Treatment of accumulative variables in data-driven prognostics of lead-acid batteries

Erik Frisk and Mattias Krysander

Department of Electrical Engineering
Linköping University, Sweden
e-mail: {frisk,matkr}@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 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 used where the prognostic algorithm has access to fleet operational data including 291 variables from 33603 vehicles from 5 different European markets. A main implementation aspect that is discussed is the treatment of accumulative variables such as vehicle age in the approach. Battery lifetime predictions are computed and evaluated on recorded data from Scania's fleet-management system and the effect of how accumulative variables are handled is analyzed.

Keywords: Battery prognostics, reliability, survival analysis, machine learning, classification

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, cooling, and kitchen equipment. High availability can be achieved by changing batteries frequently but such an approach is expensive both due to unnecessary maintenance actions and also due to the cost of the batteries. In addition battery degradation is highly dependent on the particular usage and ambient conditions.

A non-parametric and data-driven prognostics approach was developed in (Frisk et al., 2014) to compute, on an individual vehicle basis, prognostic information on remaining useful life of the lead-acid batteries. Prognostic information is computed by applying a tree based classification method called Random Survival Forests (RSF) (Ishwaran et al., 2008; Ishwaran and Kogalur, 2010) on fleet operational data from the heavy-duty truck manufacturer Scania. The approach can be classified as a reliability function based prognostic approach (Linxia and Köttig, 2014).

⋆ This work was sponsored by Scania and FFI - Strategic Vehicle Research and Innovation (Swedish Governmental Agency for Innovation Systems) and the Swedish Research Council within The Linnaeus Center CADICS.

The basic idea is to classify vehicles with similar battery degradation and for each class estimate a reliability function in a training phase. Then, when prognostics for a specific vehicle is computed, the reliability function can be obtained by identifying which class the vehicle belongs to and compute the corresponding reliability function. The data contains accumulative variables such as driven distance and vehicle age. The accumulative variables will increase over time and if these variables are used in the classification, a vehicle used in a similar way for its entire life will change class over time, which is not desirable. The main contribution of this work is to investigate how accumulated variables can be handled in RSF and how they affect the result.

The outline of the paper is as follows. First, Section 2 introduces data and briefly recalls a case study (Frisk et al., 2014) based on the same data set. Section 3 states the studied problem. Section 4 recalls how to estimate battery degradation properties based on fleet operational data by using random survival forests. One characteristic of the data set is that it contains variables that are accumulated over time and how they can be introduced in the approach is discussed in Section 5. Finally, Section 6 analyze and discuss how the prognostic result depends on the way accumulated variables are included and then some conclusions are given in Section 7.

2. BACKGROUND

There exist a number of approaches in the literature to do prognostics. A physics based approach is to look for trends in measured or estimated component health status indicators, see e.g. (Heng et al., 2009). 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 measure-

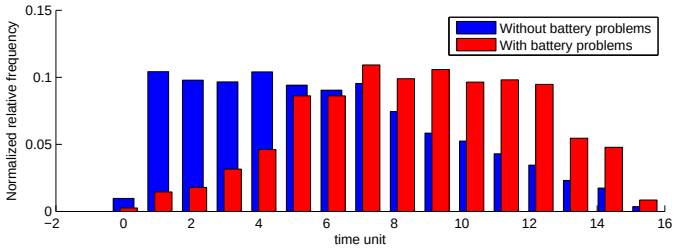


Figure 1. Normalized histogram of time stamp for vehicles with and without battery problems.

ments 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 operational 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.
- Data set 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 data set. 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 reflects 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 in (Frisk et al., 2014) it has been shown that battery usage and vehicle configuration have a big impact on battery degradation. For example, the battery failure rate is significantly different for different vehicles, e.g., a long-haulage vehicle with a large battery, kitchen equipment, and driving in cold weather may experience significantly different battery degradation behavior than a city distribution truck. A more detailed discussion is given in Section 4. Hence there clearly is potential in vehicle individual maintenance plans.

2.1 Prognostics Approach

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

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

which is a fundamental object in the prognostics analysis. Since vehicle configuration and usage is important for battery reliability, let $\mathcal{V}$ denote configuration and usage data for a vehicle and let $R^{\mathcal{V}}(t)$ denote the reliability function for that particular vehicle. In (Frisk et al., 2014)

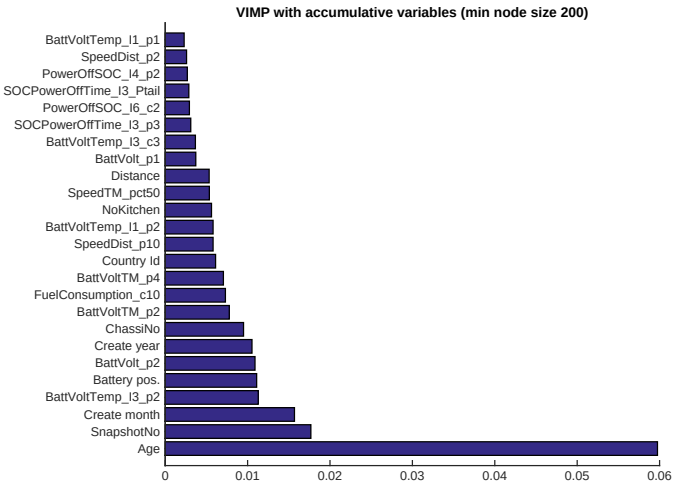


Figure 2. Variable importance.

Random Survival Forests (RSF) (Ishwaran et al., 2008; Ishwaran and Kogalur, 2010) have been used to estimate vehicle specific reliability functions. The key motives for using random survival forests for the available data are

- 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
- it handles censored data

The basic idea of the approach 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.

The RSF algorithm also automatically computes which variables that are important for clustering vehicles with similar battery degradation, i.e., which variables that are important for predicting battery degradation. Figure 2 shows a list of the 20 most important variables, when considering also accumulative variables, and their variable importance (VIMP), which is defined and discussed in more details in (Ishwaran et al., 2008, 2007). The most important variable, and its corresponding strength, is the undermost variable in Figure 2.

There are configuration variables such as battery position `Battery pos.` and country index `Country Id` and usage variables such as battery voltage `BattVolt_p2` and driven distance `Distance`. Configuration variables describe the vehicle configuration which does not change with time while usage variables changes as the vehicle is used. Variables with the suffix `x_pi` represents the frequency of a bin in histogram `x`, `x_ci` a cumulative sum of bins in histogram `x`, `x_pct50` the median of histogram `x`, `x_Ptail` the weight of the tails of histogram `x`, see (Frisk et al., 2014) for more details.

2.2 Accumulative Variables

Most of the variables in the Figure 2 are more or less constant over time if the vehicle is operated in a similar way over time. However there are a couple of exceptions. Vehicle age will of course increase with time and if this variable is used as a classification variable there is a risk of estimating

a reliability function based on vehicles observed only with a similar age. Then the reliability function estimate will only be changing values in a tight age interval. If age t would be omitted as a classification variable, it would still be used in the prediction step since t is used to evaluate the reliability function $R^{\mathcal{V}}(t)$, see Section 3. Another example of a variable that is accumulated over time is the traveled distance. Variables that are accumulated over time will be called accumulative variables and this paper investigate how to include accumulative variables in RSF.

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 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 be operational 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 how to handle accumulative variables in the classification method. The main objectives of the paper are to, in a case study with heavy-duty truck data, analyze and compare the difference in the results obtained with or without including accumulative variables in the classification approach. In particular the effects on variable importance and the estimate of the reliability function $R^{\mathcal{V}}(t)$ for a specific vehicle $\mathcal{V}$ will be studied. Vehicles with similar age and distance but different battery predictions are compared and their differences in operation and configuration are analyzed.

4. PROGNOSTICS WITH RANDOM SURVIVAL FORESTS

This section will briefly outline the algorithm used to estimate the battery prognostics function $\mathcal{B}(t; t_0, \mathcal{V})$ as defined in (2). The key step, from (3), is to estimate the reliability function (1). Thus, a reliable estimate of the reliability function $R^{\mathcal{V}}(t)$ for a specific vehicle $\mathcal{V}$ makes it possible to compute the prognostics function $\mathcal{B}(t; t_0, \mathcal{V})$.

4.1 Reliability Function Estimation

Basic techniques for maximum-likelihood estimation of reliability functions can be found in (Cox and Oakes, 1984). As will be described below, they are not directly applicable to this case but they are useful so first a brief summary of a basic result for reference purposes. Derivations and details of these expressions can be found in (Cox and Oakes, 1984). Now, assume N vehicles with age t_i and response variable c_i for $i = 1, \ldots, N$. The response variable $c_i = 0$ if the i:th vehicle does not have battery problems at time $t = t_i$ and $c_i = 1$ if the vehicle has battery problems. Then, in discrete-time, the maximum-likelihood estimator of the hazard function, i.e., immediate hazard-rate, at time-point $t = t_i$ can be found as

$$\hat{h}_i = \frac{d_i}{u_i} \qquad (4)$$

where d_i and u_i is the number of battery failures and the number of vehicles at risk at time $t = t_i$ respectively. Here it is explicitly taken into consideration that data might be right right censored, i.e., the time of battery failure is unknown but is known to be greater than the time of observation. The Kaplan-Meier (Product-limit) estimator of the reliability function $R(t)$ is then

$$\hat{R}(t_i) = \prod_{t_j < t_i} (1 - \hat{h}_j) \qquad (5)$$

This means that expressions (4) and (5) can be used to estimate the reliability function $R(t)$, and thereby the battery prognostic function $\mathcal{B}(t; t_0, \mathcal{V})$ using (3).

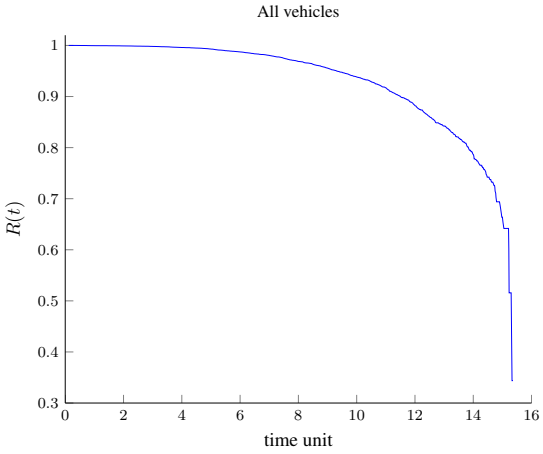


Figure 3. Reliability function estimate for the full data set.

4.2 Battery Degradation Characteristics

As described in Section 2, the battery failure rate is significantly different in different vehicles, e.g., a long-haulage vehicle with a large battery, kitchen equipment, and driving in cold weather may experience significantly different battery degradation behavior than a city distribution truck. To illustrate this, Figure 3 shows the Kaplan-Meier estimate (5) for the full data set. This estimate would be useful if it were true that the battery degradation is equal for all vehicles, no matter the vehicle configuration or usage. Figure 4 shows corresponding estimates for classes of vehicles with different battery mount position (a) and different temperature statistics (b). The blue curve in Figure 4 a/b corresponds to the full set of vehicles in the database, as shown in Figure 3. Since the estimated reliability functions significantly deviate from the blue curve for different sets of vehicles it is clear that battery degradation characteristics significantly depends on which set of vehicles that are investigated. Further, this means that there is a need to estimate the battery reliability function for each specific vehicle and (5) can not be directly applied.

4.3 Reliability Function Estimation for a specific Vehicle $\mathcal{V}$

From the discussion above, the 291 variables that are stored for each vehicle and describe vehicle configuration and

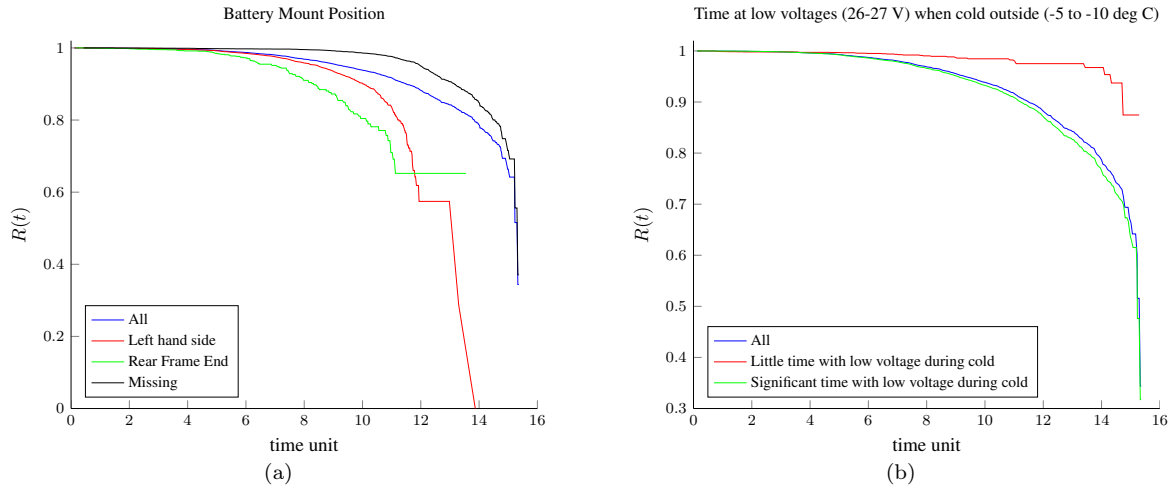


Figure 4. Reliability function estimation for different battery positions (a) and vehicles different with different amount of time with low battery voltage during cold ambient temperatures (b).

usage need to be taken into account when estimating the reliability function. As said in Section 2, the basic idea of the approach 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. The approach is based on Random Survival Forests (Ishwaran et al., 2008; Ishwaran and Kogalur, 2010). Random survival forest is a survival analysis extension of Random Forests (Breiman, 2001) which is a tree-based classifier (Breiman et al., 1984) extended with bootstrap aggregation (Breiman, 1996) techniques.

There are 291 variables stored for each vehicle and the data includes 17 histograms. The treatment of histogram variables is not described here in detail, the procedure can be found in (Frisk et al., 2014), but the key step is that 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 data size is reduced, the procedure is described in (Frisk et al., 2014), to 113 or 116 variables depending on how accumulative variables are handled. The treatment of accumulative variables is further discussed in Section 5. This corresponds to a slightly modified version of the approach from (Frisk et al., 2014) and a flowchart in Figure 5 outlines the procedure. The procedure to build a random survival forest model here then comprises the steps

(1) Collect the data
(2) Handle histogram variables as in (Frisk et al., 2014)
(3) Reduce data size as in (Frisk et al., 2014)
(4) Build the model, see Section 6.1

When built, the random survival forest model is able to predict a reliability function $R^{\mathcal{V}}(t)$, and also $\mathcal{B}(t; t_0, \mathcal{V})$ based on vehicle data $\mathcal{V}$. The experiments are conducted in R (R Core Team, 2014) using the package Random Forests for Survival, Regression and Classification (Ishwaran and Kogalur, 2013).

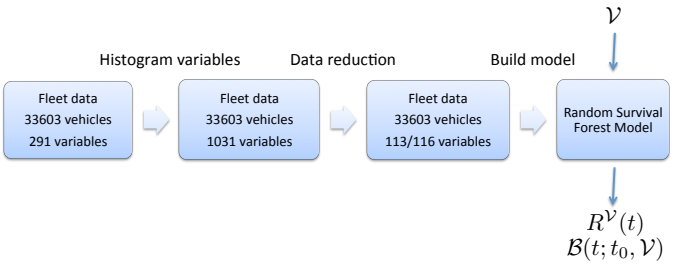


Figure 5. Flowchart of the reliability function estimation procedure.

5. ACCUMULATIVE VARIABLES

This section describes how variables, and especially the accumulative variables, are prepared for the classification step in RSF. The basic principle is the property that a vehicle operated similarly over time should remain in the same class. In that case vehicles at different age with similar operation characteristics will be collected in a class and a reliability function estimate for that type of operation can be computed. Accumulative variables do not have this property and need to be modified.

In the data there are 17 histograms with bin-values with units time, distance, count, and fuel volume which are all accumulative entities. All these 17 histograms are normalized such that the sum of bin-values equals 1. The variable `Age` is removed from the set of classification variables but the information of vehicle age is used for estimating the reliability function. The variable `Distance` is an accumulative variable and is replaced by distance per day `MilagePerDay`.

Figure 6 shows the correlation of the most correlated variables with age. There is a strong correlation between `Create month` and `Age` and this is caused by the way data has been collected. Data from all vehicles up to a certain age has been exported at one single date. This date subtracted with create month will be an upper bound and often also a good estimate of vehicle age. Even though vehicle age can be estimated based on `Create month`, it is not accumulated over time and is therefore considered as a classification variable. By using `Create month` in the classification, seasonal and component quality variations for different months can affect reliability function estimation.

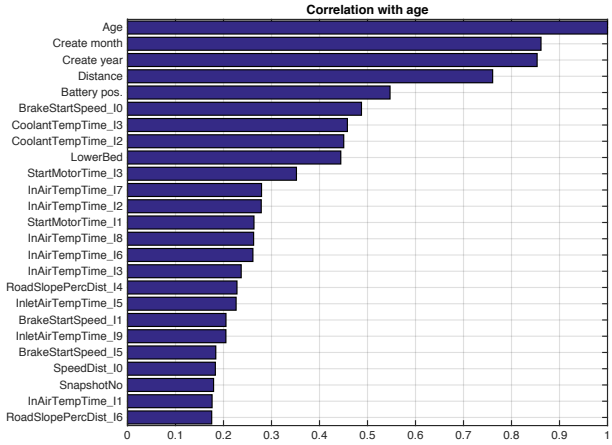


Figure 6. Most correlated variables with age.

Another non-accumulative variable that has been removed is snapshot number `SnapshotNo` which is a serial number assigned to each vehicle data download in chronologically increasing numbers. It has been removed because the data collection method has introduced a correlation with battery failure in the following way. For vehicles with working batteries the snapshot is taken within a maintenance interval from the date of data collection. The snapshot for vehicles with battery problems is taken at time of battery failure. Hence a low snapshot number will correlate with battery problem. However considering a true situation the snapshot number will not be correlated with battery failure and therefore should not influence the reliability function estimation.

To conclude this section a summary of the difference of the classification variables used here and used in (Frisk et al., 2014) will be given. In (Frisk et al., 2014) the histograms where normalized so the change from that work to this work is

- `Age`, `SnapshotNo`, and `ChassiNo` have been removed
- `Distance` has been replaced by `MileagePerDay`

6. CASE STUDY: BATTERY PROGNOSTICS

The objective now is to use the methodology described in Section 4 to estimate battery prognostic functions and analyzing the results based on the discussions in Section 5. To avoid revealing sensitive information, presented data is normalized. A fundamental property when analyzing the data set is that there is no ground truth, i.e., the true battery degradation behaviors for the set of vehicles are not known. Therefore, a discussion why the predicted battery prognostic functions are reasonable are included at the end of the section.

6.1 Building the models from data

First, the approach form Section 4 is used to build the random survival forest models. Two models will be built, one using accumulative variables, denoted $\mathcal{M}_{\text{acc}}$, and one who do not which is denoted $\mathcal{M}_{\text{no acc}}$. To build the models, the software package (Ishwaran and Kogalur, 2013) is used and 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

The discussion on these variables requires some detailed knowledge on random survival forest, and a reader not familiar with the technique can skip this part. Selection of these parameters is important for the result and in (Frisk et al., 2014), an investigation on a proper size of the terminal nodes were conducted and is also here chosen to be of size 200. Thus, each class in each tree in the grown forest consists of at least 200 vehicles. Further, when building each tree in the forest, at each node a random procedure is used to select the next split variable. A rule of thumb (Ishwaran and Kogalur, 2013) is to randomly try $\sqrt{n}$ variables where n is the total number of variables. Here, n is 113 and 116 respectively for the cases with and without accumulative variables. Thus, the number of randomly chosen variables to try at each split would ≈ 11, here 13 variables is used. To find a split value for the corresponding node in the tree classifier, a randomized procedure could be used to speed up the process. But here instead a complete search is performed.

The fourth and last of the key parameters in the algorithm is the number of trees to grow in the forest. Analysis of the prediction error rate is useful for selecting number of trees. 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 batteries in two random vehicles. Formally, the error rate is $1 - C$ where C is Harrell's concordance index (Harrell et al., 1982). Figure 7 plots the error rate as a function of number of trees for both models, i.e., the model with accumulative variables $\mathcal{M}_{\text{acc}}$ and without accumulative variables $\mathcal{M}_{\text{no acc}}$. From this plot it is clear

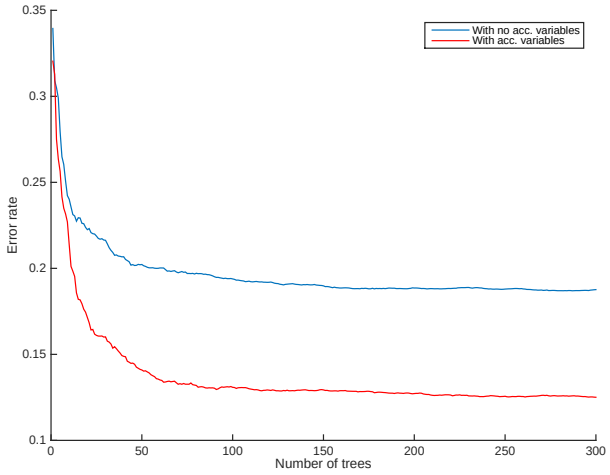


Figure 7. Error rate as a function of number of trees for two models, one with and one without accumulative variables.

that, based on the error rate, there is no reason to grow more than about 200-300 trees in the forest. Here, 300 classification trees are grown in the forest for both models. Another observation is that the model $\mathcal{M}_{\text{acc}}$ obtains a significantly lower error rate than $\mathcal{M}_{\text{no acc}}$. This is to be expected since the variables in $\mathcal{M}_{\text{acc}}$ is a superset of the variables in $\mathcal{M}_{\text{no acc}}$. However, this should not immediately be interpreted as that $\mathcal{M}_{\text{acc}}$ is a more accurate model for the reasons outlined in Sections 3 and 5.

With the parameter values chosen, building the random survival forest models $\mathcal{M}_{\text{acc}}$ and $\mathcal{M}_{\text{no acc}}$ takes about 62

minutes each. 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 takes about 25 seconds.

6.2 Variable Importance Analysis

Figure 2 and Figure 8 show variable importance for the models $\mathcal{M}_{\text{no acc}}$ and $\mathcal{M}_{\text{acc}}$ respectively. A comparison of these figures shows how variable importance get influenced by different treatments of the accumulative variables.

First remember that the variables `Age`, `SnapshotNo`, `ChassiNo`, and `Distance` are not included in $\mathcal{M}_{\text{no acc}}$ and hence not included in Figure 8. A comparison of the remaining variables show that `Create month` is the most important variable in both cases. It is interesting to note that in Figure 8 `Create year` and `Battery pos.` are higher ranked than in Figure 2. It is also interesting to note that the 3 most important variables in Figure 8 are the 3 most correlated variables with age according to Figure 6. The variable `Age` is important according to Figure 2 and in the case when `Age` is not used the most correlated variables `Create month`, `Create year`, and `Battery pos.` can provide information of age.

There are some battery related variables that are important in both cases such as `BattVoltTemp_I3_p2` and `BattVolt_p2`. It is also worth noting that `MilagePerDay` is higher ranked than `Distance`. Also `Country Id` is more important when accumulative variables are not used.

As a conclusion `Create month`, `Create year`, and `Battery pos.` are most important variables in $\mathcal{M}_{\text{no acc}}$ but further investigations need to be done in order to understand if their importance are due to quality variations of batteries over time or seasonal conditions degrading the battery different over time or if those variables are important because of their correlation with vehicle age.

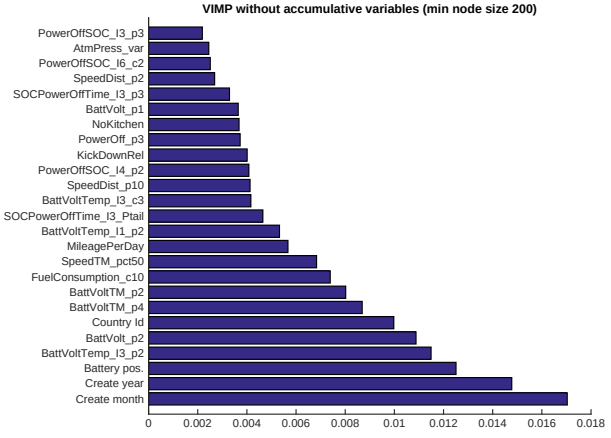


Figure 8. VIMP without accumulative variables.

6.3 Battery prognostics

Given the two estimated models, $\mathcal{M}_{\text{acc}}$ and $\mathcal{M}_{\text{no acc}}$, we can now estimate the battery prognostic function (2) given vehicle data $\mathcal{V}$. It is clear that there is an age component to battery degradation, either directly or indirectly for example due to longer exposure to low temperatures or vibrations. The objective of the prognostics approach is to find a vehicle individual maintenance plan that is not based on age or distance. A set of vehicles to analyze further is therefore needed. The set of vehicles to predict and further analyze is selected such that the vehicles have similar age and distance properties, i.e., vehicles that with a fixed maintenance schedule based on age or distance should have similar time for next maintenance.

Now, let $\mathcal{W}_0$ be the set of vehicles in the original database with *no battery problems* and let the functions age($\mathcal{V}$) and distance($\mathcal{V}$) give the age and distance traveled respectively for a given vehicle $\mathcal{V}$. Then, a set of vehicles with age about 5 time units is extracted as

$$\mathcal{W}_1 = \{\mathcal{V}; \mathcal{V} \in \mathcal{W}_0 \text{ and } 4.85 \leq \text{age}(\mathcal{V}) \leq 5.15\}$$

Let m be the mean distance traveled among the vehicles in $\mathcal{W}_1$, then the final set of vehicles $\mathcal{W}$ are the vehicles in $\mathcal{W}_1$ with distance traveled within 10% of the mean distance, i.e.,

$$\mathcal{W} = \{\mathcal{V}; \mathcal{V} \in \mathcal{W}_1 \text{ and } 0.9\,m \leq \text{distance}(\mathcal{V}) \leq 1.1\,m\}$$

The resulting set $\mathcal{W} \subseteq \mathcal{W}_0$ consists of 144 vehicles with no reported battery problems, similar age, and traveled distance.

Figure 9 shows the predicted battery prognostic function $\mathcal{B}(t; t_0, \mathcal{V})$ for the 144 vehicles in $\mathcal{W}$ using both models. From Figure 9 it is evident that there is a wide spread among the battery prognoses and this is true regardless if the accumulative variable are included in the model or not. Let T_{90} denote the maximum time where we have more than 90% confidence that the battery will be operational, i.e.,

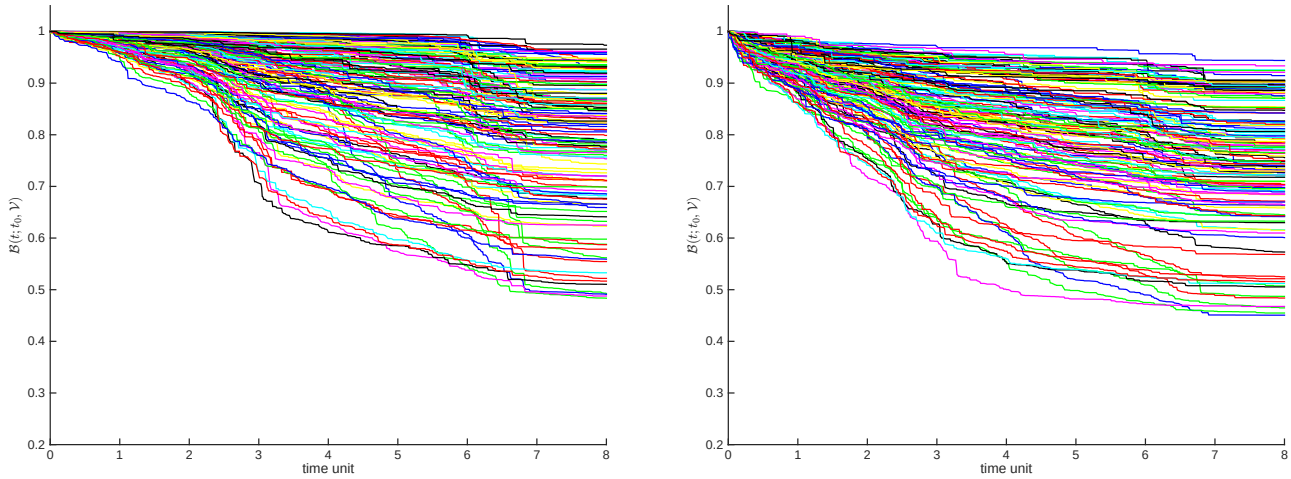
$$T_{90} = \max_t \ \mathcal{B}(t; t_0, \mathcal{V}) \geq 0.9$$

In Figure 9(a), with predictions using the model $\mathcal{M}_{\text{no acc}}$, it is clear that the T_{90} time varies from about 1.3 time units for the vehicle with the worst prognosis and more than 8 time units for the vehicle with the best prognosis. A similar situation occurs when predicting using also accumulative variables. This is interesting since the models predict what Figure 4 showed, that vehicle configuration and usage significantly influences the battery prognosis.

To further analyze the results of Figure 9, we identify the vehicles with the most extreme battery prognoses. Let $\mathcal{V}_1$ and $\mathcal{V}_2$ denote the vehicles in $\mathcal{W}$ with best and worst prognosis using the model $\mathcal{M}_{\text{no acc}}$, i.e., the functions in Figure 9(a) that are highest and lowest respectively. Figure 10 shows the battery prognostics function for vehicles $\mathcal{V}_1$ and $\mathcal{V}_2$ where it is evident that the estimated prognoses for these two vehicles are significantly different and needs different maintenance plans. Identifying the vehicles with best and worst prognosis using model $\mathcal{M}_{\text{acc}}$ instead of model $\mathcal{M}_{\text{no acc}}$ shows a similar difference. It turns out that the vehicle with best predicted prognosis is the same for both models, but the vehicle with worst predicted prognosis is different with the two models. Denote the vehicle with worst predicted prognosis using model $\mathcal{M}_{\text{acc}}$ with $\mathcal{V}_3$. Thus, three vehicles with extreme differences in battery prognosis have been identified as:

- $\mathcal{V}_1$ - best predicted prognosis using both models $\mathcal{M}_{\text{no acc}}$ and $\mathcal{M}_{\text{acc}}$.
- $\mathcal{V}_2$ - worst predicted prognosis using model $\mathcal{M}_{\text{no acc}}$.
- $\mathcal{V}_3$ - worst predicted prognosis using model $\mathcal{M}_{\text{acc}}$.

Table 1 shows some detailed data from vehicles $\mathcal{V}_1$, $\mathcal{V}_2$, and $\mathcal{V}_3$. In the table, not all variables are included, only the 25 most important according to Figure 8 and also distance and age of the vehicles. The data has been normalized such



(a) $\mathcal{M}_{\text{no acc}}$ (b) $\mathcal{M}_{\text{acc}}$

Figure 9. Estimated battery prognostic function of the 144 vehicles in $\mathcal{W}$ using both random survival forest models.

Table 1. Normalized data for the three vehicles $\mathcal{V}_1$, $\mathcal{V}_2$, and $\mathcal{V}_3$ which correspond to the vehicle with best prognosis, the worst prognosis according to model $\mathcal{M}_{\text{no acc}}$, and the worst prognosis according to model $\mathcal{M}_{\text{acc}}$.

Variable	**Vehicle** $\mathcal{V}_1$	**Vehicle** $\mathcal{V}_2$	**Vehicle** $\mathcal{V}_3$
Distance (norm)	1.00	0.93	0.87
Age (norm)	1.00	1.00	1.02
Create month	2011092	2011101	2010122
Create year	2011	2011	2010
Battery pos.	Left hand side	Left hand side	Left hand side
BattVoltTemp_I3_p2 (norm)	1.00	10.20	24.48
BattVolt_p2 (norm)	1.00	10.61	26.64
Country Id	2	1	0
BattVoltTM_p4 (norm)	1.00	5.89	12.34
BattVoltTM_p2 (norm)	1.00	23.25	84.70
FuelConsumption_c10 (norm)	1.00	0.79	0.52
SpeedTM_pct50 (norm)	1.00	1.09	1.36
MileagePerDay (norm)	1.00	0.93	0.86
BattVoltTemp_I1_p2 (norm)	1.00	9.69	28.99
SOCPowerOffTime_I3_Ptail	-	1.0	1.0
BattVoltTemp_I3_c3 (norm)	1.00	11.78	34.94
SpeedDist_p10 (norm)	1.00	0.66	0.35
PowerOffSOC_I4_p2 (norm)	1.00	16.91	28.64
KickDownRel	-	1.0	1.0
PowerOff_p3 (norm)	1.00	0.64	1.12
NoKitchen	yes	no	no
BattVolt_p1 (norm)	1.00	11.65	31.22
SOCPowerOffTime_I3_p3	-	1.0	1.0
SpeedDist_p2 (norm)	1.00	0.89	0.76
PowerOffSOC_I6_c2 (norm)	1.00	1.11	1.11
AtmPress_var (norm)	1.00	3.48	1.37
PowerOffSOC_I3_p3	-	1.0	1.0

that the vehicle $\mathcal{V}_1$ with best battery prognosis has variable values 1 except for cases when the unnormalized value is 0 or if it is a categorical variable such as `Country_Id`. The variables for vehicles $\mathcal{V}_2$ and $\mathcal{V}_3$ that are significantly different from vehicle $\mathcal{V}_1$ are mainly variables related to low battery voltage and sometimes at specified temperature intervals. For example `BattVoltTemp_I3_p2` is the relative time spent in 10-25°C with 26-27 V where the normal voltage is up to 30 V. This is a good temperature for the battery and low voltages are not expected to be common in this temperature interval. The batteries in $\mathcal{V}_2$ and $\mathcal{V}_3$ had this condition 10.20 and 24.48 times as frequently as the battery with good prognosis in vehicle $\mathcal{V}_1$. Further, comparing vehicles $\mathcal{V}_2$ and $\mathcal{V}_3$ it is clear that the main differences compared to vehicle $\mathcal{V}_1$ are in the same variables. As noted in the beginning of this section, there is no ground truth available but one conclusion so far is that for vehicles with similar age and distance, the most important differences are related to low battery voltage independent of the treatment of accumulated variables which is consistent with engineering experience. Thus, the procedure managed to automatically produce relevant battery prognostic functions, separating vehicles that otherwise would have had the same maintenance schedule.

To further analyze the effect of using accumulative variables, Figure 11 shows the predicted battery prognostic function for vehicles $\mathcal{V}_i$ using both models with and without the accumulative variables. Let $\mathcal{B}_i^{\text{no acc}}$ and $\mathcal{B}_i^{\text{acc}}$ correspond to the battery prognostic function for vehicle $\mathcal{V}_i$ using models $\mathcal{M}_{\text{no acc}}$ and $\mathcal{M}_{\text{acc}}$ respectively. From the figure it is clear that choice of model has significant influence on the

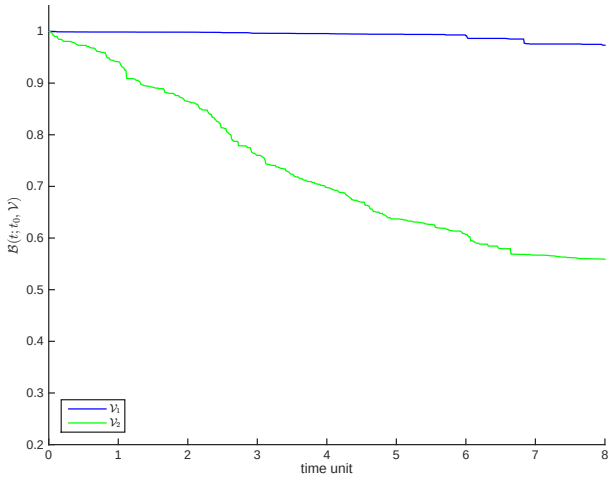


Figure 10. Battery prognostics function for vehicles in $\mathcal{W}$ with best and worst predicted prognoses using model $\mathcal{M}_{\text{no acc}}$.

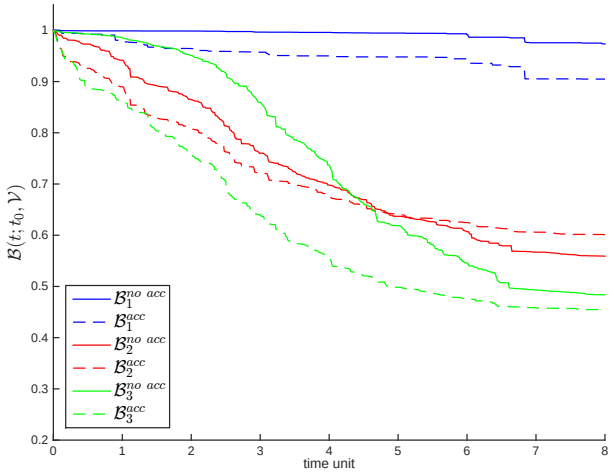


Figure 11. Battery prognostic estimates for functions for vehicles $\mathcal{V}_1$, $\mathcal{V}_2$, and $\mathcal{V}_3$ using models $\mathcal{M}_{\text{no acc}}$ and $\mathcal{M}_{\text{acc}}$.

estimate of the prognostic function, compare for example $\mathcal{B}_3^{\text{no acc}}$ with $\mathcal{B}_3^{\text{acc}}$. This indicates that choice of model is important. In this case, the model $\mathcal{M}_{\text{acc}}$ produces mostly conservative prognostic estimates, i.e., the dashed lines lies below the corresponding solid lines in the figure for most of the prediction horizon. However, this is not generally true. Consider a vehicle that is old, but has been used in a way that is not damaging for the battery, for example operational in mainly $+20°$, low speeds with low levels of vibrations etc. That vehicle would, in the $\mathcal{M}_{\text{acc}}$ model, be associated in the same class as other equally old vehicles. This is due to that the `age` variable is so important in the classifier as shown in Figure 2. This would not be true for the model $\mathcal{M}_{\text{no acc}}$ and the vehicle would be associated with vehicles with similar usage profile and configuration. This is a key difference between the different models and the main reason why $\mathcal{M}_{\text{no acc}}$ is preferable to $\mathcal{M}_{\text{acc}}$.

7. 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.

A contribution is a case study utilizing the data-driven approach random survival forests to compute probabilistic reliability properties for a battery in a specific vehicle. The case study is based on vehicle data from 33603 vehicles. A main contribution of the paper is to analyze and compare the difference in the results obtained with or without including accumulative variables in the classification approach. A first conclusion is that if the accumulative and most important variable `Age` is removed, the three most strongly correlated variables with age become most important. A second conclusion is that the estimated battery prognostic function is significantly changed if accumulative variables are omitted. A third conclusion is that when looking at vehicles with the same age and driven distance but with significant different battery predictions the main differences are in variables related to battery properties such as relative time with low voltage or relative time with a certain voltage at a specified temperature interval.

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., and Olshen, R.A. (1984). *Classification and regression trees*. CRC press.

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

Frisk, E., Krysander, M., and Larsson, E. (2014). Data-driven lead-acid battery prognostics using random survival forests. In *Proceedings of the Annual Conference of The Prognostics and Health Management Society*. Fort Worth, Texas, USA.

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

Heng, A., Zhang, S., Tan, A.C., and Mathew, J. (2009). Rotating machinery prognostics: State of the art, challenges and opportunities. *Mechanical Systems and Signal Processing*, 23(3), 724–739.

Ishwaran, H. and Kogalur, U. (2013). *Random Forests for Survival, Regression and Classification (RF-SRC)*. URL `http://cran.r-project.org/web/packages/randomForestSRC/`. R package version 1.4.

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

Ishwaran, H., Kogalur, U.B., Blackstone, E.H., and 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. and Köttig, F. (2014). 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*. R Foundation for Statistical Computing, Vienna, Austria. URL `http://www.R-project.org/`.