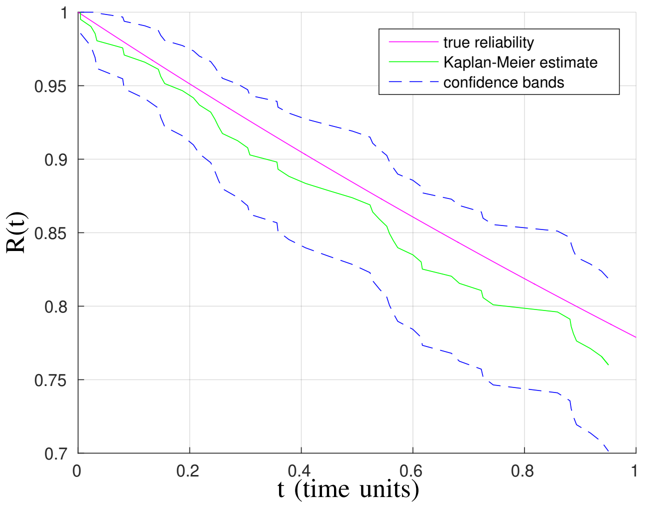


Fig. 2. True reliability function for a vehicle battery, magenta curve, Kaplan-Meier estimate for the given class of vehicles, green curve, and 95 % confidence bands with Gaussian assumption where variance estimated using Greenwood formula, blue dashed curves.

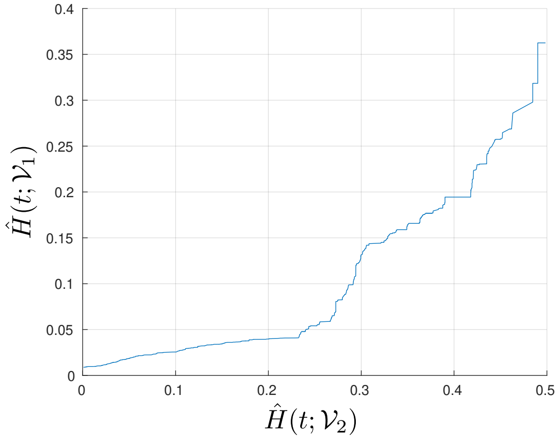


Fig. 3. Etimated cumulative hazard functions for two vehicles plotted against each other.

classes. This is an important observation and the data set covers a continuum of degradation profiles and therefore, the Kaplan-Meier and Greenwood formula are not directly applicable.

D. Summary

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

III. Lifetime prediction function model

An important first choice is which model should be used in the lifetime prediction framework. In medical studies the well known Cox regression model with the proportional hazards has proven to be useful [6, 14]. Here, instead, a non-parametric approach, Random Survival Forests (RSF), is used and a main reason concerns the proportional hazards assumption. Proportional hazards is a restrictive assumption and would limit the generality of the approach and a main objective here is to study and develop an approach that is applicable to also other components than lead-acid batteries. To motivate the choice of the non-parametric RSF model, a simple visualization of the proportional hazards assumption is done below. For a more systematic approach, see for example [16]. The hazard function, i.e., the instantaneous failure rate, is properly introduced in Section III-A, but under the proportional hazard assumption it holds that

$$H(t;\mathcal{V}_1) \propto H(t;\mathcal{V}_2)$$

where $H(t;\mathcal{V}_i)$ are the cumulative hazard functions for vehicles $\mathcal{V}_1$ and $\mathcal{V}_2$ respectively. Fig. 3 shows the non-parametric Nelson-Aalen estimates of cumulative hazard functions for two representative vehicles plotted against each other. If the proportional hazard assumption was valid, this would be a straight line, which is clearly not the case for these two vehicles. Therefore, a straightforward application of the Cox regression model is not applicable and motivates our choice of the non-parametric RSF model. Next, Random Survival Forest is briefly summarized in Section III-A and then the approach is applied to the battery prognostic case in Section III-B.

A. Random survival forests

Classification and regression trees are machine learning techniques that maps/predicts a feature or variable space X into a space of outcomes Y by means of binary trees [17] where features and outcome for a particular case are considered as a pair (x_i, y_i). Target values y_i from the outcome space could be continuous valued in case of regression and discrete in case of a classification problem. A decision tree is a non-linear estimator

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

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

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

Therefore, ensemble of trees, a Random Forest (RF) model, was successfully introduced by Breiman [18]. There are different implementations of ensemble of trees such as [19] and [20], however, the basic Breiman model is described here since the RSF model is an extension of RF. There are two techniques that are the distinctive features of the RF method, namely, bootstrap aggregation, also known as bagging, and a step that reduces correlation between trees in the forest. When number of data samples is small, bootstrap is a powerful method for estimating statistics, see [21]. By sampling from the given data samples with replacement one can construct a significantly large set of new samples that can be used to estimate target statistics. Bootstrap aggregation is an ensemble method that combines predictions from different machine learning models. In the case of trees, a number of sets of bootstrap samples are created and then a classification or regression tree model is fitted for each of bootstrap sample. As mentioned, a single tree model is sensitive to unseen data, but by combing outputs from a set of trees, grown on different bootstrap samples, the resulting output has reduced variance of a predictor compared to the single tree model. In regression, the output from a bootstrap aggregation model is the mean of outputs of all trees

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

where $\hat{\theta}_i(x)$ is a tree model fitted to the i^{th} bootstrap sample, and B is the number of trees/bootstrap samples. It was suggested by Breiman [18] that introducing randomness into the procedure of choosing variables for splitting reduces correlation between trees and increase performance of the aggregated model. Therefore, instead of choosing all m available variables for split at each node, only a fraction p of them is considered. This step also increases speed of the algorithm as it requires less variables to check at each split.

A key concept in survival analysis is the age-specific failure rate, the hazard function $h(t)$. Let T be the random failure time and t current time, then the hazard function is defined as

$$h(t) = \lim_{\delta\to 0^+}\frac{P(t \le T < t+\delta | T \ge t)}{\delta}. \tag{5}$$

The hazard function describes the probability of failure at time t given that it has survived until t. The relationship between the reliability function and the hazard function can be seen by denoting the cumulative distribution function for the random variable T with $F(t)$ and expanding (5) as

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

Then the relation between the hazard and reliability function is

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

where $H(t)$ is cumulative hazard rate.

A Random Survival Forest (RSF) model is an RF model modified for the purpose of survival analysis [7]. 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 [22]. It is a hypothesis test which compares distributions of failures of the samples that are formed by dividing data available at the splitting node into two samples which will be the part of the two child nodes. The best split corresponds to a variable with a value under which two samples have as distinctive degradation profiles as possible. The log-rank test is non-parametric and designed for censored data, a type of data encountered in survival analysis. At each terminal node, a node at which splitting no longer is performed, the Nelson-Aalen estimate of the cumulative hazard rate $H(t)$ is computed [14]. The estimated cumulative hazard rate $\hat{H}(t)$ of the whole forest is computed by averaging over tree hazard rates. The estimate $\hat{R}(t)$ of the reliability function is directly given by (6) as

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

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

B. Battery prediction model

The output from the RSF is an estimate of reliability function as in (7). Then, an estimate of the lifetime function $\hat{\mathcal{B}}^{\mathcal{V}}(t, t_0)$ can be expressed directly from (2) as

$$\hat{\mathcal{B}}^{\mathcal{V}}(t, t_0) = \frac{\hat{R}^{\mathcal{V}}(t+t_0)}{\hat{R}^{\mathcal{V}}(t_0)}. \tag{8}$$

IV. Confidence estimate for the battery lifetime prognostics function

Consider a bagged predictor (4). Such an estimator is complex, nonlinear, and deriving an explicit expression for the estimation covariance is infeasible. Then, one option is to use a bootstrap technique. Since the estimator already uses a bootstrap technique, a bootstrap strategy for estimating the variance would require to compute bootstrap of bootstraps which is computationally infeasible, [23]. Therefore, an approach that uses the original bootstrap samples used when building the model also for estimating variance is desired. One possibility for such an approach is the Infinitesimal Jackknife (IJ) variance estimate suggested in [24] for random forests. The basic approach is described in Section IV-A and then the technique will be extended to RSF and the battery lifetime function in Section IV-B. This section is technical and it is possible to go directly to Section V and Section VI and come back latter for the technical details.

A. Theoretical background on IJ variance estimation

To summarize results from [24], 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, a snapshot of the vehicle from the data set in the given study, is included in the bootstrap sample. Introduce a resampling vector as

$$\boldsymbol{P} = (p_1, p_2, \ldots, p_n) \tag{9}$$

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

$$\mathcal{L}_n = \left\{ \boldsymbol{P} : P_i \geq 0, \sum_{i=1}^{n} P_i = 1 \right\}. \tag{10}$$

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

$$\hat{V}_{\text{IJ}} = \frac{1}{n^2} \sum_{i=1}^{n} U_i^2 \tag{11}$$

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

$$U_i = \lim_{\epsilon \to 0} \frac{\hat{\theta}_{\text{BAGG}}(\boldsymbol{P}^0 + \epsilon(\boldsymbol{\delta}_i - \boldsymbol{P}^0)) - \hat{\theta}_{\text{BAGG}}(\boldsymbol{P}^0)}{\epsilon}, \quad i = 1, \ldots, n \tag{12}$$

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

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

where

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

Here, the b^{th} tree grown on the b^{th} bootstrap sample is built with the Breiman procedure, t_b^* is the output from the b^{th} tree and $\bar{t}$ is the RF output. The estimator (13) can be proven to be and an improved unbiased estimator can be derived as in [24]

$$\hat{V}_{\text{IJ-U}} = \hat{V}_{\text{IJ}} - \frac{n}{B^2} \sum_{b=1}^{B} (t_b^* - \bar{t})^2 \tag{14}$$

B. IJ variance estimate for the lifetime function

There are two main differences between IJ variance estimate of the RF model compared to variance estimate of lifetime function (8). First, the output of the RF model is either a class or regression value, but in the RSF case the output is a time dependent function, and secondly, the lifetime function is a ratio of the reliability estimates $\hat{R}^{\mathcal{V}}(t)$ as in (2).

For the first difference mentioned above, the reliability function is computed on a predefined grid of time points, i.e., time points chosen by the RSF algorithm based on the samples in the terminal node. The variance estimate $\hat{V}_{\text{IJ}}^{\text{RSF}}(t)$ of the true forest variance $\operatorname{var}\left[\hat{\theta}_{\text{RSF}}\right]$ becomes

$$\hat{V}_{\text{IJ}}^{\text{RSF}}(t) = \sum_{i=1}^{n} \widehat{\text{Cov}}_i^2(t) \tag{15}$$

where

$$\widehat{\text{Cov}}_i(t) = \frac{1}{B} \sum_{b=1}^{B} (y_{bi}^* - 1)(\hat{R}_b^{\mathcal{V}}(t) - \hat{R}^{\mathcal{V}}(t)). \tag{16}$$

Here, the reliability $\hat{R}_b^{\mathcal{V}}(t)$ is the output reliability from the b^{th} tree for a particular vehicle with data $\mathcal{V}$ and $\hat{R}^{\mathcal{V}}(t)$ is the output from the forest. These values correspond to t_b^* and $\bar{t}$ in (14) respectively. An unbiased IJ variance estimate $\hat{V}_{\text{IJ-U}}^{\text{RSF}}$ in analogy with Efron's estimate is then

$$\hat{V}_{\text{IJ-U}}^{\text{RSF}}(t) = \hat{V}_{\text{IJ}}^{\text{RSF}}(t) - \frac{n}{B^2} \sum_{b=1}^{B} (\hat{R}_b^{\mathcal{V}}(t) - \hat{R}^{\mathcal{V}}(t))^2. \tag{17}$$

For the second property, the variance estimate for the lifetime function $\hat{\mathcal{B}}^{\mathcal{V}}(t, t_0)$ from (8), which is a ratio of the outputs of the random survival forest, is estimated and summarized next.

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

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

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

$$\operatorname{var}\left[\hat{\mathcal{B}}^{\mathcal{V}}(t, t_0)\right] \approx \left(\frac{\mu_X}{\mu_Y}\right)^2 \cdot \left(\frac{\operatorname{var}[X]}{\mu_X^2} + \right. \\ \left. + \frac{\operatorname{var}[Y]}{\mu_Y^2} - 2\frac{\operatorname{cov}[X, Y]}{\mu_X \mu_Y}\right) \tag{18}$$

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

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

The random forest outputs $\hat{R}^{\mathcal{V}}(t)$ and thereby μ_X, μ_Y above, and the infinitesimal jackknife estimator (14) gives $\operatorname{var}[X]$ and $\operatorname{var}[Y]$. A result for the estimation of $\operatorname{cov}[X, Y]$ is given in Lemma 1.

Proof: From (2), the lifetime function can be expressed as the ratio of the reliability functions $\hat{R}^{\mathcal{V}}(t)$ and $\hat{R}^{\mathcal{V}}(t + t_0)$. Assume that $\hat{R}^{\mathcal{V}}(t + t_0)$ is a random variable X and $\hat{R}^{\mathcal{V}}(t_0)$ is a random variable Y. Then, the variance of the lifetime function can be estimated using a Taylor series expansion as of (18) 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 $\operatorname{var}[X]$ and $\operatorname{var}[Y]$ correspond to IJ variance estimates $\hat{V}_{\text{IJ-U}}^{\text{RSF}}(t)$ computed at time $t + t_0$ and t_0 respectively. Covariance $\operatorname{cov}[X, Y] = \widehat{\operatorname{cov}}\left[\hat{R}^{\mathcal{V}}(t + t_0), \hat{R}^{\mathcal{V}}(t_0)\right]$ is a covariance between two random variables which are represented by the values of two points from the reliability curve $\hat{R}^{\mathcal{V}}(t)$ at time $t + t_0$ and t_0. ■

The missing part and a main contribution is the derivation of $\text{cov}\,[X,Y]=\widehat{\text{cov}}\left[\hat{R}^{\mathcal{V}}(t+t_0),\hat{R}^{\mathcal{V}}(t_0)\right]$ using an infinitesimal jackknife approach. This key result is summarized in the lemma below. The proof of the lemma is given in the appendix for the interested reader. The continuation of the paper can be read without the technical details of the proof.

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 output, $\hat{R}^{\mathcal{V}}_b(t)$, from tree b is independent from one data point j from the i^{th} bag, then an asymptotic expression of the infinitesimal jackknife estimate of $\widehat{\text{cov}}\left[\hat{R}^{\mathcal{V}}(t+t_0),\hat{R}^{\mathcal{V}}(t_0)\right]$ and the corresponding bias correction are

$$\text{cov}\,[X,Y]=\text{cov}_{\text{Bias}}\,[X,Y]-\text{Bias} \tag{19}$$

where

$$\text{cov}_{\text{Bias}}\,[X,Y]=\widehat{\text{cov}}\left[\hat{R}^{\mathcal{V}}(t+t_0),\hat{R}^{\mathcal{V}}(t_0)\right]= \\ =\sum_{i=1}^{n}\widehat{\text{Cov}}_i(t_0)\widehat{\text{Cov}}_i(t+t_0) \tag{20}$$

and

$$\text{Bias}=\frac{n}{B^2}\sum_{i=1}^{B}(\hat{R}^{\mathcal{V}}_i(t_0)-\hat{R}^{\mathcal{V}}(t_0))(\hat{R}^{\mathcal{V}}_i(t+t_0)-\hat{R}^{\mathcal{V}}(t+t_0)) \tag{21}$$

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

When the prediction of the battery's lifetime in the form of the lifetime function $\mathcal{B}^{\mathcal{V}}(t,t_0)$ together with its variance estimate are available, we have a tool that is useful in maintenance planning and its usefullness is demonstrated in Section IV-C.

C. Analysis of the IJ covariance estimate

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

When $\widehat{\text{cov}}\left[\hat{R}^{\mathcal{V}}(t+t_0),\hat{R}^{\mathcal{V}}(t_0)\right]$ and bias are estimated, it is possible to plot confidence bands for an estimate of the lifetime function $\mathcal{B}^{\mathcal{V}}(t,t_0)$. Fig. 4 shows a 95% confidence band for 4 vehicles from the validation set with a Gaussian assumption for the lifetime function estimate. The RSF model used for the figure had 1000 trees. To motivate the need in estimating the $\widehat{\text{cov}}\left[\hat{R}^{\mathcal{V}}(t+t_0),\hat{R}^{\mathcal{V}}(t_0)\right]$ and the estimator bias, three types of confidence bands are plotted. Blue dashed curves are 95% confidence bands computed using the variance from (18) where biased IJ variance estimates are used, i.e., values var $\left[\hat{R}^{\mathcal{V}}(t+t_0)\right]$, var $\left[\hat{R}^{\mathcal{V}}(t_0)\right]$, and cov $\left[\hat{R}^{\mathcal{V}}(t+t_0),\hat{R}^{\mathcal{V}}(t_0)\right]$ are biased. It can be seen that when biased estimates are used in (18), the confidence bands become conservative. The black curves are 95% confidence bands computed using the variance from (18) with the unbiased IJ variance estimates var $\left[\hat{R}^{\mathcal{V}}(t+t_0)\right]$ and var $\left[\hat{R}^{\mathcal{V}}(t_0)\right]$ of reliabilities and assumption that values of $\hat{R}^{\mathcal{V}}(t)$ are independent at time point t and $t+t_0$, i.e., $\widehat{\text{cov}}\left[\hat{R}^{\mathcal{V}}(t+t_0),\hat{R}^{\mathcal{V}}(t_0)\right]=0$. The red curves are 95% confidence bands computed using variance from (18) with the unbiased IJ variance estimates var $\left[\hat{R}^{\mathcal{V}}(t+t_0)\right]$ and var $\left[\hat{R}^{\mathcal{V}}(t_0)\right]$ of reliabilities and estimated cov $\left[\hat{R}^{\mathcal{V}}(t+t_0),\hat{R}^{\mathcal{V}}(t_0)\right]$. Fig. 4 shows that for three vehicles out of four black and red curves are close to each other, however, for a vehicle in top right corner they differ significantly. This indicates the importance in finding $\widehat{\text{cov}}\left[\hat{R}^{\mathcal{V}}(t+t_0),\hat{R}^{\mathcal{V}}(t_0)\right]$ and its bias.

Confidence bands can be used for the model selection. For example, an estimate for the lifetime functions together with the confidence bands for two RSF models, one with 100 trees and one with 1000 trees, are presented in Fig. 5. It is not surprising that more trees improve the variance of the predictor. However, let us consider model selection based on the available error metric. One of the metrics measuring error available in the RSF framework is the error rate [7]. It relies on the Concordance index which counts prediction as erroneous when for two randomly selected vehicles the shorter survival time has worse predicted value of survival function. As it is shown in the previous work [25], the error rate curve starts to converge after about 100 trees and the difference between the error rates for the models with 100 and 1000 trees is negligible, and then it is tempting to stop increasing the number of trees in the RSF model. However, it is evident from Fig. 5 that the quality of the prediction in the case of 1000 trees is significantly better than in the case of 100 trees, because confidence bands are narrower. The experiment shows that adding confidence bands to the predictor helps to find better model than the one created by relying only on the error rate values.

It is evident from the results above that the unbiased covariance estimates give less conservative variance estimate of the lifetime function and, in addition, confidence bands can be used as a complimentary tool, for example, to the error rate for model selection. From now on in the paper only the unbiased covariance estimates obtained with the help of IJ technique are used when computing the confidence bands of the predictor.

V. Synthetic data set study

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

A. Parameters of RSF algorithm

Before proceeding with the description of the synthetic data, the parameters used when training the RSF model are described. There are three main parameters when building the RSF model, minimal node size, number of trees in the forest, and number of splitting variables in each node. Number of trees in the

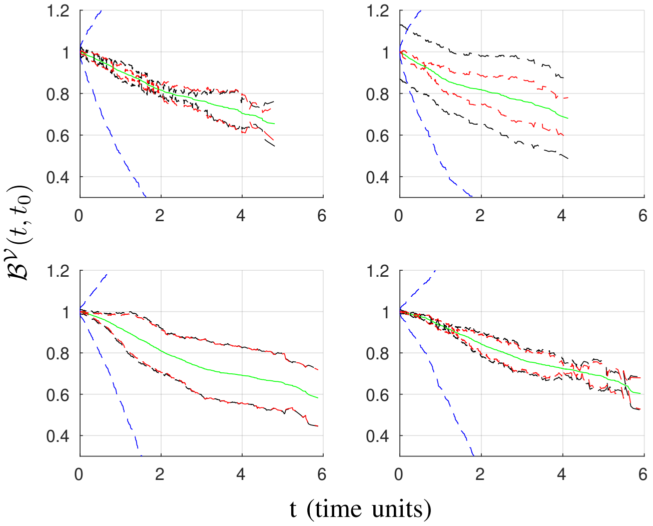


Fig. 4. IJ variance estimates of lifetime function for 4 vehicles from validation set. Green curve is an estimate of lifetime function $\mathcal{B}^{\mathcal{V}}(t,t_0)$. Blue curves are 95% confidence bands computed using variance from (18) with biased IJ variance estimates of covariance of reliabilities. Black curves are 95% confidence bands computed using variance from (18) with unbiased IJ variance estimates of covariance of reliabilities and assumption that values of $\hat{R}^{\mathcal{V}}(t)$ are independent at time point t and $t+t_0$. Red curves are 95% confidence bands computed using variance from (18) with unbiased IJ variance estimates of covariance of reliabilities.

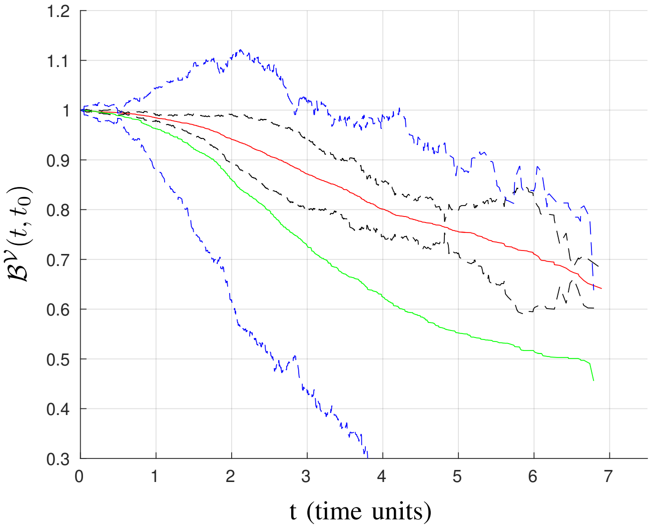


Fig. 5. Estimate of the lifetime function $\mathcal{B}^{\mathcal{V}}(t,t_0)$, green curve corresponds to the model with 100 trees and red to the model with 1000 trees, with the 95% confidence bands, blue curves correspond to the model with 100 trees and black curves to the model with 1000 trees.

forest B is chosen to be 1000 for the experiments in Section V and Section VI. The chosen value of the trees will guarantee a good quality of the prediction as shown in Section IV-C. However, number of trees in the forest can be set to values which differ from 1000 in some cases to compare the results from the different models. Minimal node size is set to value 200 and full motivation is given in [25]. The number of splitting variables m in each node is set to the default value $m=\sqrt{p}$ where p is number of all variables. The cost function used for splitting is a log-rank test [22] which is also the default value for the RSF package. Readers who are interested in the detailed description of all parameters are referred to [26].

B. Synthetic data experiments

The generated synthetic data has 6 variables and 1000 vehicles. One variable is important for prognosis as it controls the degradation of the battery. The other five variables are noisy in the sense that they do not influence the battery degradation. The battery degradation is controlled by varying the hazard rate [14], i.e., the probability of instantaneous failure at time t, according to the one important variable. The expected lifetime of the batteries with the selected nominal hazard rate is set to 10 years and it is assumed that the important variable v_1 has an impact on the battery hazard rate h as

$$h = \begin{cases} 1 \cdot h_0, & \text{if } v_1 = 1 \\ 1.5 \cdot h_0, & \text{if } v_1 = 2 \\ 2.5 \cdot h_0, & \text{if } v_1 = 3 \\ 2.9 \cdot h_0, & \text{if } v_1 = 4 \\ 3.4 \cdot h_0, & \text{if } v_1 = 5 \end{cases} \tag{22}$$

where $h_0 = \frac{1}{10}$ is the nominal hazard rate. The censoring rate is controlled to be at 80%, which is similar to the real dataset. Vehicles are uniformly distributed among the five classes, meaning each class has about 200 vehicles.

Here, in contrast to the vehicle data set, the class of each vehicle is known and then it is possible to compute the Kaplan-Meier estimate, $\hat{R}(t)$, of the reliability function, which is the maximum likelihood estimator, for every class together with confidence bands computed using the standard Greenwood formula. This corresponds to the estimates based on an ideal vehicle classifier, i.e., that perfectly separates vehicles into the 5 defined classes. The estimates for the third class of the vehicles are presented in Fig. 2. Now, let us compare the maximum-likelihood estimates with full class knowledge with the estimates from the RSF model.

First, consider a prediction for one of the vehicles in the validation set belonging to the third class. Fig. 6 shows the predictions from the forest of 1000 trees together with maximum-likelihood estimates. The magenta curve is the true reliability, green and blue curves are the Kaplan-Meier estimate and 95% confidence bands respectively, and the RSF reliability and 95% confidence bands based on IJ variance estimate are black and red curves respectively. It can be seen from Fig. 6 that the confidence bands based on IJ variance estimate is close to the the ones given by the maximum likelihood estimate, Greenwood formula. To show how variance estimate varies with different number of trees, variance estimate is computed for a vehicle at time point $t=0.2$ and $t=0.8$ for the various numbers of the trees B. Several options are tried for number of trees B, namely, $B \in \{100, 500, 1000, 2000, 5000, 10000\}$. The result is presented in Fig. 7 showing that the variance estimates, red and blue curves, converge to some non-zero positive. Green and black lines are variances received using Greenwood formula, i.e., computed under an ideal classifier

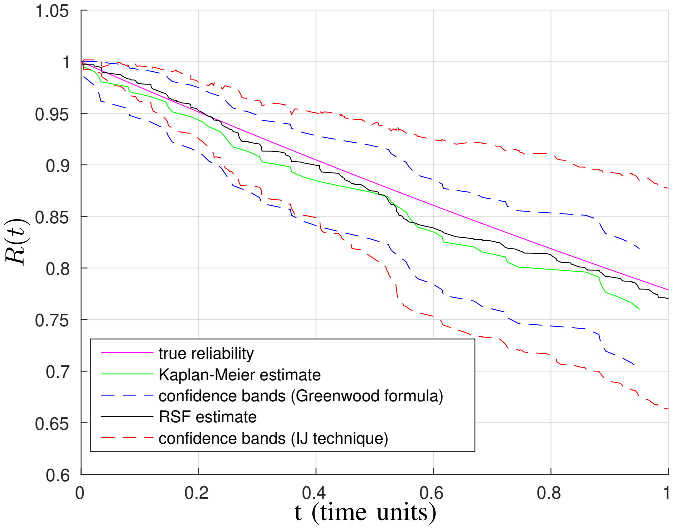


Fig. 6. Reliability with confidence bands. Theoretical values vs estimates from the RSF. Magenta curve is the true reliability curve, green and blue curves are the Kaplan-Meier estimate and 95% confidence bands with Gaussian assumption computed using Greenwood formula, black and red curves are the RSF estimate of the reliability and 95% confidence bands with Gaussian assumption estimated using IJ technique.

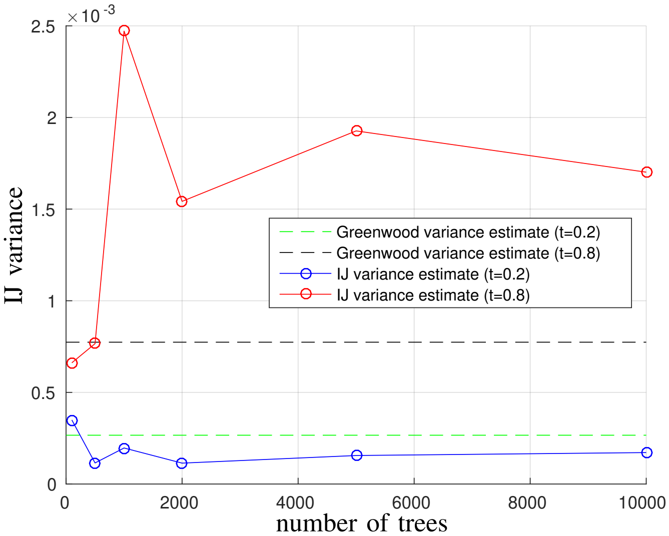


Fig. 7. Theoretical values of variances vs estimates from the RSF. Green and black lines are the true variances, the Greenwood estimates, at time point $t = 0.2$ and $t = 0.8$ respectively. Red and blue curves are the IJ variance estimates at time point $t = 0.2$ and $t = 0.8$ respectively for different number of trees B.

assumption. As it can be seen, the variance estimate at time point $t = 0.2$, blue curve, is very close to the Greenwood estimate. Variance estimate at time point $t = 0.8$, red curve, is biased with respect to the Greenwood estimate which is suspected due to censoring.

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

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

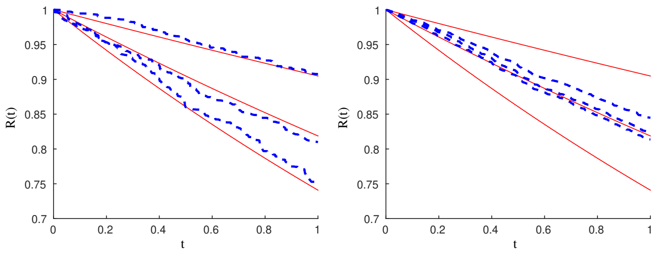
VI. Performance evaluation with several metrics

Every prognostic model should be evaluated such that their predictive performance is known. As mentioned, this is problematic since the output from the forest model is a survival or reliability function and there is no record of their true values in the data set. In a pure classification or regression problem there are established metrics to evaluate performance, however, this is not the case for survival analysis. A metric to use in the case of the RSF framework is an error rate based on the concordance index [7] which estimates the probability that, when a pair of the vehicles/batteries is randomly selected, the vehicle/battery that fails first has a worst predicted outcome. A question is if this error rate is descriptive enough. For example, the authors in [27] conclude that it is possible that the error rate is not an appropriate performance measure, because concordance index measures if the predicted survival times are in the right order and says nothing regarding how close the predicted and actual survival times are. The example given below supports this observation and shows that with similar values of the error rates models predict significantly different survival curves.

The example relies on simulated data similar to the one used in Section V. Degradation of the battery is controlled by the hazard rate h_0 which corresponds to 10 years mean battery life. As in the previous example it is assumed that there is one important variable v_1 which influences hazard rate h_0 such that three classes of vehicles exist with different degradation profiles corresponding to the new hazard rate h

$$h = \begin{cases} 1 \cdot h_0, & \text{if } v_1 = 1 \\ 2 \cdot h_0, & \text{if } v_1 = 2 \\ 3 \cdot h_0, & \text{if } v_1 = 3 \end{cases} \tag{23}$$

Two models with 2 and 100 noisy variables are considered where the censoring rate is about 80% which is similar to the value from the example in Section V. The data set is comprised from 1000 vehicles and parameters of the RSF model are chosen as in Section V-A. Fig. 8 shows the comparison of the predicted survival curves from RSF model, dashed blue curves, with theoretical values, red curves, for three randomly chosen vehicles that were not included in the training sets. It is evident from the left plot in Fig. 8 that, as expected, predictions for the model with only 2 noisy variables are significantly better than for the model with 100 noisy variables, right plot in Fig. 8. At the same time the values of the error rate for both models are close, 0.4097 for the model with 2 noisy and 0.4270 for the model with 100 noisy variables, therefore, one would expect that forest outputs would be similar as well, but this is clearly not the case. Thus, new evaluation techniques are needed to be able to say more about predictive performance of the model.



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

Fig. 8. RSF predictions for two models with different number of noisy variables. Red curves are the theoretical reliabilities for three classes and blue dashed curves are the outputs from the RSF model.

A. Performance analysis of predictive model for battery data

A vehicle used the same way should never leave its class of similar vehicles, however, with year or mileage variables present in the database, the model might be dominated by age effects which is not the intention and could possibly mask the effects of different vehicle usage. The problem of using accumulative variables like age or mileage is addressed by the authors in [28]. It was suggested that instead of using accumulative variables directly it is better to preprocess them first. For example, there are two accumulative variables in the current data set, namely, age and mileage. First, a new variable mileage per day is created and, then, two models are considered, namely, a model based on all variables except the accumulative ones and another model where the variable mileage per day has been added.

The RSF model training and validation processes is described next. Even though 56,163 vehicles are available for the study, 30,000 of them are randomly selected for training and validation purposes. The reason for this is partly limited computational resources for training an RSF model with many variables, which is the case in our study (536 variables per vehicle). In particular, significant memory resources are required. For validation, data is partitioned into training and validation sets where, out of the 30,000 vehicles, 2/3 are assigned to the training set and the remaining 10,000 vehicles to the validation set. Parameters of the RSF model are the same as described in Section V-A.

One way to evaluate performance of a predictor is to look at values of $\hat{R}^{\mathcal{V}}(t_{\text{surv}})$ survival/reliability curves at the time of either failure or censoring, and see how predictions of the two classes of vehicles differ. Fig. 9 and Fig. 10 show the histograms of reliabilities for failed, red color, and censored vehicles, blue color, for the two models with and without the variable mileage per day on the validation set. On the one hand the histograms of the two classes of vehicles are different, however, have a big overlap on the other, therefore not much can be said about the performance of the predictor.

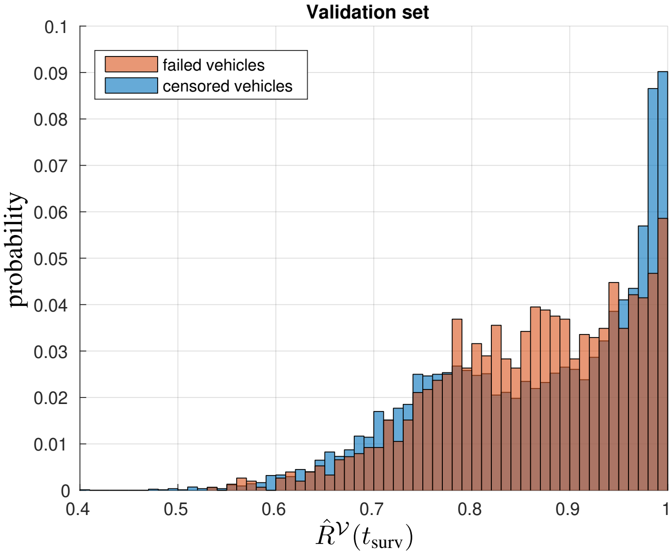


Fig. 9. Histograms of $\hat{R}^{\mathcal{V}}(t_{\text{surv}})$ for the failed vehicles, red bins, and censored vehicles, blue bins. Data set without mileage per day.

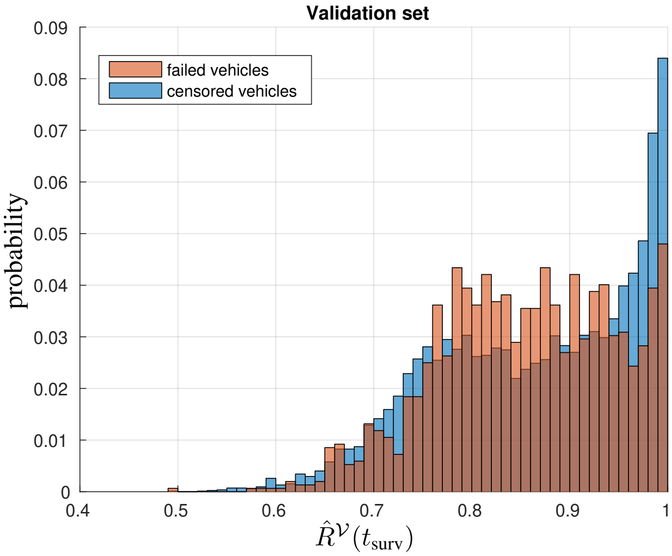


Fig. 10. Histograms of $\hat{R}^{\mathcal{V}}(t_{\text{surv}})$ for the failed vehicles, red bins, and censored vehicles, blue bins. Mileage per day variable is included.

Another approach for performance evaluation is to plot lifetime functions $\hat{\mathcal{B}}^{\mathcal{V}}(t;t_0)$ where $t_0 = t_{\text{surv}}$ and observe how they differ between the two classes of vehicles. The result of the prediction for 100 randomly selected vehicles from failed and censored classes on the validation set is depicted in Fig. 11 and Fig. 12 where red curves correspond to the failed class and blue curves to the censored. What can be seen in the figures is that, on average, predictions for both classes of vehicles are different. However, the overlap between two classes is big,

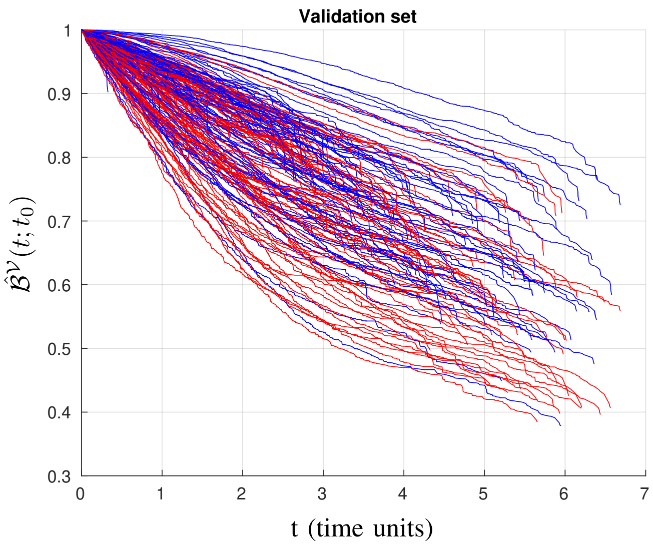


Fig. 11. Lifetime functions $\hat{\mathcal{B}}^{\mathcal{V}}(t;t_0)$ computed for 100 randomly selected vehicles from censored and failed classes on validation set. Red curves correspond to the failed vehicles and blue curves represent the censored vehicles. Data set without mileage per day variable.

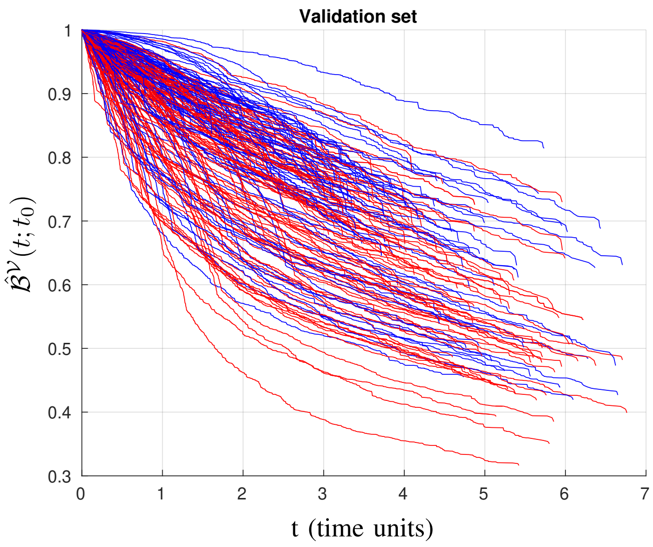


Fig. 12. Lifetime functions $\hat{\mathcal{B}}^{\mathcal{V}}(t;t_0)$ computed for 100 randomly selected vehicles from censored and failed classes on validation set. Red curves correspond to the failed vehicles and blue curves represent the censored vehicles. Mileage per day variable is included.

therefore, it would be good to find more informative measure of performance. Instead of considering the predictions of the reliability and lifetime curves at time t_{surv} when a vehicle is either censored or failed, let us consider the cross section of the respective curves at some fixed time point t which is similar for all vehicles.

Results of reliability histograms computed on the validation set after 3 time units for the two classes of vehicles and two models are shown in Fig. 13 and Fig. 14. This particular time point was selected to allow the batteries to be in operation for some time, so their different usage patterns influence the degradation and it is expected that the predictions for the two classes should differ. The difference between the histograms of the two classes is more clear now than in Fig. 9 and Fig. 10. There is still an overlap between the two histograms and one would expect them to be completely separated in the ideal case, however, it is possible that some of the censored vehicles are really close to failure, but leave the study before failure and the problem of the battery is not recorded. Therefore, left tails of the censored histograms with small values of reliabilities could be not a mistake of the algorithm, but a correct indication that a vehicle belongs to the failed class. On the other hand, a group of vehicles from the failed class which has reliability values close to 1, right tails of the failed histograms, experiences problem with battery due to the reasons that cannot be explained by information in the current data set. Thus, it is impossible for the algorithm to see that the vehicle has potential problems with a battery. One other thing to notice is that for the model that includes mileage per day variable there is a peak in the failed histogram coinciding with the peak of censored histogram, see Fig. 14. For now, it is unclear what it represents, however the results are affected by including or excluding the variable.

Histograms of the lifetime functions $\hat{\mathcal{B}}^{\mathcal{V}}(t;t_0)$ computed on the validation set for two models at time point $t = t_{\text{surv}}+1$ time unit and $t_0 = t_{\text{surv}}$ are presented in Fig. 15 and Fig. 16. Our industrial partner Scania CV is, say, interested in predictions up to 1 time unit to be used in their maintenance planner, therefore, only a time point within 1 time unit in the future is selected. Separation between histograms for failed and censored classes is not so distinctive as in the case of the reliability curves, nevertheless, similar behavior is seen for the model with the mileage per day variable where the peak of the histogram of the failed batteries coincides with the peak of the censored one, see Fig. 16. In addition, the histogram of the failed batteries for the model with mileage per day are skewed more to the right than for the model without the variable.

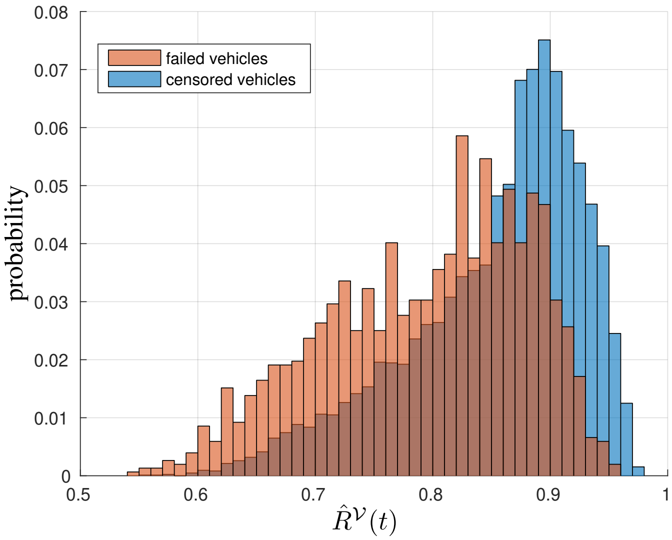


Fig. 13. Histograms of $\hat{R}^{\mathcal{V}}(t)$ for the failed vehicles, red bins, and censored vehicles, blue bins, at time point $t = 3$ time units. Data set without mileage per day variable.

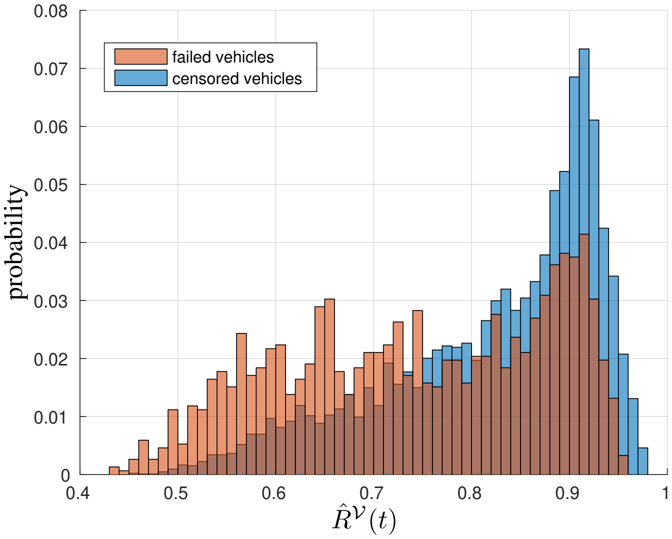


Fig. 14. Histograms of $\hat{R}^{\mathcal{V}}(t)$ for the failed vehicles, red bins, and censored vehicles, blue bins, at time point $t = 3$ time units. Mileage per day variable is included.

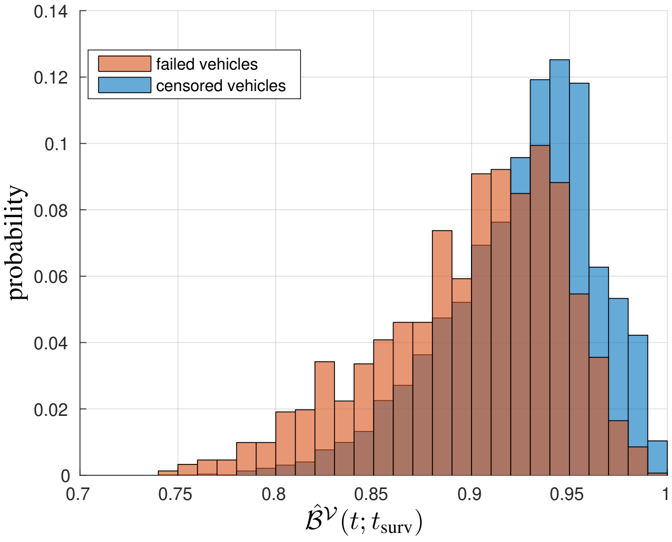


Fig. 15. Histograms of lifetime function estimates $\hat{\mathcal{B}}^{\mathcal{V}}(t; t_{\text{surv}})$ for the failed vehicles, red bins, and censored vehicles, blue bins, at $t = 1$ time unit point in future from t_{surv}. Data set without mileage per day variable.

Different evaluation methods for the prognostic performance of the RSF model are demonstrated in this section showing the difficulty in validating the results of the model's predictions, especially for the case when only one snapshot of data per vehicle is available. It is also shown in the section that predictions of the model that includes the mileage per day variable are different from the predictions of the model that does not include the variable. A detailed investigation of the differences between the two aforementioned models is given next.

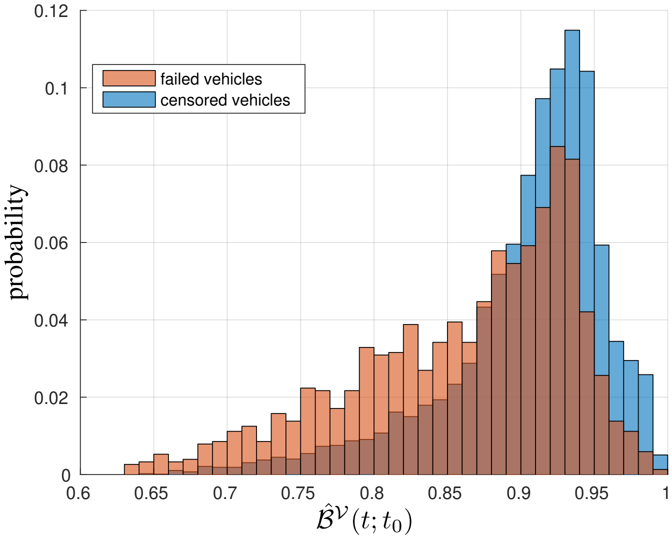


Fig. 16. Histograms of lifetime function estimates $\hat{\mathcal{B}}^{\mathcal{V}}(t; t_{\text{surv}})$ for the failed vehicles, red bins, and censored vehicles, blue bins, at $t = 1$ time unit point in future from t_{surv}. Mileage per day variable is included.

TABLE I
VALUES OF VARIABLES SELECTED AMONG 50 MOST IMPORTANT GIVEN BY VIMP FOR VEHICLE V_1 WITH BEST PROGNOSIS, V_2 WITH THE WORST PROGNOSIS FOR THE MODEL EXCLUDING MILEAGE PER DAY AND VEHICLE V_3 WITH THE WORST PROGNOSIS FOR THE MODEL INCLUDING MILEAGE PER DAY

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

B. Lifetime prognosis for vehicles with similar mileage

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

From this set of vehicles with similar mileage, vehicles with similar age is selected. There are 84 vehicles satisfying the stated requirement on mileage in the validation set. The lifetime function estimates $\hat{\mathcal{B}}^{\mathcal{V}}(t;t_0)$ with $t_0 = t_{\text{surv}}$ for the selected vehicles are presented in Fig. 17 and Fig. 18 showing the prediction for model with and without mileage per day variable respectively. First notice that the difference between the best and the worst predictions is significant which shows that how vehicles are used is important. Next, three vehicles V_1, V_2 and V_3 are selected from the set of the vehicles with similar mileage. Vehicle V_1 corresponds to the lifetime function with the best prognosis and is the same vehicle for the both models, when vehicles V_2 and V_3 with the worst prognosis are different for the two models. Age of batteries for the vehicles V_1, V_2 and V_3 are 1.3, 0.83 and 0.98 time units respectively where the vehicle V_1 with the best prognosis lived the longest among three vehicles, therefore, vehicle usage pattern plays a significant role.

Table I shows selected variables for three vehicles V_1, V_2 and V_3 among 50 most important variables for the prediction obtained using VIMP [7], Variable IMPortance, with the most important variables at the top. VIMP can be interpreted in terms of misclassification under the concordance index. As mentioned above, the index estimates the probability of correctly classifying two vehicles. Therefore, VIMP measures the increase or decrease in the concordance index on the test data if the given variable is not available for training the model. Only variables that have different values for three vehicles are left among 50 most important. First, vehicles operated in different countries that can explain the difference in the degradation profiles of the batteries as climate, quality of roads can vary. Bin 3 of the ambient temperature histogram appears important for the prediction. This bin corresponds to operation of a vehicle under the low temperatures. Vehicles V_2 and V_3 have operated more time under the low temperatures which corroborates the fact that the vehicles have worse degradation prediction than the vehicle V_1. Two bins of the atmospheric pressure histogram are important, namely, bins 7 and 8. Vehicle V_1 has much bigger value in the 7th bin compared to the values for the vehicles V_2 and V_3, at the same time has much lower value in the 8th bin. Two bins from the battery voltage histogram are also important. They correspond to the operation of the battery under high voltage. It can be seen that the vehicles with worse prediction operated more under the high voltage that can be considered as counterintuitive at first. However, it is possible that the generator that charges the battery has malfunctions that lead to overcharging and faster degradation of the battery. Overall, there is a significant amount of the variables in Table I that indicate different usage of the vehicles. This fact gives positive signs for using the RSF method as a predictive tool.

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

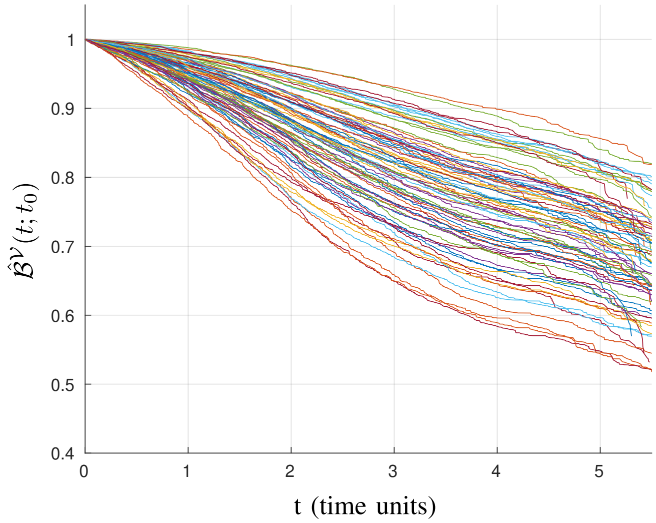


Fig. 17. Lifetime functions estimates $\hat{\mathcal{B}}^{\mathcal{V}}(t;t_0)$ for 84 vehicles which mileage values are in plus-minus 5% interval around base mileage value m and age of batteries are within 1 to 2 time unit interval. Model does not contain mileage per day variable.

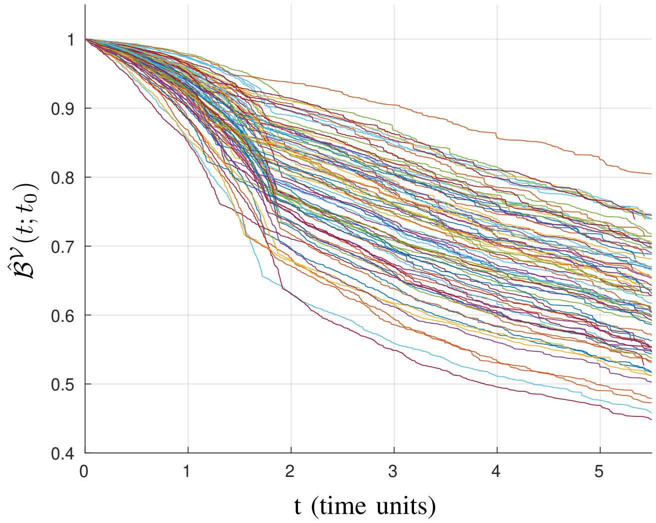


Fig. 18. Lifetime functions estimates $\hat{\mathcal{B}}^{\mathcal{V}}(t;t_0)$ for 84 vehicles which mileage values are in plus-minus 5% interval around base mileage value m and age of batteries are within 1 to 2 time unit interval. Model contains mileage per day variable.

Now, consider the lifetime function estimates which correspond to the best, worst and intermediate prediction for two models. They can be found in Fig. 19 and Fig. 20. The lifetime function estimates correspond to the solid lines in the figures, and dashed curves are 95% confidence bands with Gaussian assumption and IJ variance estimate from Section IV. It can be seen that confidence bands for the model with mileage per day variable are wider than for the model without which is a surprising result. Intuitively the more variables the better predictions, however, the result shows opposite. It means that relying on usage profile rather than on time related variables

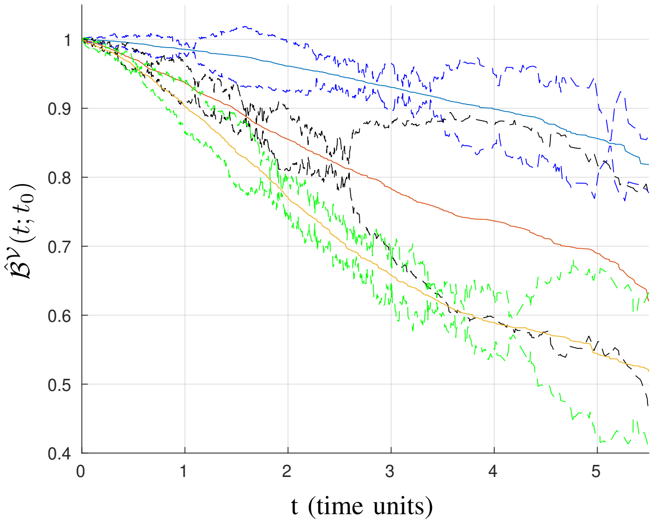


Fig. 19. Lifetime function estimates $\hat{\mathcal{B}}^{\mathcal{V}}(t;t_0)$ for the best, worst and intermediate predictions from Fig. 17, solid lines, together with 95% confidence bands with Gaussian assumption and IJ variance estimate, dashed curves. Without mileage per day model.

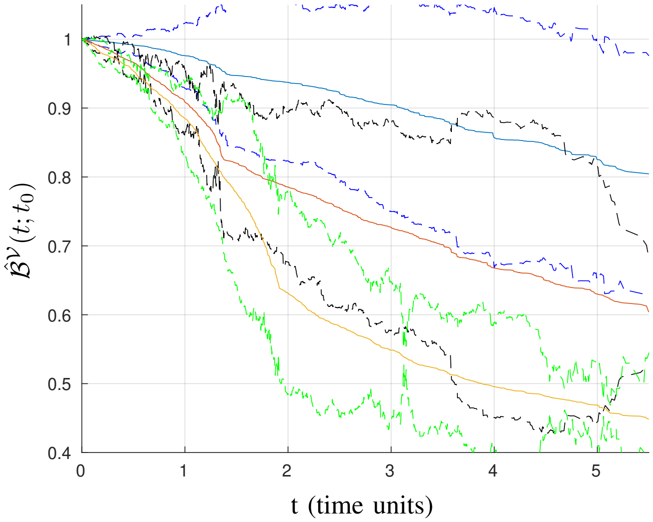


Fig. 20. Lifetime function estimates $\hat{\mathcal{B}}^{\mathcal{V}}(t;t_0)$ for the best, worst and intermediate predictions from Fig. 18, solid lines, together with 95% confidence bands with Gaussian assumption and IJ variance estimate, dashed curves. With mileage per day model.

would give more accurate predictor for the given data. More studies should be carried out to see if incorporation of time related variables can give better performance.

As a conclusion, RSF model applied to the given data set gives on average different predictions for the failed and the censored class of vehicles, results show that vehicle usage profile is important for predicting the degradation of a battery and that there is an indication not to include accumulative variables into the training RSF model as it increases uncertainty of the predictor. It is impossible to determine and validate a failure time for a battery with the given data set, because only one snapshot of data is available for every vehicle. When such information becomes available to us, the methods could and should be extended and further validated.

VII. Conclusion

It is shown in the paper that an RSF model can be applied to the static data, i.e., one snapshot only per vehicle, in the data set to predict a battery lifetime prediction function. A key difference in the data compared to many other prognostic approaches, e.g., [2, 5, 8], is that only one snapshot per vehicle is available and it is not possible to track the vehicle to predict failure time. The lifetime function (1) is proposed as an estimate of the battery lifetime and the RSF model output is the estimate of the reliability function which can be used to compute the lifetime function estimate (2). The confidence bands of the lifetime function estimate (2) are estimated by extending the existing Infinitesimal Jackknife (IJ) variance estimate approach for Random Forest method to Random Survival Forest and properties of the variance estimates are analyzed. First, confidence bands can be used for the model selection, for example, it is shown that the prediction for the forest model with 1,000 trees is significantly better than for the model with 100 trees in terms of confidence bands, however, in terms of the standard error rate, the two models are similar. Second, IJ variance estimate starts to converge for the forest with 1,000 trees or larger which means that the variance estimate of the predictor with 1000 trees is appropriate. Models with and without accumulative variables give different results and currently it seems that excluding accumulative variables gives better results based on the fact that the confidence bands become narrower. Performance evaluation is done and it has been shown that prediction for a censored and failed vehicle is different. It is also shown that the validation of the method's prediction performance in the case when only one snapshot of data per vehicle is available is difficult and requires extensive analysis and problem insight. In general, the battery lifetime function can be used to schedule and optimize the cost of the battery replacement which leads to more flexible maintenance.

Appendix
Proof of Lemma 1

This section gives the proof of Lemma 1. According to the definition of the covariance

$$\widehat{\text{cov}}_{\text{Bias}}\left[\hat{R}^{\mathcal{V}}(t+t_0),\hat{R}^{\mathcal{V}}(t_0)\right]= \\ =E\left[\left(\hat{R}^{\mathcal{V}}(t+t_0)-E\left[\hat{R}^{\mathcal{V}}(t+t_0)\right]\right)\cdot \\ \cdot\left(\hat{R}^{\mathcal{V}}(t_0)-E\left[\hat{R}^{\mathcal{V}}(t_0)\right]\right)\right] \quad (24)$$

where $E[x]$ is an expectation of a random variable x.

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

$$\hat{\theta}_{\text{RSF}}(\boldsymbol{P},t)=\hat{\theta}_{\text{RSF}}(\boldsymbol{P}^0)+(\boldsymbol{P}-\boldsymbol{P}^0)\cdot\boldsymbol{U}+ \\ +\mathcal{O}((\boldsymbol{P}-\boldsymbol{P}^0)\cdot(\boldsymbol{P}-\boldsymbol{P}^0)') \quad (25)$$

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

$$U_i(t) = \lim_{\epsilon \to 0} \frac{\hat{\theta}_{\text{RSF}}(\boldsymbol{P}^0 + \epsilon(\boldsymbol{\delta}_i - \boldsymbol{P}^0), t) - \hat{\theta}_{\text{RSF}}(\boldsymbol{P}^0, t)}{\epsilon}, \quad i = 1, \dots, n. \tag{26}$$

Taking the result in (25) into account, covariance of reliabilities in (24) becomes

$$\begin{aligned}
&\widehat{\text{cov}}_{\text{Bias}}\left[\hat{R}^{\mathcal{V}}(t+t_0), \hat{R}^{\mathcal{V}}(t_0)\right] = \\
&= E\left[\left(\hat{\theta}_{\text{RSF}}(\boldsymbol{P}, t+t_0) - E\left[\hat{\theta}_{\text{RSF}}(\boldsymbol{P}, t+t_0)\right]\right) \cdot\right. \\
&\quad \left.\cdot \left(\hat{\theta}_{\text{RSF}}(\boldsymbol{P}, t_0) - E\left[\hat{\theta}_{\text{RSF}}(\boldsymbol{P}, t_0)\right]\right)\right] = \\
&= E\left[\left((\boldsymbol{P} - \boldsymbol{P}^0)\boldsymbol{U}(t+t_0)\right)\left((\boldsymbol{P} - \boldsymbol{P}^0)\boldsymbol{U}(t_0)\right)\right].
\end{aligned} \tag{27}$$

A resampling vector for each tree has a rescaled multinominal distribution

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

with mean and covariance matrices

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

Covariance expression with the directional derivatives becomes

$$\begin{aligned}
&\widehat{\text{cov}}\left[\hat{R}^{\mathcal{V}}(t+t_0), \hat{R}^{\mathcal{V}}(t_0)\right] = \\
&= E\left[\left(\sum_{i=1}^{n}(p_i - \frac{1}{n})U_i(t+t_0)\right)\left(\sum_{j=1}^{n}(p_j - \frac{1}{n})U_j(t_0)\right)\right] = \\
&= E\left[\sum_{i=1}^{n}\left(p_i - \frac{1}{n}\right)^2 U_i(t_0)U_i(t+t_0) + \right. \\
&\quad + \sum_{i \neq j}\left(p_i - \frac{1}{n}\right)\left(p_j - \frac{1}{n}\right)U_i(t_0)U_j(t+t_0) + \\
&\quad \left. + \sum_{i \neq j}\left(p_i - \frac{1}{n}\right)\left(p_j - \frac{1}{n}\right)U_i(t+t_0)U_j(t_0)\right] = \\
&= \sum_{i=1}^{n}\frac{1}{n^2}\left(1 - \frac{1}{n}\right)U_i(t_0)U_i(t+t_0) + \\
&\quad + \sum_{i \neq j}\left(-\frac{1}{n^3}\right)U_i(t_0)U_j(t+t_0) + \\
&\quad + \sum_{i \neq j}\left(-\frac{1}{n^3}\right)U_i(t+t_0)U_j(t_0) = \\
&= \frac{1}{n^2}\sum_{i=1}^{n}U_i(t_0)U_i(t+t_0) - \\
&\quad - \frac{1}{n^3}\left[\left(\sum_{i=1}^{n}U_i(t_0)\right)\left(\sum_{j=1}^{n}U_j(t+t_0)\right)\right].
\end{aligned} \tag{28}$$

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

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

Therefore, according to the definition of directional derivative $U_i(t)$ in (12) can be expressed as

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

where $\boldsymbol{\delta}_i$ has 1 at the i^{th} position and 0 at all others. Rewriting $U_i(t)$ using knowledge about the vectors' structure gives

$$\begin{aligned}
U_i(t) &= \left(\underbrace{-\frac{1}{n}, \dots, -\frac{1}{n}}_{\text{i - 1}}, 1 - \frac{1}{n}, -\frac{1}{n}, \dots, -\frac{1}{n}\right) \cdot \begin{pmatrix} D_1 \\ \vdots \\ D_n \end{pmatrix} = \\
&= \sum_{j \neq i}\left(-\frac{1}{n}\right) \cdot \frac{\partial}{\partial p_j}\hat{\theta}_{\text{RSF}}(\boldsymbol{P}, t)\Big|_{\boldsymbol{P}=\boldsymbol{P}^0} + \\
&\quad + \left(1 - \frac{1}{n}\right) \cdot \frac{\partial}{\partial p_i}\hat{\theta}_{\text{RSF}}(\boldsymbol{P}, t)\Big|_{\boldsymbol{P}=\boldsymbol{P}^0}.
\end{aligned}$$

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

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

Thus, by substituting zeroes instead of the sums of directional derivatives in (28) we get

$$\widehat{\text{cov}}_{\text{Bias}}\left[\hat{R}^{\mathcal{V}}(t+t_0), \hat{R}^{\mathcal{V}}(t_0)\right] = \frac{1}{n^2}\sum_{i=1}^{n}U_i(t_0)U_i(t+t_0). \tag{29}$$

Following the same steps as in [24] it can be written that

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

which proves (20). Bias from (21) of $\widehat{\text{cov}}_{\text{Bias}}\left[\hat{R}^{\mathcal{V}}(t+t_0), \hat{R}^{\mathcal{V}}(t_0)\right]$ estimate is found as follows

$$\begin{aligned}
\text{Bias} &= E\left[\widehat{\text{cov}}_{\text{Bias}}\left[\hat{R}^{\mathcal{V}}(t+t_0), \hat{R}^{\mathcal{V}}(t_0)\right]\right] - \\
&\quad - \text{cov}\left[\hat{R}^{\mathcal{V}}(t+t_0), \hat{R}^{\mathcal{V}}(t_0)\right].
\end{aligned} \tag{30}$$

Here, $\text{cov}\left[\hat{R}^{\mathcal{V}}(t+t_0), \hat{R}^{\mathcal{V}}(t_0)\right]$ is a covariance between reliabilities when number of trees in the forest $B \to \infty$. One can rewrite (30) as

$$\begin{aligned}
\text{Bias} &= \sum_{j=1}^{n}\left(E\left[\widehat{\text{Cov}}_i(t_0)\widehat{\text{Cov}}_i(t+t_0)\right] - \right. \\
&\left. - \text{Cov}_i(t_0)\text{Cov}_i(t+t_0)\right) = \sum_{j=1}^{n}\left(E\left[\widehat{\text{Cov}}_i(t_0)\widehat{\text{Cov}}_i(t+t_0)\right] - \right. \\
&\quad \left. - E\left[\widehat{\text{Cov}}_i(t_0)\right]E\left[\widehat{\text{Cov}}_i(t+t_0)\right]\right) = \\
&= \sum_{j=1}^{n}\text{cov}\left[\widehat{\text{Cov}}_i(t_0); \widehat{\text{Cov}}_i(t+t_0)\right].
\end{aligned} \tag{31}$$

Taking into account expression for $\widehat{\text{Cov}}_i(t_0)$ in (16) bias becomes

$$\begin{aligned}\text{Bias} = \sum_{j=1}^{n} \text{cov}\Bigg[\frac{1}{B}\sum_{b=1}^{B}(y^*_{ij}-1)(\hat{R}^{\mathcal{V}}_b(t_0)-\hat{R}^{\mathcal{V}}(t_0));\\ \frac{1}{B}\sum_{b=1}^{B}(y^*_{bj}-1)(\hat{R}^{\mathcal{V}}_b(t+t_0)-\hat{R}^{\mathcal{V}}(t+t_0))\Bigg] =\\ = \frac{1}{B^2}\sum_{j=1}^{n}\sum_{i=1}^{B}\sum_{b=1}^{B}\Big(E\Big[(y^*_{bj}-1)(y^*_{ij}-1)\cdot\\ (\hat{R}^{\mathcal{V}}_i(t_0)-\hat{R}^{\mathcal{V}}(t_0))(\hat{R}^{\mathcal{V}}_b(t+t_0)-\hat{R}^{\mathcal{V}}(t+t_0))\Big] -\\ - E\Big[(y^*_{ij}-1)(\hat{R}^{\mathcal{V}}_i(t_0)-\hat{R}^{\mathcal{V}}(t_0))\Big]\cdot\\ E\Big[(y^*_{bj}-1)(\hat{R}^{\mathcal{V}}_b(t+t_0)-\hat{R}^{\mathcal{V}}(t+t_0))\Big]\Big). \quad (32)\end{aligned}$$

Assuming that the original sample $\boldsymbol{Y}$ is large enough, meaning $n \to \infty$, it becomes possible to suppose that $\hat{R}^{\mathcal{V}}_i(t)$ and y^*_{ij} are independent and as the result bias simplifies to

$$\begin{aligned}\text{Bias} = \frac{1}{B^2}\sum_{j=1}^{n}\sum_{i=1}^{B}\sum_{b=1}^{B}\big(E\left[(y^*_{bj}-1)(y^*_{ij}-1)\right]\cdot\\ E\Big[(\hat{R}^{\mathcal{V}}_i(t_0)-\hat{R}^{\mathcal{V}}(t_0))(\hat{R}^{\mathcal{V}}_b(t+t_0)-\hat{R}^{\mathcal{V}}(t+t_0))\Big] -\\ - E\left[(y^*_{ij}-1)\right]E\Big[(\hat{R}^{\mathcal{V}}_i(t_0)-\hat{R}^{\mathcal{V}}(t_0))\Big]\cdot\\ E\left[(y^*_{bj}-1)\right]E\Big[(\hat{R}^{\mathcal{V}}_b(t+t_0)-\hat{R}^{\mathcal{V}}(t+t_0))\Big]\Big). \quad (33)\end{aligned}$$

Random variable y^*_{ij} has the following properties

$$\begin{aligned}E\left[(y^*_{ij}-1)(y^*_{bj}-1)\right] = \text{cov}\left[y^*_{ij};y^*_{bj}\right] = \frac{1}{n} \to 0,\\ n\to\infty, b\neq j\\ E\left[y^*_{ij}-1\right] = E\left[y^*_{bj}-1\right] = 0\\ E\left[(y^*_{ij}-1)^2\right] = \text{var}\left[y^*_{ij}\right] = 1-\frac{1}{n} \to 1,\\ n\to\infty, b=j.\end{aligned}$$

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

$$\begin{aligned}\text{Bias} = \frac{1}{B^2}\sum_{j=1}^{n}\sum_{i=1}^{B}\big(E\left[(y^*_{ij}-1)^2\right]\cdot\\ E\Big[(\hat{R}^{\mathcal{V}}_i(t_0)-\hat{R}^{\mathcal{V}}(t_0))(\hat{R}^{\mathcal{V}}_i(t+t_0)-\hat{R}^{\mathcal{V}}(t+t_0))\Big]\Big) =\\ = \frac{n}{B^2}\sum_{i=1}^{B}(\hat{R}^{\mathcal{V}}_i(t_0)-\hat{R}^{\mathcal{V}}(t_0))(\hat{R}^{\mathcal{V}}_i(t+t_0)-\hat{R}^{\mathcal{V}}(t+t_0)).\\ (34)\end{aligned}$$

The expression in (34) is similar to the bias correction for the RF model found in [24] and presented in (14).

Acknowledgment

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

References

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

[2] M. Daigle and K. Goebel, "A model-based prognostics approach applied to pneumatic valves," *International Journal of Prognostics and Health Management*, vol. 2, no. 2, pp. 1–16, 2011.

[3] H. Hanachi, J. Liu, A. Banerjee, Y. Chen, and A. Koul, "A physics-based modeling approach for performance monitoring in gas turbine engines," *IEEE Transactions on Reliability*, vol. 64, no. 1, 2015.

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

[5] K. Medjaher, D. A. Tobon-Mejia, and N. Zerhouni, "Remaining useful life estimation of critical components with application to bearings," *IEEE Transactions on Reliability*, vol. 61, no. 2, 2012.

[6] D. Cox, "Regression model and life-table," *Journal of the Royal Statistical Society*, vol. 34, no. 2, pp. 187–220, 1972.

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

[8] F. Zhao, Z. Tian, E. Bechhoefer, and Y. Zeng, "An integrated prognostics method under time-varying operating conditionds," *IEEE Transactions on Reliability*, vol. 64, no. 2, 2015.

[9] L. Liao and F. Kottig, "Review of hybrid prognostics approaches for remaining useful life prediction of engineered systems, and an application to battery life prediction," *IEEE Transactions on Reliability*, vol. 63, no. 1, 2014.

[10] S. M. Rezvanizaniani, Z. Liu, Y. Chen, and J. Lee, "Review and recent advances in battery health monitoring and prognostics technologies for electric vehicle (ev) safety and mobility," *Journal of Power Sources*, vol. 256, pp. 110 – 124, 2014.

[11] J. Zhang and J. Lee, "A review on prognostics and health monitoring of li-ion battery," *Journal of Power Sources*, vol. 196, no. 15, pp. 6007 – 6014, 2011.

[12] H. Blanke, O. Bohlen, S. Buller, R. W. D. Doncker, B. Fricke, A. Hammouche, D. Linzen, M. Thele, and D. U. Sauer, "Impedance measurements on lead–acid batteries for state-of-charge, state-of-health and cranking capability prognosis in electric and hybrid electric vehicles," *Journal of Power Sources*, vol. 144, pp. 418 – 425, 2005.

[13] G. Capizzi, F. Bonanno, and G. M. Tina, "Recurrent neural network-based modeling and simulation of lead-acid batteries charge-discharge," *IEEE Transactions on Energy Conversion*, vol. 26, no. 2, pp. 435–443, June 2011.

[14] D. R. Cox and D. Oakes, *Analysis of survival data.* CRC

Press, 1984, vol. 21.

[15] E. L. Kaplan and P. Meier, “Nonparametric estimation from incomplete observations,” *Journal of the American Statistical Association*, vol. 53, no. 282, pp. 457–481, 1958.

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

[17] L. Breiman, J. Friedman, R. Olshen, and S. C., *Classification and regression trees*. Taylor and Francis, 1984.

[18] L. Breiman, “Random forests,” *Machine learning*, vol. 45, no. 1, pp. 5–32, 2001.

[19] T. Dietterich, “An experimental comparison of three methods for constructing ensembles of decision trees: bagging, boosting, and randomization,” *Machine Learning*, vol. 40, no. 2, pp. 139–157, 2000.

[20] T. Ho, “The random subspace method for constructing decision forests,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 20, no. 8, pp. 832–844, 1998.

[21] B. Efron, “Bootstrap methods: Another look at the jackknife,” *Ann. Statist.*, vol. 7, no. 1, pp. 1–26, 01 1979.

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

[23] B. Efron, “Estimation and accuracy after model selection,” *Journal of the American Statistical Assosiation*, vol. 109, pp. 991–1007, 2014.

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

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

[26] H. Ishwaran and U. Kogalur, “Random survival forests for r,” *Rnews*, vol. 7/2, pp. 25–31, 2007.

[27] H. Moradian, D. Larocque, and F. Bellavance, “L1 splitting rules in survival forests,” *Lifetime Data Analysis*, vol. 21, no. 1, 2016.

[28] E. Frisk and M. Krysander, “Treatment of accumulative variables in data-driven prognostics of lead-acid batteries,” in *Proceedings of IFAC Safeprocess’15*, Paris, France, 2015.