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

Sergii Voronov[1], Daniel Jung[2], and Erik Frisk[3]

[1,2,3] *Department of Electrical Engineering, Linköping University, Linköping, 581 83, SWEDEN*
sergii.voronov@liu.se
daniel.jung@liu.se
erik.frisk@liu.se

ABSTRACT

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

1. INTRODUCTION

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

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

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

2. PROBLEM FORMULATION

The main objective in this work is to use Random Survival Forests (RSF) (Ishwaran et al., 2008) to identify, from data, which variables are relevant for building RSF models for survival analysis. The problem of identifying important variables is usually referred to as variable selection and is a relevant topic in data-driven prognostics and machine learning in general (Guyon & Elisseeff, 2003).

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

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

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

2.1. Operational data

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

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

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

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

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

2.2. Motivation for variable selection

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

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

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

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

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

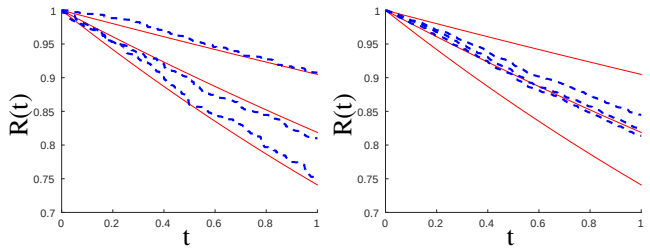
where h_0 is the nominal hazard rate. Data for 3000 vehicles is generated with a censoring rate about 80 percent. Different numbers of noisy variables are included in the synthetic data to observe how they change the RSF output.

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

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

computed by the two RSF models are compared with the theoretical values of the reliability as shown in Figure 1. One vehicle from each of the three classes was chosen and submitted to the forest to receive the predictions. It is shown in Figure 1 (a) that predictions from RSF for the case of 2

noisy variables, dashed blue curves, are following the theoretical reliability functions, red solid curves, better than the case with 100 noisy variables, see Figure 1 (b). However, the error rate, which is a common performance measure for the RSF, is similar for both cases. This means that the error rate is not a good measure in prognostic terms. It is worth to notice that in the simulation environment, information about the true reliability curves is available. However, this is not the case for the vehicle fleet database.



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

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

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

3. Random Survival Forests

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

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

RSF is a data-driven method that can be used for computing maximum-likelihood estimates of the reliability function (4). It can be used to rewrite the lifetime prediction function (1) as

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

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

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

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

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

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

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

3.1. Variable selection using VIMP

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

3.2. Variable selection using Minimal depth

As an alternative to VIMP, a candidate measure called minimal depth for variable selection in RSF has been proposed, see (Ishwaran et al., 2011) or (Ishwaran, Kogalur, Gorodeski,

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

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

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

4. Variable Depth Distribution Method

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

The VIMP and minimal depth measures are applied to the vehicle database and the results are shown in Figure 2 and Figure 3, respectively. As a reference, three variables, only containing Gaussian noise, are included in the data set. The computed VIMP is positive for half of the variables, but the VIMP curve

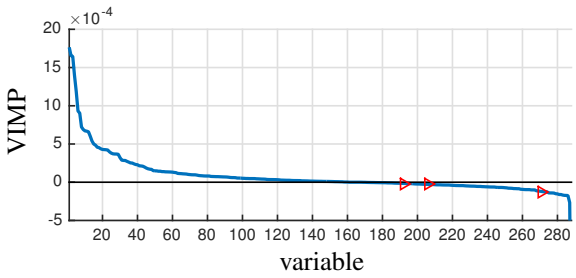


Figure 2. VIMP of variables in vehicle database sorted in ascending order.

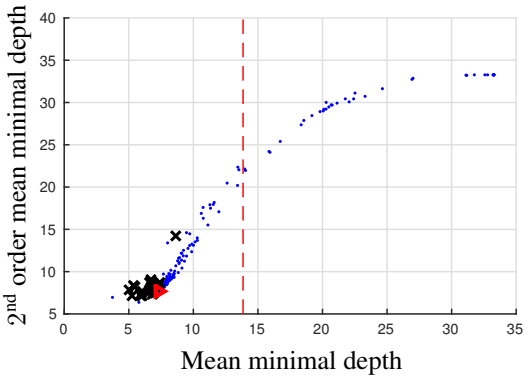


Figure 3. Minimal depth analysis of vehicle data. Black crosses correspond to 30 variables with highest VIMP and red triangles to added noise variables. Red dashed line is a threshold. Noisy variables should be located to the right of the red dashed line.

starts to flatten out after the first 30 variables with highest VIMP indicating that approximately 10% of the variables are expected to be relevant for battery lifetime prediction. The result of the minimal depth measure is presented in Figure 3 where the x axis is a mean value of the first appearance of the variable in the forest, y axis is a mean value of the second appearance of the variable in the forest, and the red dashed line is the threshold computed based on (8). The figure shows that most variables are identified as important, including the added noisy variables. Since the noisy variables are identified as important, it is an indication that minimal depth is not a suitable method for the vehicle database.

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

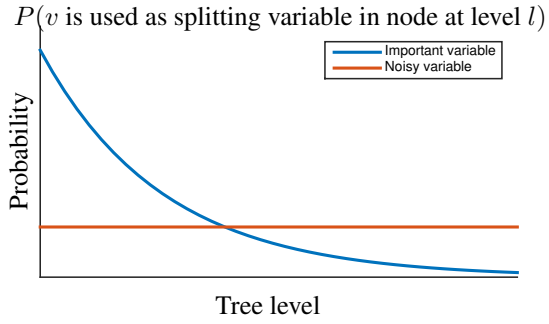


Figure 4. Illustrative example of the probability that a given splitting variable is used in a node at different tree levels.

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

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

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

The conditional probability (9) will be used as a variable importance measure. However, the true probability is not known because it depends on many different factors, for example, the parameters when generating the RSF. It can be noticed from (9) that $P(d|v) \propto P(v|d)$ and if $P(v|d)$ is known the value $P(d|v)$ could be found as well. After growing the forest, $P(v|d)$ can be estimated by first computing

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

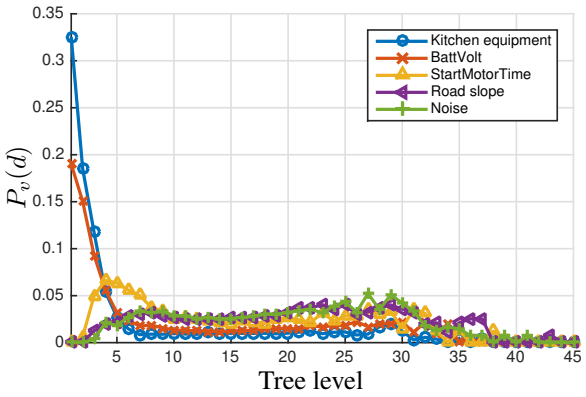


Figure 5. Examples of the estimated $P_v(d)$ for five different variables including one known noisy variable.

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

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

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

Instead of comparing the whole distribution $P_v(d)$ for each variable v, two representative features are considered, mean and skewness,

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

According to Figure 4 and Figure 5, an important variable should have high positive value of skewness and low value of mean. These two features can be used alone to identify which variables that are important. There is one drawback with this approach, namely, it is possible that a noisy variable will be selected by random at low level of a tree. It means it will have values of skewness and mean as important variable. However,

for a noisy variable, this is likely to be a rare event. Therefore, introducing information about how often a variable is used as a third dimension can help to filter out noisy variables in the area where important one should reside. Two possible candidates to express this information are:

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

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

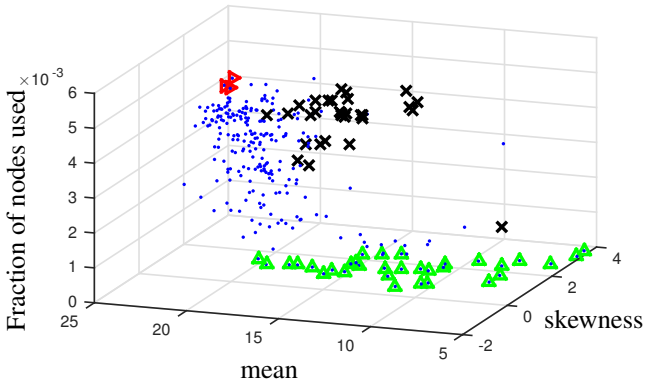


Figure 6. Skewness and mean of (10) of vehicle data combined with fraction of nodes. Black crosses correspond to 30 variables with highest VIMP, green triangles are variables rejected by minimal depth, and red triangles are added noise variables.

4.1. Real data case study

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

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

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

5. Analysis

Before continuing the analysis of the vehicle fleet data using the new variable selection method in the prognostic algorithm, the properties of the proposed measure in Section 4 are further analyzed. As mentioned in Section 4, there is no knowledge which variables are important in the vehicle database. There is an intuition that some of them could be informative, but it is not clear how many they are and what their influence is on the battery hazard rate.

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

5.1. Case study in simulated environment

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

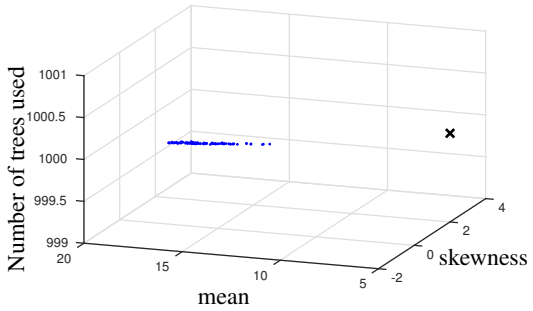


Figure 7. Simulated data from 10000 generated vehicles with censoring rate 80 percent. The important variable is marked with a cross.

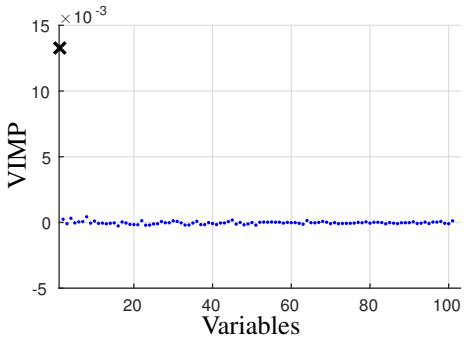


Figure 8. Computed VIMP of simulated data from 10 000 generated vehicles with censoring rate 80 percent. The important variable is marked with a cross.

hazard rate h_0 by a factor h_1 defined as

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

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

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

Figures 8 and 9 show the results for the same problem, but using VIMP and minimal depth respectively. Using VIMP, it is easy to identify the important variable, therefore, VIMP and the method proposed in the paper gives similar results in this case. In Figure 9, the red dashed line is the threshold that separates important and non-important variables according to (Ishwaran et al., 2011). Variables to the left of the threshold should be important and variables to the right are not. Figure 9 shows that the specific threshold is not able to distinguish important variables in this case. However, it is visible that it is

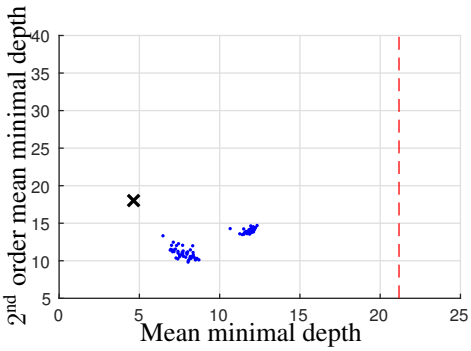


Figure 9. Minimal depth of simulated data from 10 000 generated vehicles with censoring rate 80 percent. The important variable is marked with a cross.

possible to manually select a threshold that could to that.

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

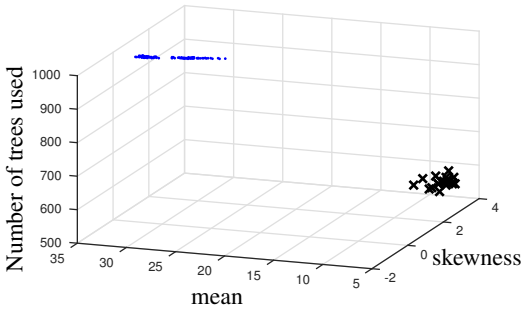


Figure 10. Simulated data from 10000 generated vehicles with censoring rate 80 percent. The important strongly correlated variables are marked as crosses.

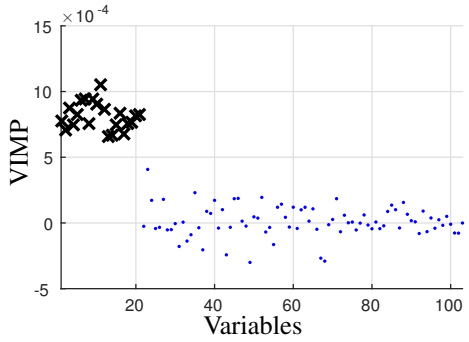


Figure 11. Computed VIMP for simulated data from 10000 generated vehicles with censoring rate 80 percent. The important strongly correlated variables are marked as crosses.

suffer of problems with correlated variables like VIMP do.

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

5.2. Strategy for variable selection

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

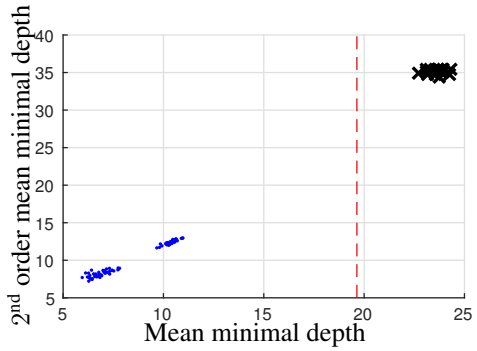


Figure 12. Minimal depth of simulated data from 10000 generated vehicles with censoring rate 80 percent. The important strongly correlated variables are marked as crosses.

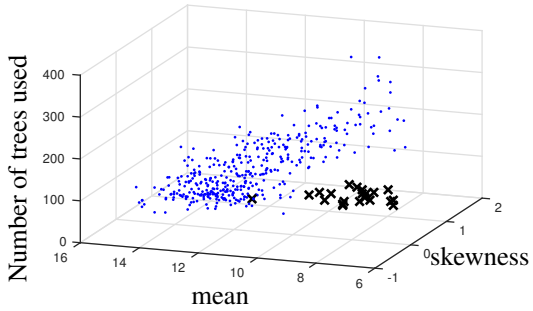


Figure 13. Simulated data from 500 generated vehicles with 21 strongly correlated important variables, 400 noisy variables, and censoring rate 80 percent. Important variables are highlighted with crosses.

hoc strategy.

Variables from the vehicle database are plotted in Figure 16 where the number of trees a variable is used in is used as the third dimension. Important variables should be used in most of the trees. Therefore, selecting a threshold that sorts out variables that are not used in many trees, for example 800, should give a first set of candidates of important variables. It could be the case that important variables are used less if there are many correlated variables, however, in that case it is expected that skewness and mean would be similar for those variables, Section 5.1. Then, there should be variables that are grouped in the skewness-mean plane which is not observed for the variables with values of number of trees less than 800. Therefore, it is assumed that there are no important variables in that area. It was shown in Section 5.1 that for some difficult cases, noisy variables are used as splitting variables more often than important variables, see Figure 13. Setting up a threshold with aforementioned strategy will not work for that case. This situation is not considered in this case study, but a more general strategy for selecting the threshold is required for a final version of the variable selection algorithm.

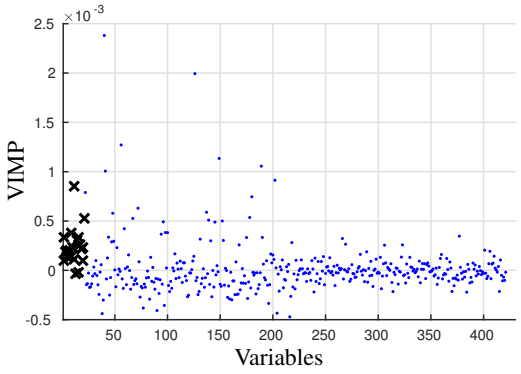


Figure 14. Computed VIMP of simulated data from 500 generated vehicles with with 21 strongly correlated important variables, 400 noisy variables, and censoring rate 80 percent. Important variables are highlighted with crosses.

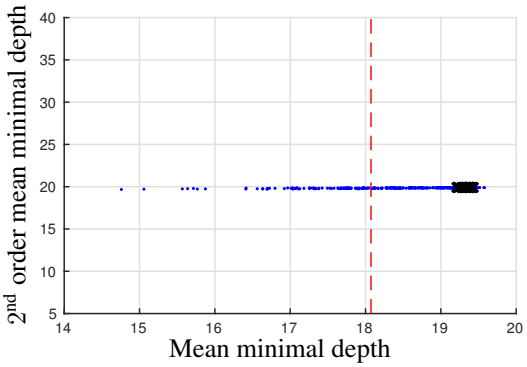


Figure 15. Minimal depth of simulated data from 500 generated vehicles with with 21 strongly correlated important variables, 400 noisy variables, and censoring rate 80 percent. Important variables are highlighted with crosses.

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

1. Set up threshold in the number of trees dimension to filter out noisy variables which are seldom used.
2. Project remaining variables in the skewness-mean plane and set up a threshold that distinguishes important variables as the subset of variables with high positive value of skewness and low value of mean.

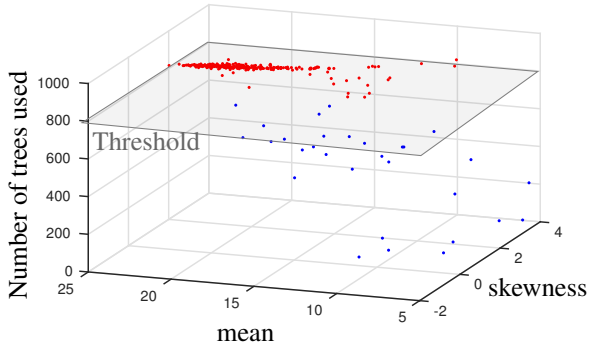


Figure 16. Setting up threshold for the vehicle database. x and y axis are skewness and mean of (10) respectively, and z axis is the number of trees in forest variable was chosen. Red points are candidates for important variables, blue dots - noisy variables, gray plane - threshold value.

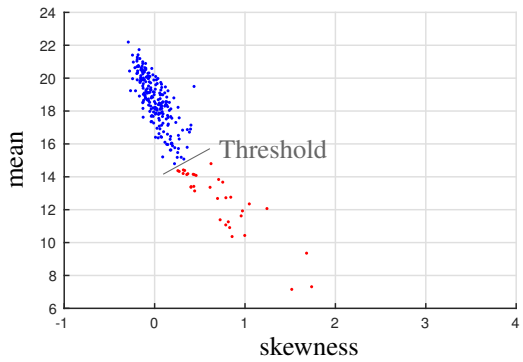


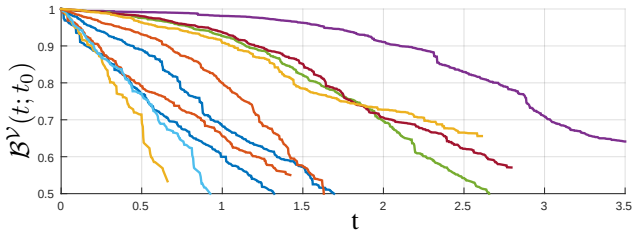
Figure 17. Setting up threshold for the vehicle database. x and y axis are skewness and mean of (10) respectively. Red points correspond to important variables, blue dots - to noisy.

6. Case study: Battery failure prognostics

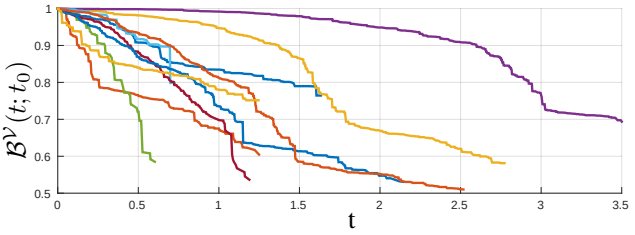
Using the manually chosen thresholds as described in Section 5 and demonstrated with the means of Figures 16-17, 34 of the variables, i.e. about 12 percent, are selected and treated as important. The performance of the RSF using the reduced set of variables is compared to using all variables. The performances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

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

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



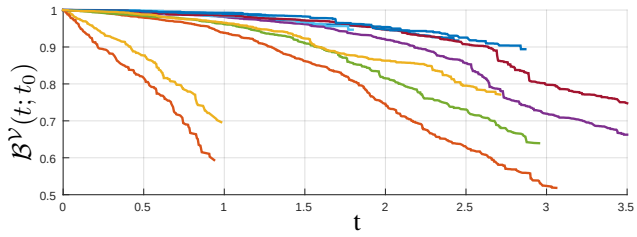
(a) RSF using all variables



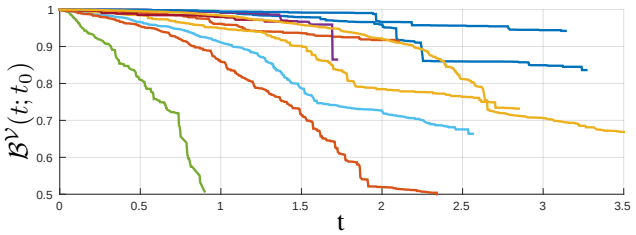
(b) RSF using selected subset of variables

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

In Figure 18 (b), vehicles are clearly more grouped compared to Figure 18 (a) where most vehicles have faster decaying lifetime prediction. The result seems reasonable since lifetime of the batteries of grouped vehicles with fast decaying lifetime prediction functions $\mathcal{B}^{\mathcal{V}}(t;t_0)$ in Figure 18 (b) are within 2 to 3 time units which is quite long life for batteries. Therefore, fast decaying lifetime prediction functions for those vehicles should be expected. Battery lifetime of the vehicle corresponding to the purple curve in Figure 18 is about 0.14 time units. However, the vehicle failed early and value of lifetime function would not allow to predict the failure, but it is possible that the cause of the battery problem is not so common in the vehicles from the database. In general, it could be seen that vehicles that lived longer are well separated from the ones that lived shorter. Of cause, it could not be used as the evaluation of the method, but as a positive sign. Note that the lifetime function of the vehicle which corresponds to the green curve in Figure 18 has changed significantly between the two figures. This vehicle operated for about 2.5 time units. It has not yet failed, but should be likely to fail soon. That is why the lifetime prediction function decays faster than for the other vehicles. It should be noticed that we need a measure for assessing predictive performance of RSF, and, when it is available, more can be said about the influence of variable selection on prognostic capabilities of the model.



(a) RSF using all variables



(b) RSF using selected subset of variables

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

7. Conclusions

A method for variable selection and variable importance analysis using random survival forests is proposed and analyzed. Main motivating factors for the approach are 1) small number of informative variables, and 2) highly correlated variables in the data set. Analyzing the feature space in Figure 6 indicates that it is possible to distinguish how VIMP and Minimal depth determines which variables that are considered important and this should be analyzed further. The proposed method is evaluated in the industrially relevant problem of heavy-duty vehicle battery failure prognostics and evaluated using real vehicle fleet data and simulated data. Simulated data shows that important variables can be distinguished from noisy variables even in difficult cases. The case study using real data shows that a prognosis model with 12% of the available variables achieves comparable error-rate with using all variables.

Acknowledgment

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

References

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

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

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

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

Ishwaran, H., Kogalur, U., Chen, X., & 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., & 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., & Zhou, D. (2011). Remaining useful life estimation–a review on the statistical data driven approaches. *European Journal of Operational Research*, *213*(1), 1–14.

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