Prediction of Railcar Remaining Useful Life by Multiple Data Source Fusion

Zhiguo Li and Qing He, Member, IEEE

Abstract—Nowadays, railway networks are instrumented with various wayside detectors. Such detectors, automatically identifying potential railcar component failures, are able to reduce rolling stock inspection and maintenance costs and improve railway safety. In this paper, we present a methodology to predict remaining useful life (RUL) of both wheels and trucks (bogies), by fusing data from three types of detectors, including wheel impact load detector, machine vision systems, and optical geometry detectors. A variety of new features is created from feature normalization, signal characteristics, and historical summary statistics. Missing data are handled by missForest, a Random Forests-based nonparametric missing value imputation algorithm. Several data mining techniques are implemented and compared to predict the RUL of wheels and trucks in a U.S. Class I railroad railway network. Numerical tests show that the proposed methodology can accurately predict RUL of the components of a railcar, particularly in a middle-term range.

Index Terms—Missing value imputation, predictive model, rail wayside detectors, random forests, remaining useful life, rolling stock failure.

I. INTRODUCTION AND PROBLEM STATEMENT

SYSTEM failures of rolling stock components are crucial to railway safety and efficient operations. Analysis of Federal Railroads Administration (FRA) accident data for the period 1999 to 2008 identified that the four largest US Class I railroads spent an average of about $38 million per year on equipment-caused mainline derailments [1]. Toward effective equipment inspection, the railroad industry has developed wayside detectors to monitor the conditions of railcar components. Such systems employ a variety of sensing technologies to measure force, heat, sound, geometry and so on. The technological maturity of these systems also improves from 1930s to 2000s [2], while new technologies are still under development. The North American railroad industry is increasingly moving to wayside detection to reduce rolling stock inspection and maintenance costs [3]. Detailed surveys of these and other wayside inspection technologies have been conducted in the previous studies [4]–[9]. It is anticipated that, through automated wayside detection, the railroad industry is expected to: i) reduce significant inspection efforts, ii) perform more efficient fleet maintenance focused on specific systems or components directly related to the measured poor performance, iii) decrease service disruptions and service delays with appropriate labor and materials provided for necessary repairs, and iv) provide insights into the root causes for poor or degraded performance [3].

Unplanned failures, such as derailment, cost significantly to railway operations. Prediction of railcar defects will give operators sufficient time to inspect the potential failures before they really occur. Existing practice of failure prediction and prevention is managed by various alarms in the United States. Such alarms are created based on two levels of existing regulations, defined by FRA and American Association of Railroads (AAR) respectively. For example, there are differences in allowable wheel profile tolerances between FRA and AAR rules. Table I summarizes the limits of some wheel profile parameters (shown in Fig. 1), defined by both FRA and AAR. Comparing the MV readings to existing FRA and AAR regulations will produce two levels of violations - FRA level and AAR level violations.
AAR level violations are handled at Train Termination Point, whereas FRA level violations need to be handled immediately, as soon as the inspection is carried out. Such rule violations will lead to a severe interruption of existing rail operations. Therefore, the existing railroads have developed their own rules to generate different levels of alarms before any violation occurs. Most of such rules are composed of threshold-based combined conditions and generated from railroad repair experiences. They are effective when the component is about to fail, because railroads strive to maintain a very low false alarm rate to reduce the inspection costs. In addition, some forecasting models for investigating the behavior of the railway components have been developed based on intelligent monitoring systems (e.g., [10]). However, such rules and models do not provide the values of remaining useful life (RUL) of a particular component whose measurements are close to, but still under, the thresholds. If the exact RUL of each component is known, the railroad can make better use of their resources to maintain the rolling stock in a state of good repair with the reduced number of maintenance “set-outs” and increased rail network average velocity.

In recent years, RUL prediction has drawn intensive attention, and different methods and techniques have been developed in academia and industries [11]–[14]. A recent review can be found in [14], where the authors classified the RUL prediction approaches into four major categories: experimental methods, data-driven methods, physical model-based methods, and hybrid methods. Due to the rapid development of cyber infrastructure and sensing technology, an abundance of data from service and maintenance of engineering systems are now readily available, and thus data-driven methods are developed and widely used for RUL prediction, especially in the cases that the physical models for failure mechanisms are difficult, if not impossible, to derive. Some examples of the recent development in the data-driven approaches include [15] and [16], where the nonlinear Support Vector Regression (SVR) models were trained to estimate and predict the wear level of cutting tools and RUL of rolling element bearings, respectively.

To avoid failures/faults, accurate prediction of RUL on the basis of the current equipment conditions and operation history is essential for making a decision on preventive maintenance or condition-based maintenance. RUL can be defined as a conditional random variable

$$T - t|T > t, \mathbf{X}(t)$$  \hspace{1cm} (1)

where $T$ represents the time to failure, $t$ is the current working age and $\mathbf{X}(t)$ denotes the historical operation profile up to the current time (condition monitoring data). Two major data types used in prediction of RUL are event data and the condition monitoring data [17]. Usually the event data record the information of failure/fault events, and thus failure times (i.e., time-to-failure) can be extracted from the data. The usage conditions and environmental stresses (working load, pressure, temperature and vibration, etc.) are recorded in the condition monitoring data. Being able to assess the conditions of the equipment is the basis for predicting potential failures and RUL.

Among all defect types, wheel failures cause about one half of all train derailments. Given massive amount of data collected from electronic wayside detectors, railcar failure prediction has recently attracted great attention. In the literature, there are two general methods to predict railcar component failures: (1) multi-body simulation (MBS) of railway vehicle dynamics, and (2) statistical learning and data mining based methods. In the first category of “vehicle dynamics”, Braghin et al. [18] proposed a wheel wear prediction model which is composed of simulation of vehicle dynamics, local wheel-rail contact model, and local wear model. The proposed model has been validated through comparison with full-scale experimental tests carried out on a single mounted wheelset under laboratory conditions. Pombo et al. [19] developed an estimation tool consisting of the use of a sequence of pre- and post-processing packages, in which the presented methodologies are implemented, and interfaced with commercial multi-body software. The disadvantage of this model is that such MBS based models require complex computation to test even one scenario. So it is difficult to implement these models on a variety of railcar equipment types and rail conditions in real-time. In the second category of “data mining”, Yang and Létourneau [20] first applied data mining based methods for prognostics of train wheel failures using measurements from WILD detectors. They proposed a Multiple Classifier System (MCS) capable of predicting 97% of wheel failures while maintaining a reasonable false alert rate (8%). However, this study just predicts the occurrence of wheel failures without calculating the occurrence times, and only a single detector type was considered in modeling. Hajibabai et al. [21] developed a logistic regression model to classify wheel failures based on WILD and Wheel Profile Detector (WPD). They claimed that the classification accuracy is 90% with 10% false alarm rate. In that study, only wheels within 30 days of train stop were classified as “bad wheels”. So their proposed model cannot predict the potential bad wheels with time-to-failure values greater than 30 days. Also, neither truck measurements nor truck failures is taken into account in that work.

As one can see, previous studies have shown that wayside detectors can successfully predict the occurrence of railcar component failures. However, there are still some significant challenges unaddressed:

1. Underlying sophisticated correlations between different component failures are still not disclosed. According to the measuring components, traditional wayside detection data can be classified into a number of different categories: trucks, axles, wheels, and bearings. Among all railcar components, wheels, journal bearings, and truck components represent the greatest derailment risk and result in approximately 78% of all equipment-caused derailment costs [1]. However, those components are not isolated from each other. It is well-known in the field that a defective wheel on one side can easily damage a good wheel on the other side. So the wheels are usually replaced in a set of two in the repair workshop, even if there is only one defective wheel found. Furthermore, defective wheels not only result in “hot” journal bearings, but also lead to truck hunting eventually. Also, a misaligned truck can cause asymmetric wheel wear, which dramatically shorten
the life of a new set of wheels. Therefore, some different types of failures are strongly correlated to each other. It is more reasonable to combine them together for prediction analytics than to treat them separately.

2. Large measurement variations are found in the datasets. The measurement variations are referred to as discrepancies of the readings on the same component due to both internal factors (e.g., axle load, and brake status) and external factors (e.g., temperature, weather and humidity). Under these various circumstances, simple threshold based rules can generate lots of false alarms. For example, rim thickness measurements from MV are very precise and hence very volatile. During winter, snow on the wheel will affect the accuracy of the rim thickness readings over time without any underlying causes or wheel dimension changes.

3. The entire wayside detection system generates sparse, heterogeneous and irregular data due to a limited number of detectors in the network. Each installation of wayside detection systems can cost up to one million dollars [2]. Therefore, wayside detectors are typically sparsely located (except for hot bearing detectors). Moreover, since locations of the wayside detector are not evenly distributed in the network, the number of readings varies a lot for different trains. Some trains which commute on a detector monitored line segment generate a large amount of data, while others only have very few readings. If we track one railcar travelling through the network, its detection data are irregular and heterogeneous. The interval between two readings of the same component could be less than one day or larger than six months. These irregular measurements may lead to a large amount of missing values in the measurement data. How to handle the missing data becomes very essential for the data analysis.

To solve these problems, this paper aims to predict RUL of both railcar wheels and trucks, from three types of wayside detectors, namely Wheel Impact Load Detector (WILD), Machine Vision (MV) systems, and Optical Geometry Detector (OGD). In this study, we will utilize the data-driven approaches for the RUL prediction. In summary, the contributions of this paper lie as follows:

- fusing measurements from multiple types of wayside detectors including WILD, MV and OGD;
- handling irregular and heterogeneous measurements with large variations;
- imputing missing data due to a limited number of readings; and
- predicting the exact RUL of both trucks and wheels.

The remaining of the paper is organized as follows. Section II briefly reviews the wayside detectors and the measurement data. In Section III, the methodology developed for RUL prediction is illustrated, and the Random Forests based missing value imputation method and predictive models are introduced. After that, we discuss the data preprocessing and missing value imputation in Section IV. The experimental results for RUL prediction using the state-of-the-art regression models are presented and compared in Section V, and the results are discussed in Section VI. Finally, we summarize the study and discuss the future work directions in Section VII.

### II. DETECTORS AND DATA OVERVIEW

In this section, we will introduce the three types of wayside detectors used for the RUL prediction as well as the detector measurement data. The key measurement features from MV, OGD, and WILD are presented in Table II.

<table>
<thead>
<tr>
<th>Detector type</th>
<th>Feature</th>
</tr>
</thead>
<tbody>
<tr>
<td>WILD</td>
<td>Wheel flange height</td>
</tr>
<tr>
<td></td>
<td>Wheel flange thickness</td>
</tr>
<tr>
<td></td>
<td>Wheel rim thickness</td>
</tr>
<tr>
<td>MV</td>
<td>Wheel diameter</td>
</tr>
<tr>
<td></td>
<td>Wheel tread hollow</td>
</tr>
<tr>
<td></td>
<td>Brake shoe upper thickness</td>
</tr>
<tr>
<td></td>
<td>Brake shoe lower thickness</td>
</tr>
<tr>
<td>OGD</td>
<td>Truck hunting peak-to-peak (PTP)</td>
</tr>
<tr>
<td></td>
<td>Truck hunting amplitude</td>
</tr>
<tr>
<td></td>
<td>Truck inter-axle misalignment (IAM)</td>
</tr>
<tr>
<td></td>
<td>Truck rotation measurement</td>
</tr>
<tr>
<td></td>
<td>Truck tracking error</td>
</tr>
<tr>
<td></td>
<td>Truck shift measurement</td>
</tr>
<tr>
<td>WILD</td>
<td>Wheel average downward load reading</td>
</tr>
<tr>
<td></td>
<td>Wheel peak downward load reading</td>
</tr>
<tr>
<td></td>
<td>Wheel average lateral load reading</td>
</tr>
<tr>
<td></td>
<td>Wheel peak lateral load reading</td>
</tr>
<tr>
<td></td>
<td>Difference between peak and average</td>
</tr>
<tr>
<td></td>
<td>downward load reading</td>
</tr>
<tr>
<td></td>
<td>Truck hunting index</td>
</tr>
</tbody>
</table>

WILD uses strain-gauge-based technologies to measure the performance of a railcar with wheels and trucks in a dynamic mode. WILD is built into the track to detect defective wheels, which can be distinguished from normal wheels by their high dynamic weights. WILD weighs each wheel on the train several times when the wheel passes by a detector in a certain distance. The system reports the features to describe dynamic impact load at the wheel level, such as vertical KIPS (a thousand pound of force) and lateral KIPS. In particular, whenever a wheel’s impact reaches 140kips or more, the train is immediately stopped and then required for the wheel replacement in the nearest siding [20]. MV technology employs real-time video imaging to capture parameter measurements of rolling stock components, such as wheels, brakes, springs, draft gears and so on. The key wheel features captured by MV include flange height, flange thickness, rim thickness, wheel diameter, etc. Some of the limits have been summarized in Table I. OGD is a laser-based monitoring system that measures performance of a railcar’s axle and wheel suspension assembly, commonly known as the “truck.” As a train passes by the OGD on tangent track, the system identifies trucks that have become warped or misaligned, and reports tracking error, truck rotation, inter-axle misalignment and shift, shown as Fig. 2. Also, OGD measures the amplitude and frequency of the waveforms developed by
the axles passing by the detector, which can be treated as useful information for identifying truck hunting.

About 500 GB of raw wayside detection data were collected in a US Class I rail network in a period of 27 months, from January 2010 to March 2012. After initial data exploration, we found that the features from wheels and trucks are correlated with each other. For example, flange thickness is negative correlated (the correlation coefficient is \(-0.52\)) with tracking errors of its truck readings. According to the measuring components, the datasets can be divided into five different levels: i) train level, ii) equipment (railcar) level, iii) truck level, iv) axle level, and v) wheel and bearing level. On average, one train level record will generate 79 equipment records, 175 truck records, 413 axle records, and 826 wheel and bearing records, respectively. Therefore, one may predict potential failures for each wheel, axle, truck, railcar, or even the whole train. It is critical to conduct failure prediction on a particular level of components, in order to maximize the benefit and performance of the approach. The best trade off maximizes usefulness of information for the target organization and the performance of the approach (false errors, rate of detection, etc.) [20]. This study selects the truck level for prediction. Therefore, the model will predict when a given truck is likely to suffer a truck/wheel failure or not. Individual wheel level predictions are not required due to the fact that defective wheels can be easily identified given an exact truck ID. Therefore, wheel and axle data are eliminated from the analysis, and thus the size of entire data set is reduced dramatically.

In addition to detection data, bad order data and teardown data are also collected for failure validation. When potential major defects are identified, bad order data consisting of order dates and types are generated either by wayside detectors or by visual inspections at a yard or terminal. Once a bad order is issued, the car is scheduled to be separated from the train and further checked or repaired in the workshop, where repair actions and reasons are recorded in the teardown data. Such data can be used to validate if the bad order is generated by false alarms and what subsystems/components have been repaired or replaced exactly.

III. METHODOLOGY

In this section, we will discuss the methodology used for predicting the RUL of railcars. Fig. 3 illustrates the flow chart of the proposed method.

The first step in the method is integrating the data sources and then extracting the data subset of bad trucks/wheels. The existence of a large amount of missing values in the data requires an appropriate missing value imputation approach, because statistical or data mining algorithms often depend on a complete data set. Here we select to use missForest [24], a Random Forests model based nonparametric method which can deal with missing values in different variable types (i.e., continuous and categorical variables). The details and advantages of missForest and Random Forests methods will be briefly described in the following subsections.

Next, we create three types of new features from the original ones through feature normalization, detector signal characteristics extraction, and summary statistics generation. We need to normalize some measurement features because these measurements vary in a large range with different train speeds and truck weights, especially for WILD. Once the normalized features are generated, the corresponding original ones are removed from the feature set. It is common that there is a large amount of noises in the collected detector signals, so the slopes and smoothing trends of the measurement signals may represent the signal characteristics better and thus can serve as important predictors for RUL. Moreover, to investigate the effects of the detector signals in a preceding time period (e.g., 90 days), we also need to calculate the summary statistics for the original signals. These summary statistics, for example, include average, median, maximum, minimum, and the 90th percentile of a measurement feature.

With the whole feature set consisting of original signals, normalized signals, extracted signal characteristics, and summary statistics, the predictive model will be built for RUL prediction. In this study, we fit several popular statistical and machine learning regression models to the training data and...
Given accuracy. A detailed description of RF can be found in [23]. RF is an ensemble learning method which maintains the advantages of decision trees while can increase prediction accuracy and outperform the other models used in this study. The details of the experiments will be discussed in Section V. The Out-Of-Bag (OOB) estimate of error rates can be obtained based on the training data as follows:

1. Each bootstrap is split into InBag data and OOB data;
2. For each bootstrap sample, predict the OOB data using the tree grown with InBag data;
3. Finally aggregate the predictions for the OOB data from \( n_{\text{tree}} \) trees. Calculate error rates using the observed and predicted values of the OOB data, and call it the OOB estimate of error rates.

B. MissForest

MissForest is a Random Forests based nonparametric imputation method, which can effectively handle mixed-type data and model interactive and nonlinear effects of the features. MissForest is an iterative imputation scheme in that the method trains a RF model on observed values first, predicts the missing values next, and then proceeds iteratively. Readers are referred to [24] for more details of the approach and experiment results. Here we briefly review the algorithm.

Given a \( n \times p \) dimensional matrix \( X = [X_1, X_2, \ldots, X_p] \), where \( n \) is the number of cases (records) and \( p \) stands for the number of predictor variables. First, we make an initial guess for the missing values in \( X \) using mean imputation or another imputation method. Then the variables \( X_j, j = 1, 2, \ldots, p \), are sorted in the ascending order of the missing value proportion. Each variable \( X_j \) can be divided into \( X_j^{\text{mis}} \) and \( X_j^{\text{obs}} \), where \( \text{mis} \) stands for missing values and \( \text{obs} \) stands for observed values, and the row indexes of the missing values in \( X_j \) is \( I^{\text{mis}} \), and the other features \( X(-j) \), where \( (-j) \) means the \( j \)th variable is excluded from \( X \), can be separated into \( X^{\text{mis}}(-j) \) and \( X^{\text{obs}}(-j) \) correspondingly with \( X^{\text{mis}}(-j) = X(-j)[I^{\text{mis}}, \cdot] \). For each feature \( X_j \), a random forest is fit to the data with response \( X_j^{\text{obs}} \) and predictors \( X^{\text{obs}}(-j) \), then, the missing values are imputed by the predicted missing values of \( X^{\text{mis}}(-j) \) from the trained random forest applied to \( X^{\text{mis}}(-j) \). This repeated imputation procedure will be stopped when the stopping criterion is met.

Compared to other prevalent imputation methods, like \( k \) nearest neighbors (KNNimpute) [25], saturated multinomial model [26], and multivariate imputation by chained equations (MICE) [27], MissForest does not need to make distributional assumptions of the data or feature subsets. Inherited from compare their performance using the test data. At the first place, regression models are built using the full feature set in the training data. To avoid model overfitting which leads to poor predictive performance (i.e., large prediction errors) on the test data, statistically significant features are then determined using some feature selection approaches. With the significant features, the final models will be generated by re-fitting the regression models to the training data, and thus be used for predicting the RUL values on the test data. It turns out that Random Forests and Quantile Regression Forests have similar regression models to the training data, and thus be used for prediction. Further, RF has fewer parameters, and is faster to train.

A. Random Forests

Random Forests (RF) is a combination of tree predictors (i.e., a collection of CART-like trees [22]) for data exploration, data analysis, and predictive modeling. Compared with decision trees, RF is an ensemble learning method which maintains the advantages of decision trees while can increase prediction accuracy. A detailed description of RF can be found in [23]. Given \( p \) features (predictor variables), the Random Forests algorithm is illustrated in Fig. 4:

1. Draw \( n_{\text{tree}} \) bootstrap samples from the training data.
2. Grow an unpruned decision tree for each sample: at each node, randomly sample \( m_{\text{try}} \) features (\( m_{\text{try}} \leq p \)) and choose the best split from the sampled features.
3. Predict new test data using each of the \( n_{\text{tree}} \) trees and then aggregate the results (i.e., take the average for regression).

From the algorithm, we can see that two types of randomness are contained in RF: each tree is grown on a different random subsample of the training data; and each node is split using a randomly chosen subset of features at that node. These result in

- Random Forests is less prone to overfitting. The testing performance of Random Forests does not decrease (due to overfitting) as the number of trees increases [23]. Hence after certain number of trees the performance tends to stay in a certain value. On the contrary, Decision Trees can overfit the training data and thus generate poor results when applied to the test data.
- RF is not very sensitive to outliers in the training data.
- Feature importance can be estimated during training, and the importance rank can be used for feature selection.
- Cross validation is unnecessary as the method generates an internal unbiased estimate of the generalization error (Out-Of-Bag estimate of error rates).
- RF usually has about the same accuracy as SVMs and neural networks, but can handle larger numbers of predictors. Further, RF has fewer parameters, and is faster to train.
RF, missForest can successfully handle mixed-type missing data imputation, and it outperforms other methods especially in the high-dimensional data with complex interactions and nonlinear relations. Furthermore, the computational efficiency of missForest is high.

C. Quantile Regression Forests

Quantile Regression Forests (QRF) is a Random Forests based regression model that can be used to estimate the conditional quantiles for high-dimensional predictor variables. While RF provides an accurate estimation of the conditional mean of the target variable in the regression, Meinshausen [28] shows that the full conditional distribution of the target can be obtained through RF, and thus the quantiles of this distribution can be inferred. The quantiles can be used to detect outliers, obtain prediction intervals for the predicted conditional mean from RF, etc.

QRF can be viewed as a marriage between quantile regression [29] and Random Forests. In general, the \( \alpha \) -quantile is defined by

\[
Q_{\alpha}(x) = \inf \{ y : F(y|X = x) \geq \alpha \}
\]

(2)

where \( X \) is a \( n \times p \) dimensional matrix containing all the predictor variables, \( Y \) is the response variable, \( F(\cdot) \) is the distribution function, and \( 0 \leq \alpha < 1 \). Estimates of the conditional quantiles \( Q_{\alpha}(x) \) are obtained by plugging the estimate of \( F(y|X = x) \) into Eq. (2). In QRF, all observations for every leaf of every tree are used to infer the distribution, while RF only takes their average.

In the context of condition-based maintenance, we may use QRF to estimate quantiles of RUL on the lower side of the distribution, e.g., the 10th percentile or the 25th percentile. Such quantiles may help railroads better adjust the conservativeness of their maintenance strategy.

IV. DATA PREPROCESSING AND MISSING VALUE IMPUTATION

In this study, we work on raw multi-detector dataset including MV, OGD, and WILD, bad order data, and tear down data collected over a rail network from January 2010 to March 2012. As the first step, we merge these three data sources to extract the bad truck/wheel data for RUL prediction. Because we aim to calculate the RUL values, only those trucks which have truck/wheel repair or replacement records were identified and stored in the integrated dataset. For a given date, as long as there is one record from the detectors (MV, OGD, or WILD) or from the bad order data validated by the teardown data, a row of observation will be generated in the dataset. Then the blank cells in the measurements from MV, OGD, or WILD in this row will be labeled as missing values.

Next, as the date when a bad order (validated by the teardown data) is generated for a wheel/truck is considered as its end-of-life or failure time, the RUL value for each record of this truck/wheel is calculated as the difference (in days) between the date of the record and the coming next bad order date. The regression models fitted in this study use RUL values as the response variable. Note that a single truck ID in a car can have multiple bad truck/wheel failures. For a truck ID, the dataset between two adjacent failure dates is called a data segment. The period, from the beginning of the records to the first failure date, is also marked as a data segment.

As stated above, a large number of missing values exist in the detector measurement data due to several aforementioned reasons. We impute the missing detector measurements by truck ID using missForest method. To assess the performance of missForest method for the railcar detector data, we randomly introduce missing values given a specified missing value proportion. That is, we first obtain a complete subset of the measurement data, and then randomly replace the entries in this subset up to a specified amount. For the missing value proportions of 0.3 to 0.7, Table III compares the performance of missForest method for the railcar detector data, we ran.

The definition of NRMSE for the continuous variables is

\[
NRMSE = \sqrt{\frac{\text{mean}((M - \bar{M})^2) / \text{var}(M)}{\frac{1}{p} \sum_j \text{var}(x_j) + \text{var}(y)}}
\]

(3)

where \( M \) is the complete data set, and \( \bar{M} \) is the imputed data set. Note mean(\( \cdot \)) and var(\( \cdot \)) are the empirical mean and variance for the missing values only. If the value of NRMSE is close to 0, the performance of the missing value imputation method is good, while the bad performance leads to a value of NRMSE around 1. From Table III, we can see that missForest outperforms MICE for all missing value proportions and shows an acceptable NRMSE level (much less than 1) even when the missing value proportion is as high as 70%. In the experiments, we remove the truck IDs with the proportion of total missing value greater than 60% to improve the prediction power.

With the complete data set, now we can generate the new features as listed in Table IV. At first, we normalize the WILD features by dividing the measurements by car speed, truck weight, and both, respectively. The original WILD measurements
will be excluded from the predictive model building. By fitting linear regression lines and cubic smoothing splines [31] to the original measurements from MV and OGD, and the normalized features from WILD, we can obtain the signal characteristics, including the signal slope and the smoothing trend. The slopes and smoothing splines are calculated in the date range from the immediately preceding failure event to the end of a data segment by each truck ID.

In this study, for a given feature $X_j$, $j = 1, 2, \ldots, p$, the cubic smoothing spline with knots at the unique values of $t_i$, where $t_i$ is the date of the $i$th record for the wheel, $i = 1, 2, \ldots, s$, where $s$ is the number of records in a data segment for a given truck ID, minimizes the penalized residual sum of squares

$$\sum_{i=1}^{s} (X_{ij} - f(t_i))^2 + \lambda \int_{t_1}^{t_s} \{f''(t)\}^2 dt \quad (4)$$

where $f(t)$ stands for all functions with two continuous derivatives, $X_{ij} = X_j(t_i)$, and $\lambda$ is a fixed penalty constant. When $\lambda$ increases, the resulting curve will be smoother. The automatic selection of smoothing parameter $\lambda$ involves the construction of cross-validation sum of squares $CV(\lambda)$ and the optimal $\lambda$ is selected over a suitable range to minimize $CV(\lambda)$. The readers are referred to [31] for details. An example plot of the fitted linear regression line and cubic smoothing spline is shown in Fig. 5. From this figure we can see that the smoother correctly describes the relationship of the target variable (normalized wheel average kips from WILD) and the record date even with the existence of an outlier.

Next, the summary statistics are calculated for each measurement in a preceding 90-day window from the date of each row in the integrated dataset, as shown in Fig. 6. Here we use a 90-day leading window based on the discussion with the engineers because this window size can provide enough lead time to perform predictive maintenance while avoid early removing of wheels/trucks.

Finally, we have the data ready for predictive model building. Due to the large number of features in this data set (i.e., all features listed in both Tables II and IV), we apply feature selection on the augmented data representation to automatically identify the significant features and remove irrelevant features.

V. EXPERIMENTS

During the period of January 2010 to March 2012, 745 bad trucks were identified and labeled by bad order codes relevant to wheels (79%) and trucks (21%). This dataset contains 21,855 records and 19 original measurement features from MV, OGD and WILD detectors as summarized in Table II. After data preprocessing, there are totally 248 features in the final dataset and these include the original MV and OGD measurements, normalized WILD measurements, signal characteristics, and summary statistics. In the following analysis, this final dataset with imputed missing values is used for comparing the regression models.

For model building and validation, the data are grouped into the training data and test data, while the training data contain about 75% truck IDs. Because the measurements for the same truck ID are not independent of each other, all data originated from a single truck ID is classified into either the training data or the test data. Random sampling for the whole failure data is not used here, since the model evaluation results with random sampling will be overly optimistic.

To select the best predictive models for RUL, we fit the training data with six popular statistic and machine learning regression models: RF, QRF, Decision Trees (DT), KNN Regression (KNN), Support Vector Regression (SVR), and the Principal Component Regression (PCR).

In the experiments, the R language and its packages are used for data processing and model building, and the values of the algorithms are set as follows:

- RF: $\text{nodelsize} = 20$, $\text{ntree} = 500$, $\text{mtry} = p/2$, where $p$ is the number of predictor variables;
- QRF: $\text{nodelsize} = 20$, $\text{ntree} = 500$;
- SVM: $\text{cross} = 10$, which means that a 10-fold cross validation is performed on the training data;
- KNN: $k = 100$, where $k$ is the number of neighbors considered;
- PCR: $\text{segments} = 10$ which are used in the cross validation;
- DT: the default control parameters are used. Please refer to the R package ‘rpart’ for details.

We use the absolute percentage error (APE) of the test data to compare the model performance. If there are $n_t$ rows in the
TABLE V
SUMMARY OF ABSOLUTE PERCENTAGE ERRORS FOR SIX REGRESSION MODELS

<table>
<thead>
<tr>
<th>Method</th>
<th>min</th>
<th>1st Quantile</th>
<th>median</th>
<th>mean</th>
<th>3rd Quantile</th>
<th>max</th>
</tr>
</thead>
<tbody>
<tr>
<td>RF</td>
<td>1.02%</td>
<td>11.85%</td>
<td>22.76%</td>
<td>28.25%</td>
<td>39.36%</td>
<td>88.03%</td>
</tr>
<tr>
<td>QRF</td>
<td>1.12%</td>
<td>11.52%</td>
<td>22.52%</td>
<td>28.42%</td>
<td>39.84%</td>
<td>102.40%</td>
</tr>
<tr>
<td>DT</td>
<td>1.07%</td>
<td>13.22%</td>
<td>25.49%</td>
<td>32.34%</td>
<td>44.10%</td>
<td>124.70%</td>
</tr>
<tr>
<td>KNN</td>
<td>1.22%</td>
<td>15.41%</td>
<td>31.50%</td>
<td>41.57%</td>
<td>56.28%</td>
<td>151.80%</td>
</tr>
<tr>
<td>SVM</td>
<td>1.12%</td>
<td>11.63%</td>
<td>24.30%</td>
<td>30.68%</td>
<td>43.54%</td>
<td>111.60%</td>
</tr>
<tr>
<td>PCR</td>
<td>1.46%</td>
<td>15.63%</td>
<td>31.36%</td>
<td>41.01%</td>
<td>54.71%</td>
<td>170.10%</td>
</tr>
</tbody>
</table>

where $\left| \cdot \right|$ is the absolute value operator, RUL stands for the actual RUL value, and RUL$^{\text{pred}}$ is the predicted RUL from a regression model. Based on (5), we can see that the small RUL values close to 0 can create very large APE values and thus generate a very skewed distribution. In practice, there is no justification for preferring a high precision for small values of RUL. Therefore, those records with RUL values not greater than 30 days are removed from the data. To eliminate the effects of extreme predicted values, the 2.5% trimmed APE (both the lowest 2.5% extreme values and the largest 2.5% extreme values are excluded) are calculated and summarized in Table V.

One important advantage of RF and QRF models is that the importance of the features can be calculated and used for feature selection. There are two types of feature importance measures [35]: The first importance measure for each feature (i.e., predictor variable), mean increase in Mean Squared Error (MSE) or mean decrease in prediction accuracy, is computed as the total differences in out-of-bag prediction error before and after the values of this feature are permuted among the training data, averaged over all trees, and then normalized by the standard deviation of the differences. The second measure, mean increase in residual sum of squares (RSS), is the total increase in RSS from splitting on the variable, averaged over all trees. In this study, we use mean increase in MSE to determine the statistically significant features. We also use feature selection approaches for other models, e.g., LASSO [36].

In the final model, there are 60 important features and most of them are new features created from the original ones. The ranking of corresponding original features using the mean decrease in the prediction accuracy is shown in Table VI. Here “Normalization” means the corresponding feature is normalized in the data preprocessing. To summarize, the measurements from WILD are the most important ones for predicting RUL, while those from MV are the least important.

VI. DISCUSSION

This study employs three types of wayside detectors including truck condition detector (OGD) and wheel condition detectors (WILD and MV). We found that all three types of...
detectors are useful to predict railcar’s RUL. However, WILD detectors provide more insightful information than the other two detectors. One reason is that WILD is contact-based measurements and previous studies showed that such measurements can provide a good reliability in the readings from the systems [7]. The other reason is that WILD is more widely installed in North America than the other two types of detectors. In this case, a truck could have more WILD readings than the other readings. Therefore, compared with OGD and MV, it has less missing data from WILD and better smoothers can be estimated. In addition to WILD, OGD provides good complementary readings on vehicle performance level. It can detect angle of attack, rotation, lateral positions and hunting amplitudes of wheelsets in a truck. The results of this study also indicate that almost all OGD readings can significantly contribute to the truck failure prediction. MV could also provide useful failure information. However, it is well-known that some MV measurements are correlated. For example, flange thickness and flange height are negatively correlated, and rim thickness and flange height are also negatively correlated. So it is sufficient to use only a subset of MV readings for modeling the wheel wear.

This paper combines wheel failure and truck failure for prediction, since either failure will need to set out the entire vehicle in the workshop. By aggregating measurements onto truck level, we are able to reduce the total amount of raw data significantly, and it will not damage our prediction performance and accuracy. We do not consider aggregating onto equipment (railcar) level, since one railcar could contain more than ten trucks. Prediction on equipment level is expected to involve much more visual inspection time than on truck level.

If only one failure type, either wheel failure or truck failure, is considered, the proposed methodology is still valid. We implemented the RUL prediction method on predicting RUL of wheels and trucks separately. The significant factors found in both prediction models are composed of mixed WILD, OGD and MV readings, similar as what we have found in this paper. Therefore, This study confirms that wheel and truck failures are interconnected with each other. A new set of wheels can therefore, have more WILD readings and MV readings, similar as what we have found in this paper.

VII. CONCLUDING REMARKS

Railway wayside detectors can automatically identify potential railcar component failures, and improve railway safety and reduce rolling stock inspection and maintenance costs. Through this paper, a Random Forests based methodology was developed to assess the current health and predict RUL of both trucks (bogies) and wheels of a railcar by fusing measurements from the three types of wayside detectors, WILD, OGD and MV. In the final predictive model, the statistically significant features are from all three detector types. However, according to the number of the most important features, we can rank the significance of a detector as follows: WILD > OGD > MV. Further, we find that MissForest, a Random Forests based nonparametric imputation method, can effectively handle missing data in wayside detector readings. The median average percentage error (APE) of RUL prediction is 22.5%, which is adequate for middle-term (60–180 days) proactive maintenance planning. After obtaining RUL estimates, maintenance actions can be planned within the lead time, and the identified defective railcar can be set out in appropriate timing for necessary maintenance activities.

In this paper, we only examined three types of wayside detectors due to data availability. Nowadays more condition monitoring technologies for railcars have been developed rapidly. Railroads may apply similar data fusion methods proposed in this paper for other detectors to prognosticate railcar failures.

Further data analysis is still needed to fully disclose underlying relationships between different types of failures. Such relationships can facilitate maintenance procedures and strategies to reduce operation costs in current railway practice.

REFERENCES
