Heavy-duty truck battery failure prognostics using random survival forests

Sergii Voronov, Daniel Jung, and Erik Frisk

Department of Electrical Engineering, Linköping University, Sweden
{*sergii.voronov, daniel.jung, erik.frisk*}*@liu.se.*

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

Keywords: Battery failure prognosis, Random survival forests, Variable selection

1. INTRODUCTION

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

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

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

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

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

2. PROBLEM MOTIVATION

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

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

where T is the random variable failure time of the battery and ν the vehicle data at $t = t_0$. The function $\mathcal{B}^{\nu}(t;t_0)$ is a function of t and gives the probability that the battery will function at least t time units after t_0. The data ν is recorded operational data for a specific vehicle which is further described in Section 2.1.

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

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

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

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

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

2.1 Operational data

In this work a vehicle fleet database is provided, where one snapshot of data is available from each vehicle including information regarding how the truck has been used and the configuration of the specific truck. There is also information if the battery has failed or not. The database contains a lot of information from the truck, not always related to battery

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

degradation, meaning that it is not known what available information is relevant for this specific task. Therefore, it is relevant to identify which variables are relevant for predicting battery lifetime. Previous works considering this vehicle data set are presented in (Frisk et al., 2014) and (Frisk and Krysander, 2015).

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

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

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

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

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

2.2 Variable selection using Random Survival Forests

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

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

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

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

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

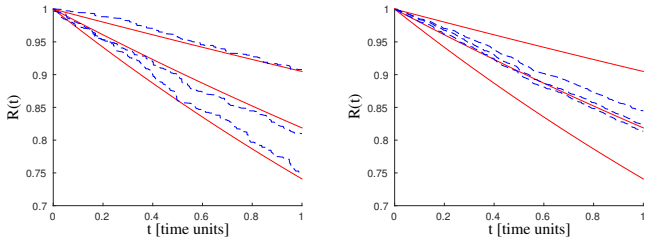
$$h = \begin{cases} 1 \cdot h_0, & \text{if } v_1 = 1 \\ 2 \cdot h_0, & \text{if } v_1 = 2 \\ 3 \cdot h_0, & \text{if } v_1 = 3. \end{cases} \tag{4}$$

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

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

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

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



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

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

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

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

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

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

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

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

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

3. RANDOM SURVIVAL FORESTS

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

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

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

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

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

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

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

3.1 Prediction error

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

3.2 Measures of variable importance and RSF

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

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

As an alternative to VIMP, a candidate measure called minimal depth for variable selection in RSF has been proposed (Ishwaran

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

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

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

4. VIMP AND MINIMAL DEPTH EVALUATION

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

The minimal depth approach is applied to the vehicle database using the recommended configuration described in (Ishwaran et al., 2011). The result of the minimal depth approach is shown in Fig. 4. The x-axis is the mean minimal depth and the y-axis show the mean value of the second minimal depth. Second minimal depth is the distance to the root from a node, in another branch of the tree, where the variable appears the second time. Important variables are thus expected to appear in the lower left corner and non-important in the upper right. The computed threshold, based on (7), is shown as the red vertical line. Most variables are located below the threshold, including the known

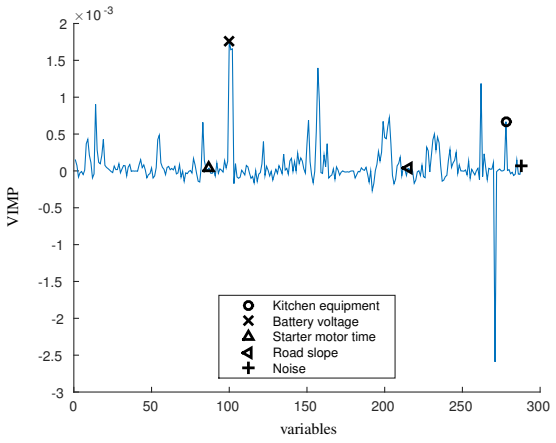


Fig. 3. VIMP values for all variables from vehicle database where x axis corresponds to variables from vehicle database, y axis shows VIMP value for a particular variable. Large positive values of VIMP corresponds to important variables.

noisy variable, meaning that the variable selection is not able to distinguish the important variables from noisy variables.

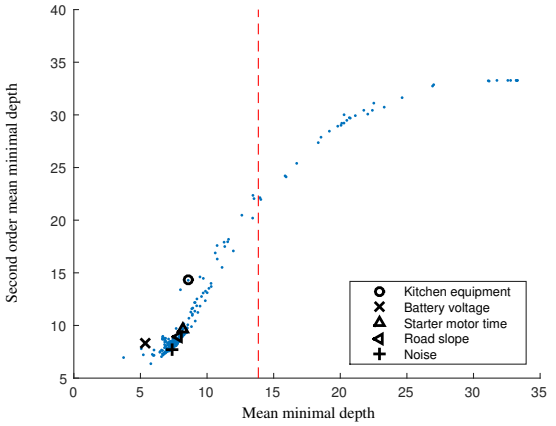


Fig. 4. Minimal depth approach applied to vehicle database where x axis corresponds to mean value of the first appearance of a variable in a tree, y axis corresponds to mean value of the second appearance of a variable in a tree. Dashed red line is a threshold that separates important and non-important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

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

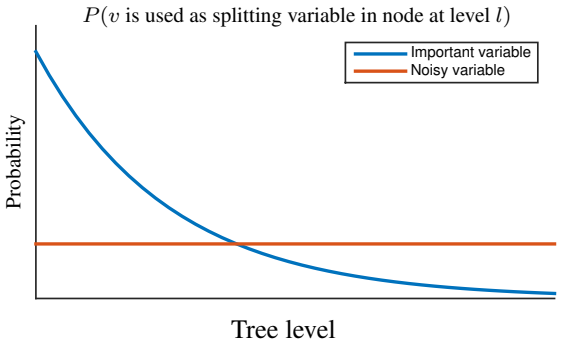


Fig. 5. Illustrative example of the probability that a given splitting variable is used in a node at different tree levels.

5. MEASURE FOR VARIABLE SELECTION

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

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

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

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

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

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

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

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

which will be used when analyzing the RSF.

Generating an RSF for identifying important variables differs from generating an RSF for battery life prediction. To identify important variables it is useful to generate the RSF such that the chance of having significant variations between variables is increased. Thus, each tree in the forest is allowed to grow deep to have as many levels and branches as possible. Therefore, the minimal terminal node size was chosen to be two. This parameter choice is not suitable for battery life prediction where instead a minimal terminal node size of 200 was used. In the later case, the focus is in quality of prediction and taking into account the fact that there are no time series for each vehicle, it should be associated with a class of vehicles with similar usage profile. Therefore, small values of minimal node size could be a bad choice. However for variable selection, trees are required to be as deep as possible, because quality of (9) depends on it. Based on experience, but also to compare the results with the minimal depth approach, 1000 trees was selected to be generated in the RSF.

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

6. IDENTIFYING IMPORTANT VARIABLES FOR BATTERY FAILURE PROGNOSTICS

The proposed variable importance measure estimate (9) is here used to analyze data from the vehicle database. There is a number of different histogram variables describing how the vehicle is used where each bin represents how much a variable has been measured within a specific interval. As a first step, the analysis will evaluate if it is possible to identify important operating regions and vehicle configurations which are correlated with battery degradation. The results are

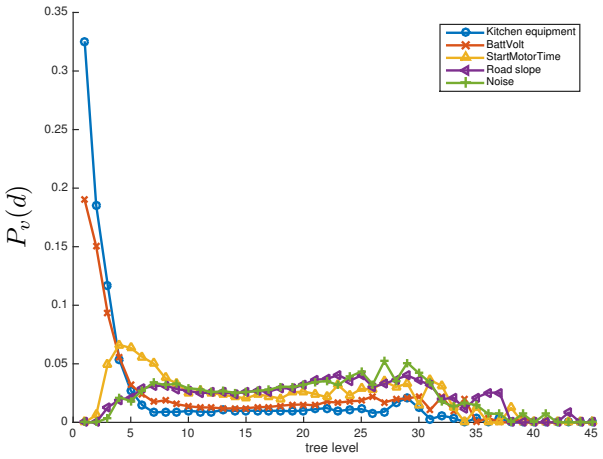


Fig. 6. Probability mass functions $P_v(d)$ for 5 variables from vehicle database calculated according to (9).

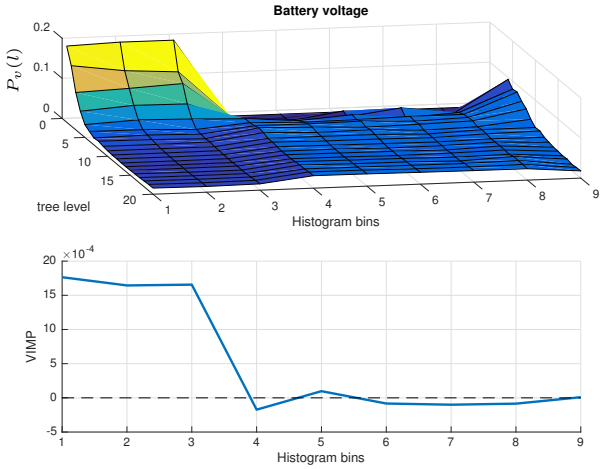


Fig. 7. Variable importance analysis of battery voltage histogram variables.

discussed based on expert knowledge for variables related to battery voltage, fuel consumption, starter motor usage, ambient temperature, and configuration variables. Then, in a second step an automatic procedure is outlined and applied to the full set of variables.

Based on the observations in Fig. 6, the shape of $P_v(d)$ for high d is more noisy due to varying sizes of the different trees. Thus, for better visualization (9) is plotted for each histogram variable but only for tree levels $d \leq 20$. Fig. 7 upper plot shows $P_v(d)$ for different histogram bins of battery voltage variable when the battery is used, where bin 1 represents low battery voltage and bin 9 high battery voltage. The three lowest histogram bins have higher values at lower tree levels indicating that the time the battery in the truck is having low voltage is important for battery health prediction. It is also visible that the bin 9 significantly higher at lower tree levels, compared to the bins 4-8, meaning that high voltages are also relevant for battery health prognostics. When comparing the result to VIMP, Fig. 7 lower plot, it is visible that both methods identifies low voltages as important, high positive values of VIMP mean important variables, but the high voltage is not identified by VIMP.

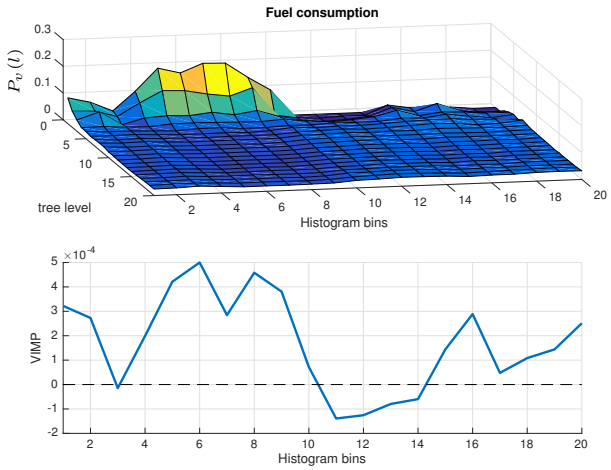


Fig. 8. Variable importance analysis of fuel consumption speed histogram variables.

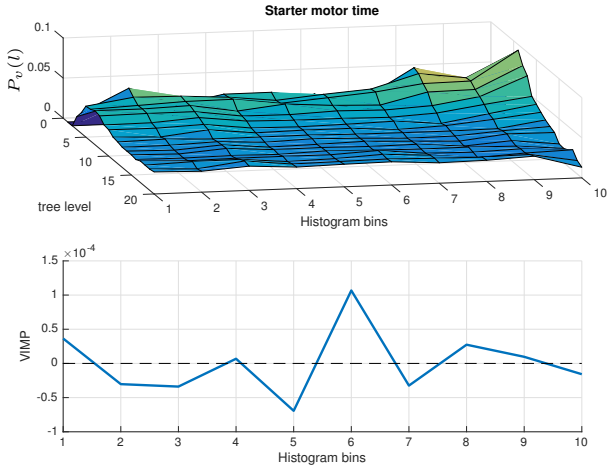


Fig. 9. Variable importance analysis of starter motor time histogram variables.

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

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

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

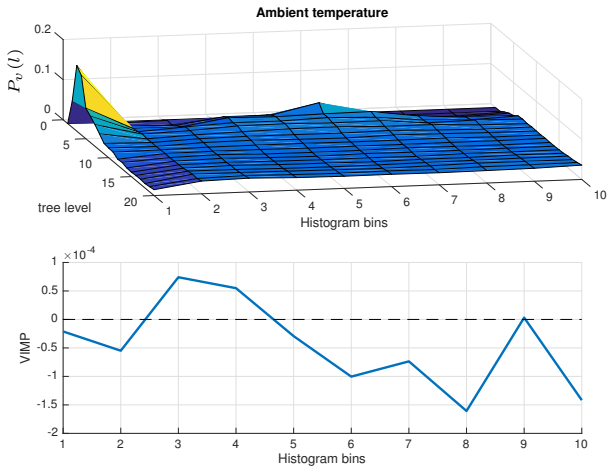


Fig. 10. Variable importance analysis of ambient temperature histogram variables.

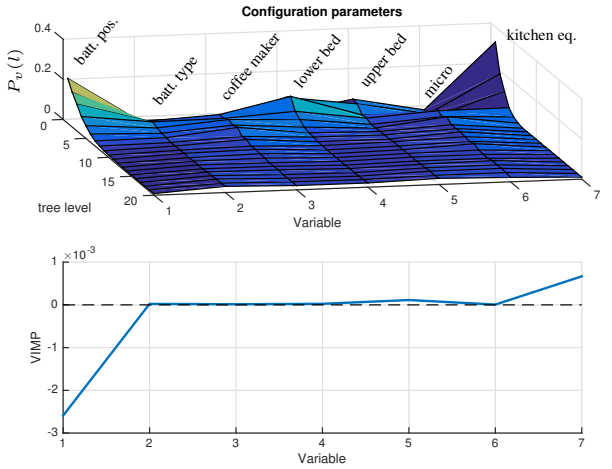


Fig. 11. Variable importance of different truck configuration variables.

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

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

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

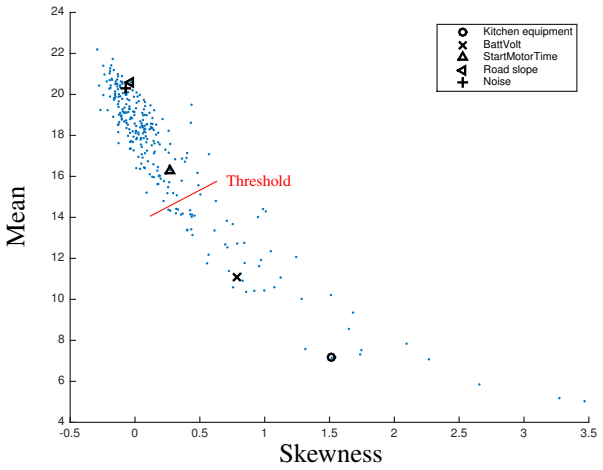


Fig. 12. Skewness and mean of (8) plotted for each variable in vehicle database. A manually selected threshold is used to select a subset of important variables which reside down to the right from the threshold.

failure prognostics and extracts information not obtained from VIMP or minimal depth metrics.

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

6.1 Variable selection using shape of depth distribution

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

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

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

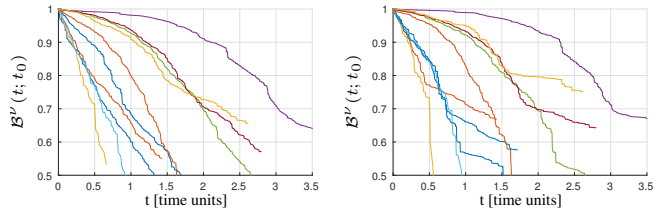
The corresponding skewness and mean for each of the variables in Fig. 6 are marked in Fig. 12 showing that the variables thought to be important are located down to the right while the noise variable is located up to the left.

7. EVALUATING RSF MODEL FOR BATTERY HEALTH PROGNOSIS

Based on Fig. 12, a manually selected threshold is defined to select a subset of variables that are most important to generate a new RSF. The performance of the RSF using the reduced set of variables is compared to using all variables. The selected subset

of variables includes 50 variables out of 283 variables. For both sets of variables, an RSF is generated with 1000 trees and a minimal terminal node size of 200. The error rate for the case with all features is 0.2011, and for the reduced set 0.2186 which are comparable in size. Note that, as observed in Section 2.2, this does not necessarily mean similar predictor performance.

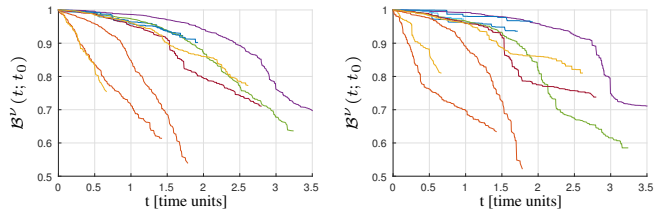
For the analysis, 10 vehicles with battery failures and 10 without are selected randomly. These vehicles are then used as inputs in the RSF to compute the life functions $\mathcal{B}^{\nu}(t;t_0)$ and the results are shown on Fig. 13 and Fig. 14 for vehicles with battery problems and healthy ones, respectively. It should be noted that time units were used on x axis for both figures, original time was scaled to hide true life-time of batteries to not reveal sensitive information for industrial partner.



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

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

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



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

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

8. CONCLUSIONS

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

ACKNOWLEDGEMENTS

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

REFERENCES

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

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

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

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

Frisk, E., Krysander, M., and Larsson, E. (2014). 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.

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

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

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

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

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

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