Data-Driven Lead-Acid Battery Prognostics Using Random Survival Forests

Erik Frisk[1], Mattias Krysander[2], and Emil Larsson[3]

[1,2,3] *Department of Electrical Engineering, Linköping University, Sweden*
frisk@isy.liu.se
matkr@isy.liu.se
lime@isy.liu.se

ABSTRACT

Problems with starter batteries in heavy-duty trucks can cause costly unplanned stops along the road. Frequent battery changes can increase availability but is expensive and sometimes not necessary since battery degradation is highly dependent on the particular vehicle usage and ambient conditions. The main contribution of this work is a case-study where prognostic information on remaining useful life of lead-acid batteries in individual Scania heavy-duty trucks is computed. A data-driven approach using random survival forests is proposed where the prognostic algorithm has access to fleet management data including 291 variables from 33603 vehicles from 5 different European markets. The data is a mix of numerical values such as temperatures and pressures, together with histograms and categorical data such as battery mount point. Implementation aspects are discussed such as how to include histogram data and how to reduce the computational complexity by reducing the number of variables. Finally, battery lifetime predictions are computed and evaluated on recorded data from Scania's fleet-management system.

1. INTRODUCTION

To efficiently transport goods by heavy-duty trucks it is important that vehicles have a high degree of availability and in particular avoid becoming standing by the road unable to continue the transport mission. An unplanned stop by the road does not only cost due to the delay in delivery, but can also lead to 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. High availability can be achieved by changing batteries frequently but such an approach is expensive both due to frequent visits to a workshop and also due to the cost of the batteries. In addition, as will be shown, battery degradation is highly dependent on the particular usage and ambient conditions.

The main contribution of this work is a case-study, with methodological development and analysis results, based on fleet-management data from heavy-duty truck manufacturer Scania. A non-parametric and data-driven prognostics approach is used to compute, on an individual vehicle basis, prognostic information on remaining useful life of the lead-acid batteries in the vehicle. This information is then used to make dynamic and vehicle individual maintenance plans. The proposed approach mainly uses existing techniques but also some methodological development is done, in particular for handling histogram information and data reduction. The approach can be classified as a reliability function based prognostic approach (Linxia & Köttig, 2014).

The outline of the paper is as follows. First, Sections 2 and 3 introduces the case study and illustrates the characteristics of the studied problem and what problems that need to be solved to obtain a feasible solution. Section 4 then discusses the key step in the approach, how to estimate battery degradation properties based on fleet management data. One characteristic of the dataset is that it contains histogram variables and how they are introduced in the approach is discussed in Section 5. The fleet management dataset is large and Section 6 discusses how to extract the most important parts of the data to be used with the approach discussed in Section 4. Finally, Section 7 discusses how the proposed approach can be used in a prognostics and condition based maintenance setting and then some conclusions in Section 8.

2. PROBLEM BACKGROUND

There exist a number of approaches in the literature to do prognostics. One common approach is to look for trends in

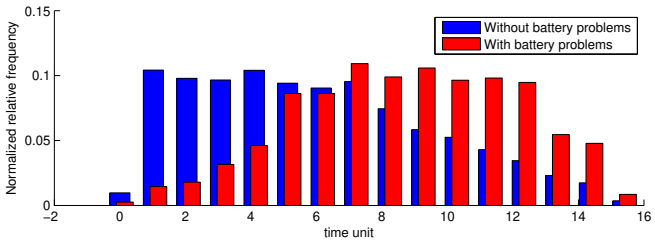


Figure 1. Normalized histogram of time stamp for vehicles with and without battery problems.

measured or estimated component health status indicators. Then, extrapolating computed health status indicators give indications on the amount of useful life left in the component. Such an approach requires reliable degradation models or measurements closely related to battery health, neither of which are available in this work. An alternative to a physics based approach where the battery health is estimated directly is to rely on recorded data from a large number of vehicles. This paper explores a data-driven approach where the prognostic algorithm has access to fleet management data and some characteristics of the data are

- 33603 vehicles logged from 5 different markets.
- 291 variables are logged for each vehicle.
- No time series, only aggregated data like traveled distance, year of delivery, histogram of ambient temperatures.
- Heterogeneous data; mix of numerical values such as temperatures and pressures with categorical data such as battery mount point or wheel configuration.
- Dataset includes histogram variables.
- Significant missing data rate ($\approx 15\%$).
- Each vehicle with a replaced battery has logged time of failure.
- There are many vehicles where battery failure has not occurred before the time of observation, i.e., data are right censored.

Figure 1 shows normalized relative frequency of logged time in the dataset. The red bars show the time of failure for vehicles with battery problems and the blue bars show time of logged data for vehicles with *no* battery problem. The histogram for vehicles with no battery problems thus reflect the last time data was logged from the vehicle which approximately is the age of the vehicle. Time is originally in days but has been scaled to *time units* to avoid revealing sensitive information. A first observation is that some batteries fail much earlier than others and that there clearly is potential in vehicle individual maintenance plans.

Let T be the random variable of failure time. Then the reliability function, sometimes referred to as the survival function, is the probability that $T \geq t$, i.e.,

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

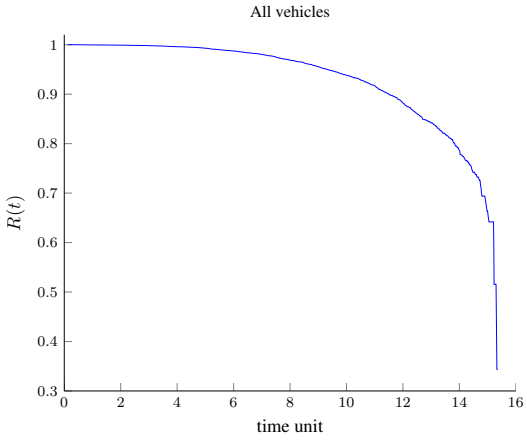


Figure 2. Reliability function estimate for the full dataset.

which is a fundamental object in the prognostics analysis. See Section 7 for further discussion on this. Estimating the reliability function from the data is basic survival data analysis and a non-parametric maximum-likelihood approach is used (Cox & Oakes, 1984). The reliability function estimate, based on the full dataset, is shown in Figure 2. This estimate would be most useful if it were true that the battery degradation is equal in all vehicles, no matter the vehicle configuration or usage. To investigate how much battery degradation characteristics change with vehicle configuration and usage, Figures 3 and 4 compare reliability function estimates for different subsets of vehicles. In Figures 3(a) and (b), different battery sizes and battery mounting positions are compared respectively. The reliability function estimate for battery size 140 Ah is based on very few vehicles, which is the reason for the jagged estimate. It is clear that battery size does not change the estimates significantly while battery mount position seems to have bigger impact. The battery size and battery position are both vehicle configuration parameters, naturally also usage parameters can have significant influence on battery degradation. Figure 4 shows reliability function estimates for vehicles with different amount of time with low battery voltage during cold ambient temperatures. Here it is clear that battery degradation significantly correlates with low temperatures and low voltages. The conclusion so far is then that truck battery degradation is dependent on vehicle usage and configuration. For each vehicle, 291 variables are recorded and it is not immediately clear which variables that are most important to describe different types of battery degradation profiles.

3. PROBLEM FORMULATION

The problem studied in this paper is to compute a probabilistic measure of the remaining useful life of a particular vehicle with a well functioning battery at a specified time $t = t_0$. As before, let T be the time of failure for the battery in a specific vehicle and let $\mathcal{V}$ denote usage and configuration data for the

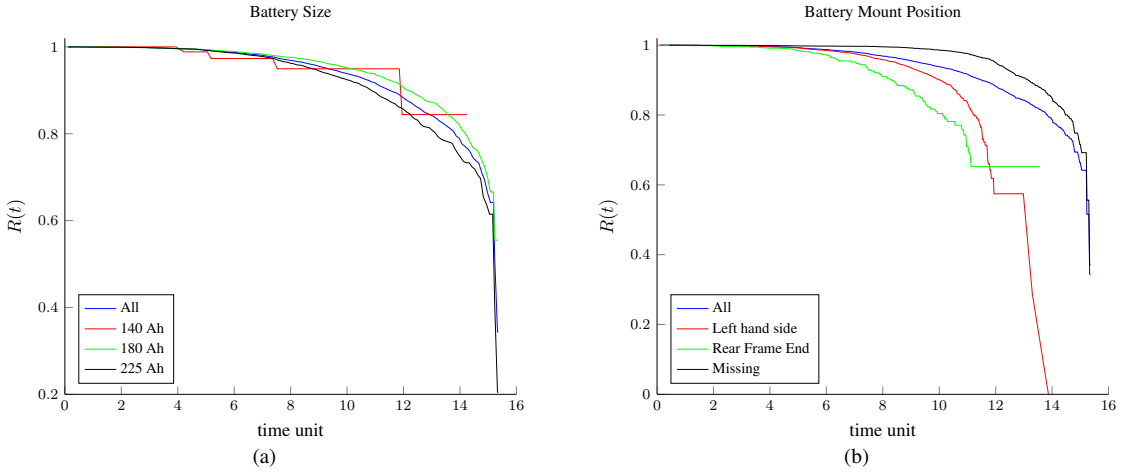


Figure 3. Reliability function estimation for different battery sizes (a) and different mounting positions (b).

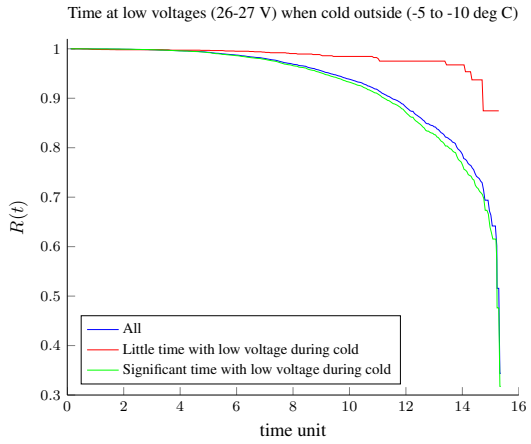


Figure 4. Reliability function estimate for vehicles with different amount of time with low battery voltage during cold ambient temperatures.

vehicle. The objective is to estimate the function

$$\mathcal{B}(t; t_0, \mathcal{V}) = P(T \geq t + t_0 | T \geq t_0, \mathcal{V}),\ t \geq 0 \qquad (2)$$

which describes, for a specific vehicle $\mathcal{V}$, the probability that the battery will at least t time units after t_0. This function is closely related to the reliability function $R(t)$. Let $R^{\mathcal{V}}(t)$ be the reliability function for a specific vehicle $\mathcal{V}$, then

$$\begin{aligned} \mathcal{B}(t; t_0, \mathcal{V}) &= P(T \geq t + t_0 | T \geq t_0, \mathcal{V}) = \\ &= \frac{P(T \geq t + t_0 | \mathcal{V})}{P(T \geq t_0 | \mathcal{V})} = \frac{R^{\mathcal{V}}(t + t_0)}{R^{\mathcal{V}}(t_0)} \end{aligned} \qquad (3)$$

The basic problem is then to, given the usage data for a vehicle $\mathcal{V}$, estimate $R^{\mathcal{V}}(t)$ and then compute $\mathcal{B}(t; t_0, \mathcal{V})$ according to (3). A key problem is that out of the 291 variables, it is not clear which ones that best capture different battery degradation characteristics. The main objectives of the paper are then to, in a case study with heavy-duty truck data,

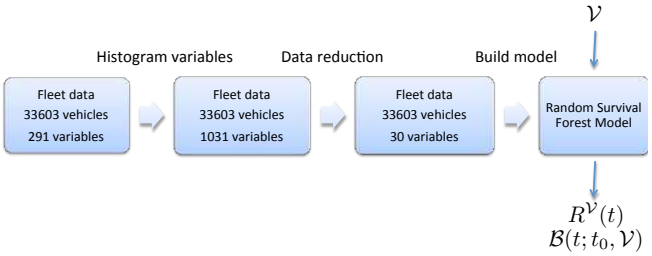


Figure 5. A flowchart describing the proposed approach.

- Determine, using machine-learning techniques, which of the 291 logged variables that are most useful for clustering vehicles with respect to battery lifetime prediction. Also analyze how to properly handle histogram variables.
- Estimate the reliability function $R^{\mathcal{V}}(t)$ for a specific vehicle $\mathcal{V}$.
- Estimate battery lifetime predictions as in (2) and evaluate on recorded data from Scania's fleet-management system.

The approach proposed for this problem is outlined in the flowchart in Figure 5. The flowchart illustrates how the original dataset first is extended with information about the histogram, which is described in Section 5. This leads to a significant growth in data size, which for complexity reasons results in a need to reduce the data before building models. The data reduction, here meaning selection of the 30 most important variables, is described in Section 6. Then, a random survival forest model is built as described in Section 4. With this model, a vehicle $\mathcal{V}$ and its associated 30 variables can be fed into the random survival forest model to compute prognostic information, which is illustrated in Section 7.

4. RELIABILITY FUNCTION ESTIMATION

Estimation of the reliability function (1) for a specific vehicle, based on a set of variables, is one of the main objective

of this work since then the function $\mathcal{B}(t; t_0, \mathcal{V})$ can be computed according to (3). As noted in Section 2, if it were a good assumption that battery degradation in all vehicles were independent on vehicle configuration, usage, and ambient conditions, a direct estimation of the reliability function using the very basic survival analysis techniques in (Cox & Oakes, 1984) would be appropriate. However this independence assumption is not realistic since it was shown how the failure rate of the battery varies significantly dependent on vehicle usage, configuration, and ambient conditions.

Thus, the 291 variables that are stored for each vehicle and describe vehicle configuration and usage need to be taken into account. One possibility is to use a parameterized approach where the failure rate of the batteries

$$h(t; \mathcal{V}) = P(T = t | T \geq t, \mathcal{V})$$

is written as a function of the variables $\mathcal{V}$. One common choice then is the proportional hazards model with log-linear hazards (Cox & Oakes, 1984) for which there exists well-established theory and tools. This approach is not used here, mainly because of the high rate of missing data which can not be handled directly, but also to avoid the proportional hazards assumption.

Instead, the basic idea of the approach used here can loosely be stated as utilizing a classifier to cluster vehicles with similar battery degradation properties. Then a non-parametric estimate for the reliability function $R^{\mathcal{V}}(t)$ is computed for a specific vehicle $\mathcal{V}$ using only the vehicles in the corresponding vehicle cluster.

A candidate tool that fits this situation well is Random Survival Forests (Ishwaran, Kogalur, Blackstone, & Lauer, 2008; Ishwaran & Kogalur, 2010). Random survival forest is a survival analysis extension of Random Forests (Breiman, 2001) which is a tree-based classifier (Breiman, Friedman, Stone, & Olshen, 1984) extended with bootstrap aggregation (Breiman, 1996) techniques. The key motives for using random survival forests in this work is that

- it handles heterogeneous data; both discrete and continuous valued variables
- it handles missing data
- it is non-parametric, i.e., does not rely on a specific hazard function parameterization like proportional hazards

There are 291 variables stored for each vehicle and the data includes 17 histograms. As will be described in Section 5, additional variables are derived to take these histogram variables into account. This results in a total of 1031 variables for each vehicle. To keep computational complexity down when building the random survival forest, Section 6 describes how to select the 30 most important variables. For this section it is not important exactly which variables that are used, it is enough to state that 30 variables were selected and used in the classifier.

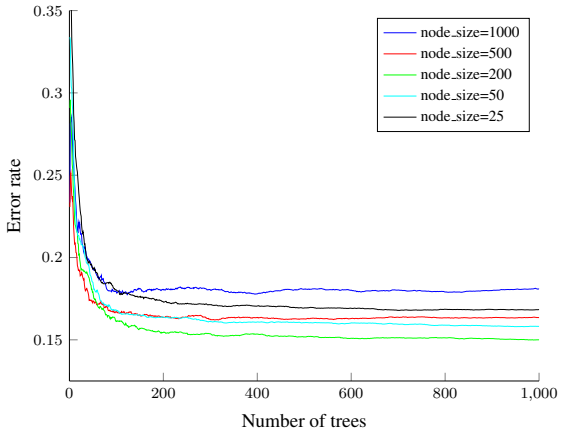


Figure 6. Error rate for the forest when node size is changed.

The experiments is conducted in R (R Core Team, 2014) using the package Random Forests for Survival, Regression and Classification (Ishwaran & Kogalur, 2013). There are 4 main parameters to be chosen in the software package

- number of trees to grow in the forest
- minimum size of terminal nodes
- number of random split variables
- number of random split values

Selection of these parameters is important for the result, and therefore there will be a short discussion on the choices made in this study. The remainder of this section requires knowledge of random survival forests, and for in-depth description of each parameter the reader is referred to (Ishwaran et al., 2008) and (Ishwaran & Kogalur, 2013).

The error rate measures how well the forest ranks two random individuals in terms of survival, and 0 is perfect and 0.5 is no better than guessing. The error rate can be interpreted as the probability of correctly ranking the survival of a batteries of two random vehicles. Formally, the error rate is $1 - C$ where C is Harrell's concordance index (Harrell, Califf, Pryor, Lee, & Rosati, 1982). Figure 6 plots the error rate as a function of node-size and number of trees. From this plot it is clear that, based on the error rate, there is no reason to grow more than about 200-300 trees in the forest and that the error rate is fairly insensitive to the selection of node size. The variance of the reliability function estimate dependends on the number of datapoints, i.e., too small terminal node sizes would give unreliable results. Based on Figure 6, the minimum terminal node size is chosen to 200.

The number of random variables to evaluate in each node of the tree classifier should not be too low, since then there is a lower probability of actually finding the best variable. Also, to get diversity among the trees in the grown forest, the number of variables should not be too high. As mentioned above and

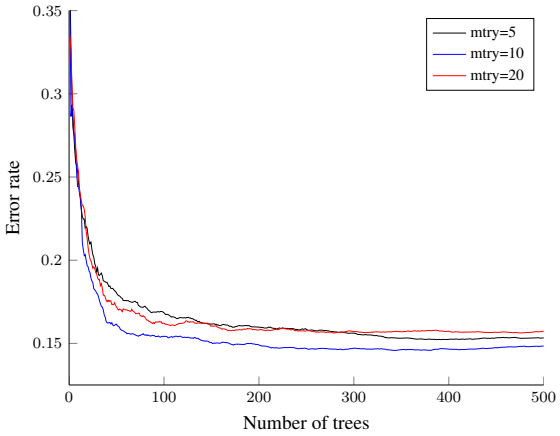


Figure 7. Error rate as a function of number of trees in the forest for three different number of random split variables to try in each node.

discussed further in Section 6, 30 variables are used in the analysis and Figure 7 shows the error rate for three different number of random variables explored in each tree node. Based on Figure 7, the number of random split variables to try in each node is selected to 10. The final parameter is the number of split values to try for each variable in each node. Due to the heterogeneous nature of the data, the package is configured for an exhaustive search for the best split value.

With the parameter values chosen, training the random survival forest with 200 trees, based on 30 variables for 33603 vehicles, takes about 15 minutes on the computer used for the experiments. The computer used has 128 GB of RAM and 2 Intel Xeon Processor X5675 (12M Cache, 3.06 GHz) resulting in 12 cores and 24 logical processors. In the experiment, 20 of the 24 logical processors were allocated in the tree computation. Note that training the forest is a one-time task, at least until more data becomes available, and predicting the reliability for a given vehicle is immediate.

5. HISTOGRAM VARIABLES

There are histograms in the available vehicle usage data and an example can be seen in Figure 10(a), which shows the fraction of time with a certain battery voltage. The frequencies of the observations in the intervals, the bin-values, are stored in the vehicle data. Thus, each bin-value is a variable that can be used for reliability function estimation.

By considering bin-values as independent variables, it is not taken into account that the bin-values represent frequencies of observations in intervals with known boundaries and that a histogram is an approximated probability distribution of a *single variable*. The mean and variance of a histogram are examples of properties that considers the underlying histogram variable and also take interval boundaries into account. Thus, could provide additional information for the reliability function estimation. To investigate properties of a histogram, a number of additional quantities, i.e., new variables, are derived for each histogram.

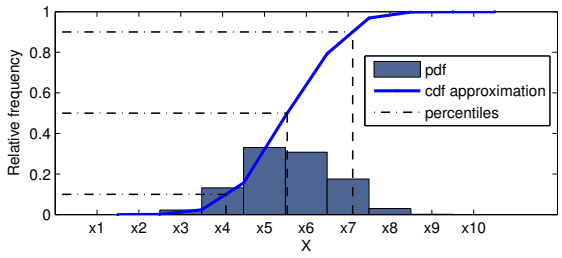


Figure 8. Histogram for a variable x.

Consider a histogram with n bins. Let p_i and x_i be the number of observations in and the center value of bin $i \in \{1, 2, \ldots, n\}$. The histograms are normalized such that the sum of bin values is one, i.e., $\sum_{i=1}^{n} p_i = 1$.

The variables considered for such a histogram are the bin values p_i for $i \in \{1, 2, \ldots, n\}$, the cumulative sum $c_i = \sum_{k=1}^{i} p_k$ for $i \in \{1, 2, \ldots, n\}$, the mean value of the histogram variable defined as $\mu = \sum_{i=1}^{n} p_i x_i$ and the variance

$$\sigma^2 = \sum_{i=1}^{n} p_i (x_i - \mu)^2$$

Furthermore the 10th, 50th (median) and 90th percentiles are computed from the cumulative distribution function based on a uniform distribution in each bin. Figure 8 illustrates the meaning of these values.

It is also natural that the tails, i.e., extreme cases of the distributions are of special importance. For example, a large number of starts with low battery voltage and almost none with high battery voltage could indicate battery problems. The following two variables have been included in the analysis to study the importance of the tails of the distribution.

Let the bin values of the mean histogram over all vehicles be denoted by $\bar{p}_i$ for $i \in \{1, 2, \ldots, n\}$. The number of bins that is considered as the left tail of the histogram n_- is computed from the mean histogram as $n_- = \max_n \sum_{i=1}^{n} \bar{p}_i < 0.05$. The number of bins considered as the right tail n_+ is computed analogously. Now, the tail variables considered for a histogram variable of a vehicle are computed as

$$\text{Ptail} = \sum_{i=1}^{n_-} p_i + \sum_{i=n-n_++1}^{n} p_i$$

and

$$\text{Mtail} = \sum_{i=1}^{n_-} p_i - \sum_{i=n-n_++1}^{n} p_i$$

6. VARIABLE IMPORTANCE

The dataset originally contains 291 variables where each bin in the histograms is counted as one variable. With the addition of the derived histogram variables described in Section 5 we obtain 1031 variables. To run the random survival forest algorithm considering the 291 variables takes 5 hours on the same machine that was described in Section 4. With 1031 variables, the computations did not finish in a reasonable time. To investigate parameter tuning of the forest, the algorithm has to be run with a number of different parameter settings. Then, also the run time with 291 variables is too long. To reduce computational complexity, the tree algorithms were run with 30 variables and this section describes how these variables have been selected.

To obtain accurate reliability functions it is important to use variables that are good at predicting battery failures. The predictive power of a variable will be called variable importance and this number can then be used to select the most important variables.

6.1. Method

Two different methods for computing variable importance have been investigated. The first method is based on the receiver operating characteristics curve, ROC-curve, and considers one variable at a time and the second is a multivariate analysis based on the error rate described in Section 4 computed by the random survival forest package.

Single variable analysis

The single variable analysis is based on the ROC-curve that shows the performance of a binary classifier. To introduce the ROC-curve, consider a hypothesis test concerning the battery of a vehicle with hypotheses

$$H_0 : \text{no battery problem}$$
$$H_1 : \text{battery problem}$$

For a variable x consider the test with threshold J and rejection region $\Phi(J) = \{x|x > J\}$ such that

$$\begin{aligned} x \notin \Phi(J) &: \text{accept } H_0 \\ x \in \Phi(J) &: \text{reject } H_0 \end{aligned} \tag{4}$$

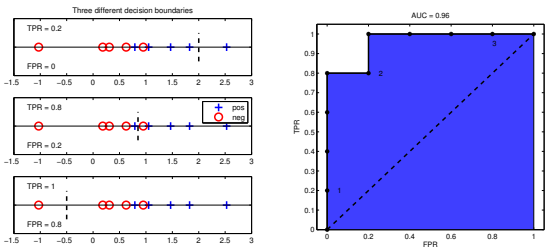
Two important properties of the test is the probability of detection, i.e.

$$P(D) = P(\text{reject } H_0 | H_1 \text{ is valid})$$

that ideally should be 1 and the probability of false alarm

$$P(FA) = P(\text{reject } H_0 | H_0 \text{ is valid})$$

which ideally is 0. Both the detection and false alarm probability is dependent on the threshold J and the ROC-curve is



(a) Example of variable values and 3 different thresholds.

(b) The ROC-curve for the dataset in Figure 9(a).

Figure 9. Example of an ROC-curve.

a plot of probability of detection $P(D)$ as a function of false alarm probability $P(FA)$. The curve is obtained by varying the threshold J.

An example of an ROC-curve is shown in Figure 9. Figure 9(a) shows the observations of a hypothetical variable used for classifying battery problems. The red circles are observations for vehicles without battery problems and the blue crosses observations from vehicles with battery problems. The value from vehicles with battery problems tends to be bigger than the values for vehicles without battery problem thus the variable could be used to separate those cases. The three different plots shows with a dashed vertical line different thresholds J and the true positive rates (TPR), i.e., the probability of detection, and the false positive rates (FPR), i.e., the probability of false alarm is shown.

The ROC-curve is shown in Figure 9(b) and is obtained by estimating the probabilities $P(D)$ and $P(FA)$ for thresholds J of different values. The numbers 1-3 refers to the 3 different thresholds shown in Figure 9(a). Consider for example the threshold in the second plot of Figure 9(a). Since 4 out of the 5 cases with battery problems are above this threshold the detection probability is estimation is $P(D) = 0.8$ and since 1 out of 5 cases without battery problems is above the threshold $P(FA) = 0.2$. This point is marked with a 2 in Figure 9(b). Variable importance for a variable x is then computed as the area under the ROC-curve (AUC) as

$$\text{AUC}(x) = \int_0^1 \text{ROC}(x)\, dx$$

For the example the AUC is 0.96.

The AUC is between 0 and 1. A value below 0.5 indicates that the observations from vehicles with battery problems are in general smaller than the observations of vehicles with fault free battery. In this situation a battery fault should be detected if the variable is below the threshold instead, i.e., to change the rejection region in equation (4) to $\Phi(J) = \{x|x < J\}$ and the AUC becomes 1 subtracted with the unmodified AUC. Hence all variables will get an AUC between 0.5 and 1 where a bigger value indicates a more important variable.

Multivariate analysis

Variable importance can also be computed using the error rate described in Section 4 as suggested in (Ishwaran et al., 2008, 2007). Variable importance for a specific variable x is evaluated by subtracting the error rate using all variables from the error rate obtained without using x. The error rate without x is evaluated on the original trees grown with x and whenever a split for variable x is encountered a daughter node is randomly assigned.

Advantages with this way of computing variable importance compared to the AUC-method is that the error rate is more closely related to our primary goal, i.e., to estimate the reliability function accurately and that the correlation of variables is considered. A disadvantage is the computational complexity of growing the trees needed to evaluate the error rates.

6.2. Case study results

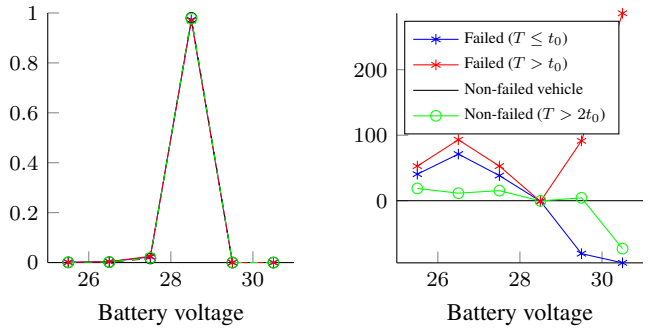
As said in the beginning of Section 6 the 30 most important of the total 1031 variables was selected as a trade-off between computational complexity and prediction performance. Since variable importance based on error rate requires the computation of a forest, the simpler AUC score has been used for the selection. The selection has been done in two steps. In the first step, the two most important variables of each histogram have been selected considering a variable correlation condition described later. In the second step, the 30 most important variables are selected among all non-histogram variables and the variables selected in first step. Since variable importance based on error rate is more closely related to reliability function prediction a comparison of the AUC-based ranking and error rate ranking is given in the end for of this section for the 30 selected variables.

Analysis of histogram variables

For each histogram stored in the dataset the variables described in Section 5 have been computed and the importance of them ranked according to the AUC.

Figure 10 shows an example of the mean histogram representing the relative time spent with a certain battery voltage when the battery temperature has been in the range of 10 to 25°C. To see how battery health effects the battery voltage the vehicles has been divided into 3 groups: vehicles with battery failure $T \leq t_0$, vehicles with battery failure $T > t_0$, and vehicles without any observed battery failure. Within the last set of vehicles also those with a long censoring time $T > 2\,t_0$ is shown separately. Figure 10(b) shows the relative deviation from the mean histogram under the fault free case. It can be seen that battery voltage is low more often for vehicles with battery failures.

Figure 11 shows variable importance based on AUC-score. The variables are introduced in Section 5 where pct stands



(a) Battery voltage histogram.

(b) Relative deviation in percent from fault free distribution.

Figure 10. Histogram for variable BattVoltTempI3, i.e., battery voltage when the battery temperature is in the range of 10 to 25°C.

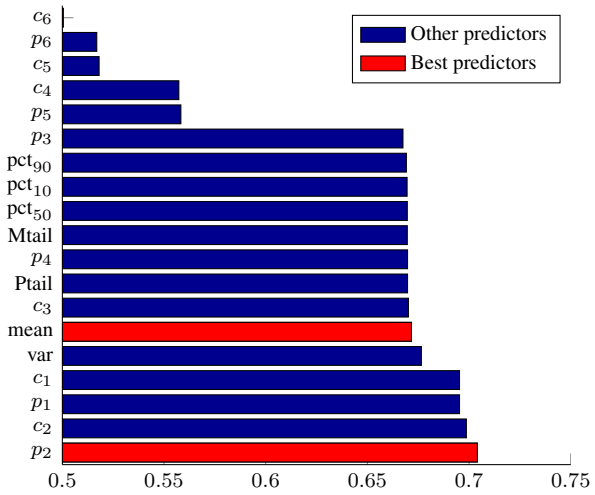
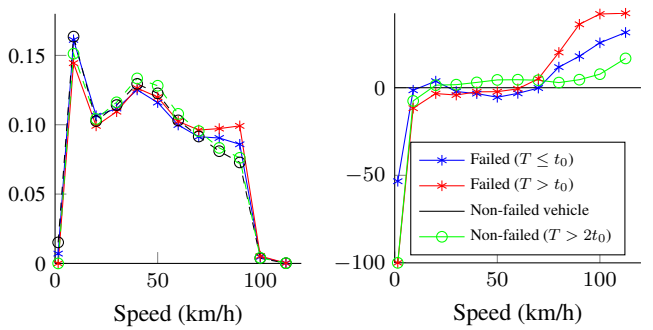


Figure 11. Importance of variables defined by the histogram for BattVoltTempI3 shown in Figure 10.

for percentile and Mtail and Ptail for minus and plus tail respectively. The most important variable of the histogram is p_2 which corresponds to the relative time with battery voltage between 26 and 27V. It can be seen that p_2 seems reasonable by looking at Figure 10(b) where the vehicle with failed batteries have a higher value than for the vehicles with non-failed batteries.

The next most important variable is c_2, i.e., the sum of the first two bins. Obviously c_2 is rather correlated with p_2 and to avoid the inclusion of highly correlated variables the most important variable is selected and the most important variable with a correlation with the most important variable less than 0.4. In this case, the mean value of the histogram will be the second selected variable.

For this histogram the original variable, p_2 was most important but the next histogram is an example where some of the derived variables are most important. Figure 12 shows a histogram for



(a) Histogram of speed in km/h when beginning to brake.

(b) Relative deviation in percent from fault free distribution.

Figure 12. Histogram for variable BrakeStartSpeed, i.e., initial vehicle speed when beginning to brake.

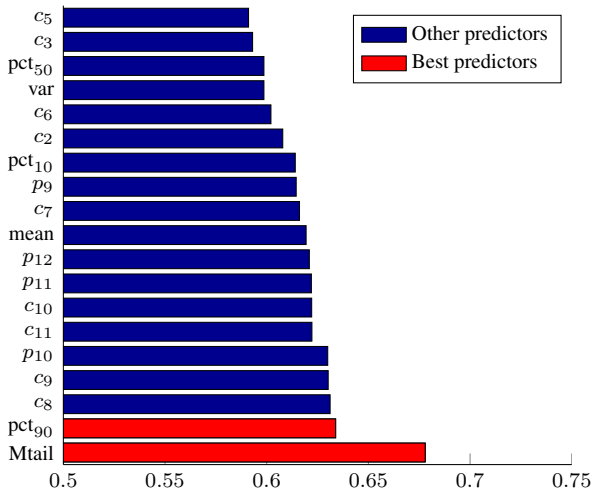


Figure 13. Importance of variables defined by the histogram for BrakeStartSpeed shown in Figure 12.

vehicle initial speed when beginning to brake. Figure 12(a) shows the histogram and Figure 12(b) the relative deviation from the mean histogram including only the vehicles without battery problems. Figure 13 shows variable importance for the variables related to the histogram in Figure 12. The most important variables here are the derived variables Mtail and pct90 and it can be seen in Figure 12 that vehicles with battery failures are more often beginning to brake at higher speeds.

As a summary of the histogram analysis, a number of variables has been derived for each histogram and two of the most important variables has been selected for each histogram when considering variable correlation. In the following analyses, only the two selected variables for each histogram will be considered together with all non-histogram variables.

Analysis of all variables

The remaining set of variables includes the selected histogram variables and the non-histogram variables and contains 117 variables. The 30 most important variables of these 117 variables are selected by using the AUC-based score and the top 18 are shown in Figure 14(a). The variables are categorized as bin variables p_i, non-histogram variables, or derived histogram variables. Among the selected 30 variables there are 5 non-histogram variables, 12 bin variables, and 13 derived histogram variables. Hence, some of the derived variables for the histograms are important. The individual variables with most predictive power are the total distance driven, time of delivery, and the number of days in use. The two most important bin variables are BattVoltTempI2_p2 which corresponds to low battery voltage at relatively low temperatures -5 to 10°C and BattVolt_p2 which corresponds to low battery voltage in general. The most important derived histogram variable concerns low ($< 20\%$) and high ($> 80\%$) state of charge when estimated after 8-24h without battery load. The variable importance based on error rate has also been computed of the top 30 variables in Figure 14(a) and the result is shown in Figure 14(b) where the top 18 variables are shown. Both rankings are quite similar. For example among the top 10 most important variables in each ranking 9 are the same. Thus even if the simpler AUC-based score has been used for variable selection the similarities with the more advance error rate based score is promising.

7. PROGNOSTICS AND CONDITION BASED MAINTENANCE

The main objective so far has been to compute the battery lifetime prediction function $\mathcal{B}(t; t_0, \mathcal{V})$ through estimation of the reliability function $R^{\mathcal{V}}(t)$ as described in Section 4 and then use (3).

With the reliability function and the battery lifetime prediction function, there are several ways to pass information to a condition based maintenance planner. One simple and direct way is to schedule the time for next maintenance T_{maint} no later than a time where the probability of a non-functioning battery is less than a certain threshold value. Formally,

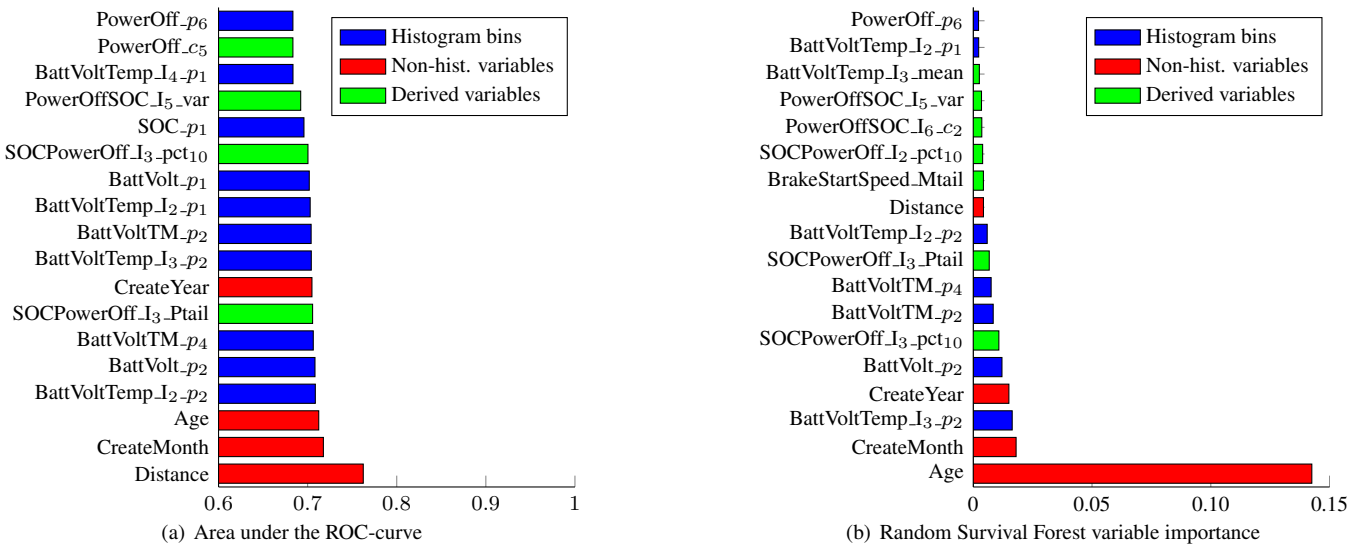
$$T_{\text{maint}} \leq \arg\min_t(\mathcal{B}(t; t_0, \mathcal{V}) < J) \tag{5}$$

where J is some predefined threshold. Another possibility is to compute the expected remaining useful life of the battery for a specified vehicle. Let $f(t)$ be the battery lifetime distribution. By definition it holds that $f(t) = -\frac{d}{dt}R(t)$ and then by partial integration

$$E(T) = \int_0^\infty t f(t)\,dt = -\int_0^\infty t\frac{d}{dt}R(t)\,dt = \int_0^\infty R(t)\,dt$$

This expression then gives that the expected remaining useful life of a battery in a vehicle $\mathcal{V}$, given that life up to $t = t_0$ is observed, is given by

$$E(\text{RUL}(t_0, \mathcal{V})) = \frac{1}{R^{\mathcal{V}}(t_0)}\int_{t_0}^\infty R^{\mathcal{V}}(t)\,dt - t_0$$



(a) Area under the ROC-curve

(b) Random Survival Forest variable importance

Figure 14. Individual predictive power for the most influential variables based on the area under the ROC-curve and ranking based on variable importance in the random survival forest.

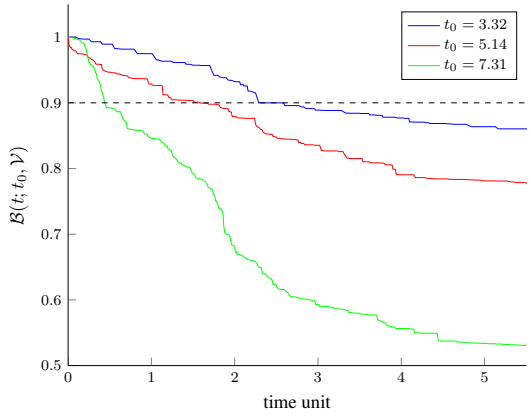


Figure 15. Function $\mathcal{B}(t; t_0, \mathcal{V})$ for three different vehicles with $t_0 = 3.32$, 5.14, and 7.31 time units respectively.

Although the expectation of remaining useful life is attractive, it involves integrating the estimated reliability function to infinity. Unfortunately, the estimated reliability functions has a high degree of uncertainty for large values of t. This is due to that there are very few recorded data points for large t and therefore this approach is not pursued further here. Instead, the battery lifetime prediction function is used as in (5).

Figure 15 shows the estimated $\mathcal{B}(t; t_0, \mathcal{V})$ function for three different vehicles selected from the set of all logged vehicles. For example, the figure shows how the probability of battery failure is increasing with increasing number of days in use. With a threshold of $J = 0.9$, the corresponding maintenance time T_{maint} should be no later than 2.29, 1.59, and 0.44 time units respectively. It is clear from Figure 15 that the expected battery lifetime prediction varies significantly for different vehicles. But that is to be expected since the three vehicles has been in operation significantly different amount of time.

In Figure 15 there are no confidence intervals or standard-error estimates. This is unfortunate since it is then difficult to assess how reliable the estimate of the reliability function is. To our knowledge, there is no standard way of estimating standard errors for bagged learners and random forests. Estimating confidence intervals for random survival forests is an active research area and one possible approach is described in (Wager, Hastie, & Efron, 2014).

To further investigate the impact on battery degradation from different usage profiles, ambient conditions, and vehicle configurations, Figure 16 shows the estimated battery lifetime prediction function for 20 vehicles with almost the same time in operation, about $t_0 = 5$ time units. Here it is clear that, even with similar time in operation the expected lifetime of the battery varies significantly. For example, comparing the vehicle with the worst predicted outcome with the vehicle with the best predicted outcome, the former vehicle has about 3% longer time in operation, which can not alone explain the big difference in predicted battery degradation. However, looking at the time with low battery voltages and low ambient temperatures, exactly as was done in Figure 4, it shows that the vehicle with worse battery lifetime prediction has spent significantly more time in that operating point. This also suggests that the dataset predicts that it is not sufficient to consider calendar time and mileage to get efficient vehicle individual maintenance plans.

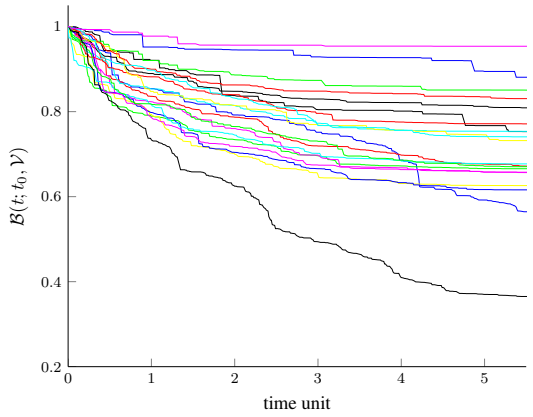


Figure 16. Battery lifetime prediction function $\mathcal{B}(t;t_0,\mathcal{V})$ for 20 vehicles with $t_0 \approx 5$ time units.

8. CONCLUSIONS

High degree of availability and reliability is important in many businesses and in particular heavy-duty trucks and the lead-acid battery is one important component to maintain. The battery is a difficult component to predict since degradation heavily relies on usage profile, vehicle configuration, and ambient conditions.

The main contribution is a case-study utilizing a data-driven approach to compute probabilistic reliability properties for a battery in a specific vehicle thus making condition-based maintenance feasible. The case-study is based on vehicle data from 33603 vehicles. A second contribution is the exploration of Random Survival Forests (RSF) for battery prognostics, and it is shown why RSF is a suitable tool in this application. A third main contribution is the study of which variables in the vehicle data that are important to characterize battery degradation. In particular a procedure is proposed how to include histogram data in the analysis.

The approach is evaluated using fleet-management data from truck manufacturer Scania and it is successfully shown how probabilistic reliability information can be estimated for the battery in individual trucks.

ACKNOWLEDGMENT

The authors acknowledge the initial work in the project done by Patrik Önnergren and Johanna Rosenvinge. This work was sponsored by Scania and VINNOVA (Swedish Governmental Agency for Innovation Systems) and the Swedish Research Council within The Linnaeus Center CADICS.

REFERENCES

Breiman, L. (1996). Bagging predictors. *Machine learning*, *24*(2), 123–140.

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

Breiman, L., Friedman, J., Stone, C. J., & Olshen, R. A. (1984). *Classification and regression trees*. CRC press.

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

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

Ishwaran, H., & Kogalur, U. (2013). Random forests for survival, regression and classification (rf-src) [Computer software manual]. manual. Retrieved from `http://cran.r-project.org/web/packages/randomForestSRC/` (R package version 1.4)

Ishwaran, H., & Kogalur, U. B. (2010). Consistency of random survival forests. *Statistics & probability letters*, *80*(13), 1056–1064.

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

Ishwaran, H., et al. (2007). Variable importance in binary regression trees and forests. *Electronic Journal of Statistics*, *1*, 519–537.

Linxia, L., & Köttig, F. (2014, March). Review of hybrid prognostics approaches for remaining useful life prediction of engineered systems, and an application to battery life prediction. *IEEE Transactions on Reliability*, *63*(1), 191–207.

R Core Team. (2014). R: A language and environment for statistical computing [Computer software manual]. Vienna, Austria. Retrieved from `http://www.R-project.org/`

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

BIOGRAPHIES

Erik Frisk was born in Stockholm, Sweden in 1971. He received a PhD degree in 2001 from Linköpings University, Sweden. Currently he has a position as an associate professor at the Department of Electrical Engineering at Linköping University. His current research interests in the field of model based diagnosis, prognosis, and autonomous vehicles.

Mattias Krysander was born in Linköping, Sweden in 1977. He received a M.S. in electrical engineering in 2000, and a Ph.D. degree in 2006, both from Linköpings University, Sweden. Currently he has a position as an associate professor at the Department of Electrical Engineering at Linköping University. His current research interests in the field of model based diagnosis and prognosis. Fault isolation and sensor placement