# Unifying Railcar Monitoring Sensor Data, Maintenance Records, and Railcar Usage Information through Big Data Processing for Optimizing Railcar Maintenance and Safety

Hamid Sharif, Ph. D.
Charles Vranek Professor
Electrical and Computer Engineering Department
University of Nebraska-Lincoln

Michael Hempel, Ph. D.
Research Assistant Professor
Electrical and Computer Engineering Department
University of Nebraska-Lincoln

A Report on Research Sponsored by

University Transportation Center for Railway Safety (UTCRS)

University of Nebraska-Lincoln

August 2018

The University of Texas Rio Grande Valley / 1201 West University Drive / ENGR Portable 1.100 / Edinburg, Texas 78539-2999
Phone: +1 (956) 665-8878 / Fax: +1 (956) 665-8879 / railwaysafety@utrgv.edu / railwaysafety.utrgv.edu

## TECHNICAL REPORT DOCUMENTATION PAGE

| | | |
|---|---|---|
| **1. Report No.**<br>26-1121-0018-010 | **2. Government Accession No.** | **3. Recipient's Catalog No.** |
| **4. Title and Subtitle**<br>Unifying Railcar Monitoring Sensor Data, Maintenance Records, and Railcar Usage Information through Big Data Processing for Optimizing Railcar Maintenance and Safety | | **5. Report Date**<br>August 2018<br>**6. Performing Organization Code** |
| **7. Author(s)**<br>Hamid Sharif, Ph.D.<br>Michael Hempel, Ph.D. | | **8. Performing Organization Report No.**<br>26-1121-0018-010 |
| **9. Performing Organization Name and Address**<br>Nebraska Transportation Center<br>University of Nebraska-Lincoln<br>Prem S. Paul Research Center at Whittier School<br>2200 Vine Street<br>Lincoln, NE 68583-0853 | | **10. Work Unit No.**<br>**11. Contract or Grant No.**<br>DTRT13-G-UTC59 |
| **12. Sponsoring Agency Name and Address**<br>University Transportation Center for Railway Safety<br>The University of Texas Rio Grande Valley<br>1201 W University Drive<br>Edinburg, Texas 78539 | | **13. Type of Report and Period Covered**<br>Final Report (October 2016 – June 2018)<br>**14. Sponsoring Agency Code** |
| **15. Supplementary Notes**<br>Conducted in cooperation with the U.S. Department of Transportation, Federal Highway Administration. | | |


**16. Abstract**

With this project, we investigated the use of Big Data Analytics to make rail transportation safer, by preventing derailments due to equipment failure. Railroads typically schedule railcar maintenance on best-practice intervals, which may not include the plethora of information available from their maintenance logs, track data, sensors information, bills of lading, manufacturer history, etc. This project explored the use of this data to adapt maintenance scheduling to reduce cost and increase safety. We showed the great potential inherent in this approach.


| | | | |
|---|---|---|---|
| **17. Key Words**<br>Railroad, Maintenance, Big Data, scheduling, Random Forest, Rail Safety | **18. Distribution Statement**<br>No restrictions. | | |
| **19. Security Classif. (of this report)**<br>unclassified | **20. Security Classif. (of this page)**<br>unclassified | **21. No. of Pages**<br>31 | **22. Price** |

Form DOT F 1700.7 (8-72) Reproduction of completed page authorized

Table of Contents

List of Figures

List of Tables

## Disclaimer

The contents of this report reflect the views of the authors, who are responsible for the facts and the accuracy of the information presented herein. This document is disseminated in the interest of information exchange. The report is funded, partially or entirely, by a grant from the U.S. Department of Transportation's University Transportation Centers Program. However, the U.S. Government assumes no liability for the contents or use thereof.

Abstract

With this project, we investigated the use of Big Data Analytics to help make rail transportation safer, by preventing derailments due to equipment failure. Railroads typically schedule railcar maintenance on best-practice intervals, but that may not include the plethora of information available from their maintenance logs, track data, sensor information, bills of lading, manufacturer history, etc. This project explored the use of this data to adapt maintenance scheduling to reduce cost and increase safety. We showed the great potential inherent in this approach. Train accidents can be attributed to human factors, equipment factors, track factors, signaling factors, and miscellaneous factors. Big Data Analytics techniques can be utilized to provide insights into possible accident causes, thus resulting in improving railroad safety and reducing overall maintenance expenses as well as spotting trends and areas of operational improvements. We proposed a comprehensive Big Data approach that provides novel insights into the causes of train accidents and find patterns that led to their occurrence. The approach utilizes a combination of Big Data algorithms to analyze a wide variety of data sources available to the railroads, and is being demonstrated using the FRA train accidents/incidents database to identify factors that highly contribute to accidents occurring over the past years. The most important contributing factors are then analyzed by means of association mining analysis to find relationships between the cause of accidents and other input variables. Applying our analysis approach to FRA accident report datasets we found that railroad accidents are correlating strongly with the track type, train type, and train area of operation. We utilize the proposed approach to identify patterns that would lead to occurrence of train accidents. The results obtained using the proposed algorithm are compatible with the ones obtained from manual descriptive analysis techniques.

# Chapter 1 Introduction

As a result of our frequent communications and close collaborations with the Union Pacific Railroad, one major problem that requires in-depth study and analysis is **the need for new approaches to better understand railcar component failures to improve the safety and reliability of rail equipment maintenance and operation.** We firmly believe that Big Data analytics is key to significantly improved equipment reliability and reduced failure rates. The railroad industry separately collects records on equipment maintenance procedures, railcar design and component lists, train movement and bills of lading. But the key to better insights is in combining all of the available data and extracting new parameters, new knowledge.

The overall approach to solving this problem is shown in the following figure:


Fig 1 - Envisioned overall architecture for Big Data Analytics for Maintenance Optimization

This seed project aimed at exploring some of the core components of the overall architecture. For us, the focus was on demonstrating that Big Data Analytics can be used to extract new insights and that there is true potential for improved maintenance scheduling through forecasting of component failure. If we know how long before a component fails, then we can leverage this analysis across the entire railcar to determine the optimal point in time to conduct maintenance.

This research effort aimed at addressing a timely and urgent need in transportation safety: preventing costly and devastating derailments through optimized equipment maintenance using Big Data Analytics. Safety continues to be of a primary concern within the North American railroad industry, highlighted by efforts in freight train Wireless Sensor Network monitoring and Positive Train Control (PTC). Despite these efforts, statistics by the Federal Railroad Administration (FRA) Office of Safety Analysis [1] show that from 2010 through 2015 over 1000 derailments occurred directly linked to rolling stock equipment failure, causing over $240 million in losses.

Current methods for equipment maintenance rely on fixed schedules, which either are too frequent and result in unnecessary operational expenses, or are not frequent enough and result in high equipment failure rates. Despite producing detailed records for all maintenance efforts, incidents, etc., this data remains largely unutilized in the optimization of operational processes such as maintenance scheduling, supplier quality ranking, parts optimization based on past comp information includes operational data, accidents/incidents data, track maintenance data, safety data, inventory and highway-rail crossing data, and inspection and maintenance data [5]. Traditionally, these data sets have been stored in multiple databases and analyzed independently using traditional descriptive analysis techniques. However, these databases can be brought together and analyzed using Big Data Analytics techniques in order to uncover hidden patterns and find correlations that might not be easily discovered from analyzing data separately. In addition, Big Data analysis would allow the usage of predictive and perspective analysis techniques to forecast future safety measures and provide insights into possible accident causes, manufacturer issues, and more as shown in the Fig. 2 below.


Fig 2 - Freight Trains Input Data for Big Data Analytics

## Chapter 2 Our Approach

Big Data Analytics tools can combine railroad accidents/incidents database with operational and maintenance databases and allow for prediction of train failures before they occur. It could also allow for efficient scheduling of train and track maintenance thus enhance rail safety and reduce the costs caused by unnecessary maintenance. There are predictive Big Data algorithms that are well known for their accuracy including Random Forest (RF) and association mining algorithm. RF is the most popular algorithm in conducting in-depth study of Big Data [6]. It has classification and regression capabilities and high-performance efficiency. RF also gives estimates of what variables in the input data are more important in achieving certain responses [7]. This latter property is very significant as it enables selecting the important features and build a simple model based on these features, thereby reducing the computational cost.

Association mining algorithms, on the other hand, analyze the input data set for frequent patterns [8]. They automatically find the patterns that would take a long time to find manually using descriptive analysis techniques. The advantage of association algorithms over RF algorithms is that associations can exist between any of the input variables. While the RF algorithm builds rules with only a single conclusion, the association algorithms attempt to find many rules, each of which may have a different conclusion. Association algorithms use the support and confidence criteria to identify the most important relationships. Support is an expression of how frequently the variables appear in the input data, whereas confidence expresses how often that relationship has been found to be true within the data set. The main drawback in association algorithms is the computational efficiency as they require extensive processing time to find patterns within a potentially large search space [9-11].

In this work, we develop a comprehensive Big Data algorithm that utilizes the importance measurement feature from RF algorithm and the pattern detection capability of association mining algorithms. The importance measure helps in choosing the most important variables in the input data and thus increase the computation speed of the association mining algorithms. Data optimization is made possible through Big Data Analytics. However, the particular nature of the railroad applications, combined with the myriad different reporting formats in use by the railroads, their supplies, and at various operations centers poses significant challenges to current data analytics approaches. Our team studied how to address the various research challenges that currently limit Big Data Analytics. We researched required methodologies and demonstrate Big Data Analytics' capabilities using synthetic or real-world data provided by Union Pacific.
We believe that this effort is a vital component in further enhancing railroad operational safety and prevent derailments and the resulting significant monetary and environmental damages.

2.1 Methodology

The proposed algorithm utilizes both RF and association mining algorithms. RF allows selection of the most important variables in the input data subject to a specific response and feeds them to the association algorithm that discovers the connection between the variables. Here is the algorithm pseudocode:

Let N be the number of rows in the input data, M be the number of columns and K is a subset of the possible categories

Determine $m \subseteq M$ such that m has high impact on deciding K, using the importance feature from RF algorithm

Find $X \rightarrow Y$ where $X \subseteq m$, $Y \in K$ and $X \cap Y = \emptyset$

Find support $\sigma(X \rightarrow Y)$ and the confidence $C(X \rightarrow Y)$[11]

Choose $X \rightarrow Y \ni \sigma(X \rightarrow Y) > 0.1$ and $C(X \rightarrow Y) \geqslant 0.8$

RF used in step 2 is an aggregation of decision trees where every node in the tree is used as a binary condition on a single variable of the input data set. The condition at each node splits the variables into two groups, such that each group contains data that provides a similar response. The measure of the optimal splitting condition is based on Gini impurity. When training RF with the input data set, the decrease in the weighted impurity caused by each variable of the input data set is computed. The impurity reduction caused by each variable is averaged and the variables are ranked according to this measure. Variables that can remove more impurity are ranked as more important than the ones that remove less impurity. We can think of the important variables (m) as the ones who contributed the most to the rules formed by RF algorithm and thus a change in their value would degrade RF prediction ability as measured by out-of-bag (OOB) techniques [11].

The implication relationship in step 3 is the association mining rule where X and Y are called antecedent and consequent, respectively. In step 4 we select the rules from the set of all possible rules found by the association mining algorithm constraints to the thresholds on support and confidence measures. A rule is identified as important if the confidence and the support are within 0.8 and 0.1, respectively.

# Chapter 3 Implementation and Results

## 3.1 Input Data

The proposed algorithm was implemented in RStudio [12] by leveraging both RF and "arules" packages. In order to assess the algorithm efficiency, we tested the algorithm on the Federal Railroad Administration (FRA) accident/incidents database and compared the obtained results with the ones from manual descriptive analysis.

The input data set used is from the Federal Railroad Administration accident data sets [13] obtained for the period from January 2013 to December 2016. It contains information regarding a variety of conditions or circumstances that may have contributed to the occurrence of the reported accidents. The data accounts for damages to on-track equipment, signals, track, track structures, and roadbed. It comprises 50 columns (M), which are the fields from the "F.6180.54" form, and 9864 rows (N) that represent the number of accident/incident reports filed over the mentioned time period. According to the data base, there are five major classes (K) of train accidents, namely: human factors (H), equipment factors (E), track factors (T), signaling factors (S), and miscellaneous factors (M). The number of accidents in each accident cause category is shown in Fig. 3.

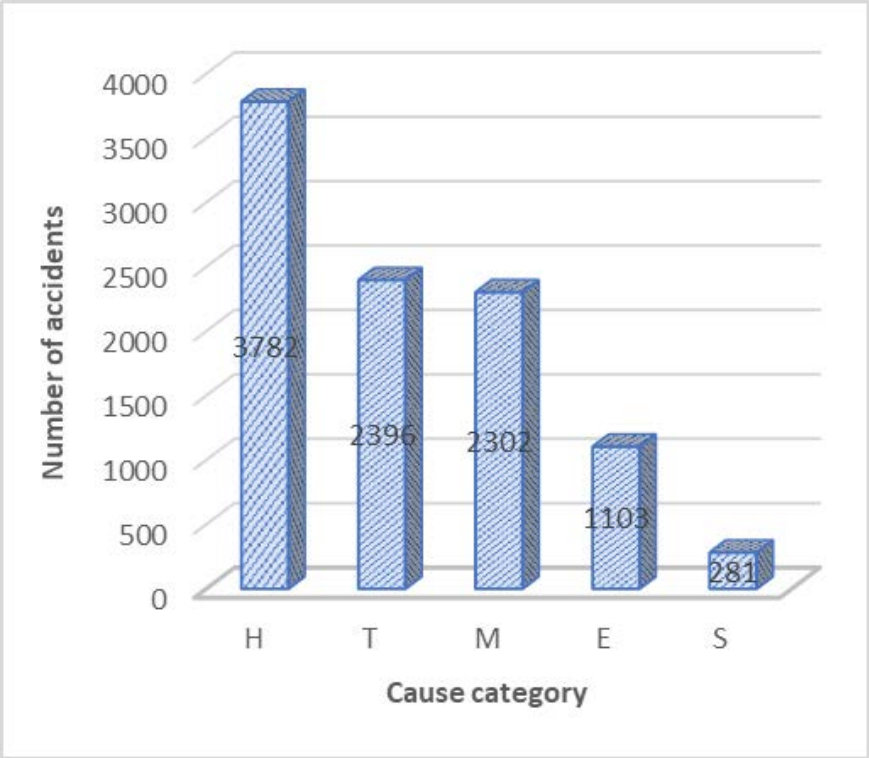


Fig 3 - Accident causes category versus number of accidents

3.2 Important Features Selection

The input data is applied to the RF algorithm in order to find the variables that contributed the most to the cause of these accidents, based on the mean decrease in Gini impurity.

Fig. 4 displays the 30 most important variables in the input data on the y-axis and the mean decrease in Gini score on the x-axis. A higher value of mean decrease in Gini score implies a higher importance of the associated variable. For example, the grade crossing ID number (GXID) and the DRUG in Fig. 4 are the most important variables in predicting the cause of accident. Table 1 lists the most important variable and their description.

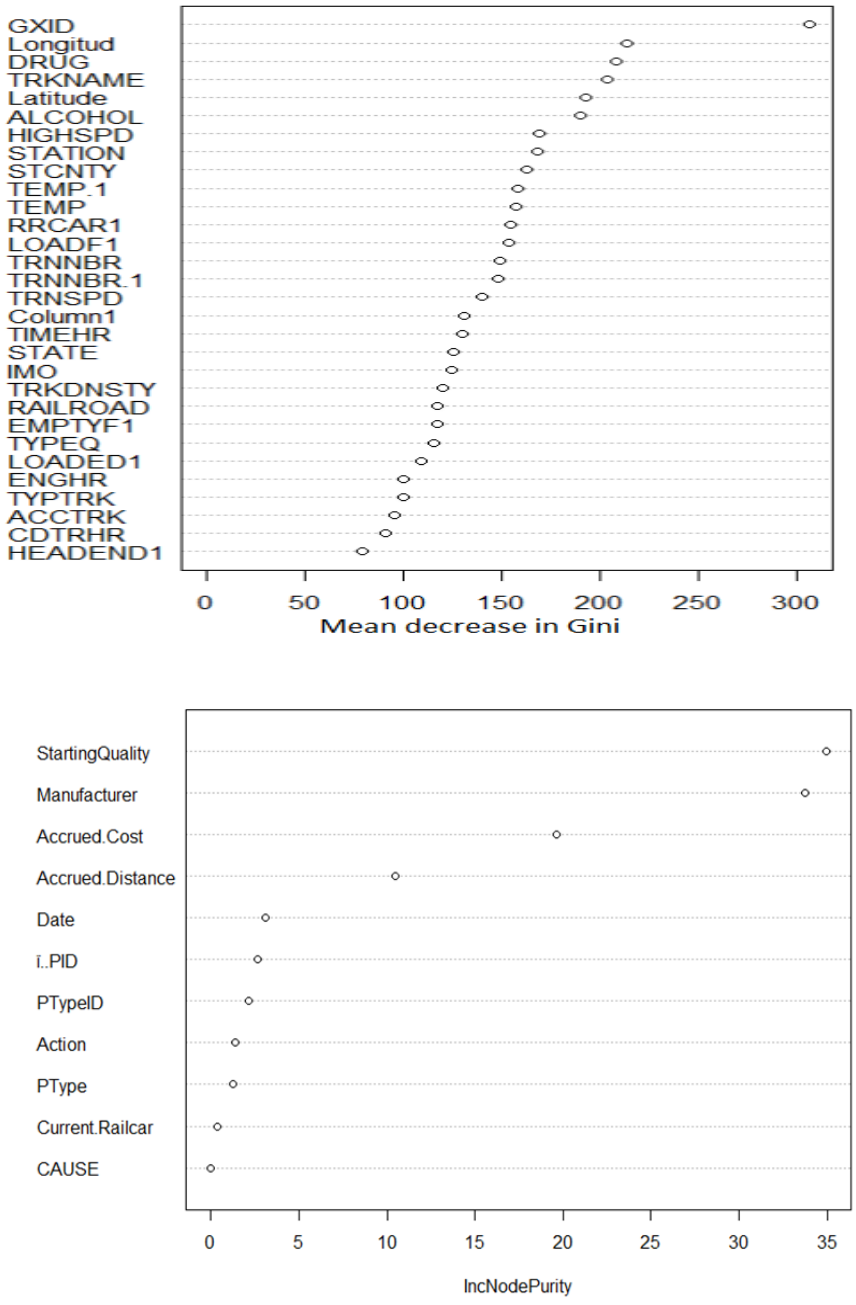


Fig 4 - Importance plot

Table. 1: Most Important Variables

| Feature acronym | Description |
|---|---|
| GXID | Grade crossing ID number: 0= No grade crossing, 1= Grade crossing |
| DRUG | Number of positive drug tests: 0=No positive drug test reported, 1=positive drug test reported |
| Longitude | Longitude in decimal degrees |
| TRKNAME | Track name |
| Latitude | Latitude in decimal degrees |
| ALCOHOL | Number of positive alcohol tests 0=No positive alcohol test reported, 1=positive alcohol test reported |
| HIGHSPD | Maximum speed reported for equipment involved |
| STATION | Nearest city and town |
| RRCAR1 | Car initials (first involved) |
| STCNTY | FIPS State & County code |
| TEMP | Temperature in degrees Fahrenheit |
| TRNNBR | Train ID number |
| LOADF1 | Number of loaded freight cars |
| TRNSPD | Speed of train in miles per hour |
| Column1 | Gross tonnage, excluding power units |
| TIMEHR | Hour of incident |
| IMO | Month of incident |
| STATE | FIPS State code |
| EMPTYF1 | Number of empty freight cars |
| RAILROAD | Reporting railroad |
| TRKDNSTY | Annual track density - gross tonnage in millions |
| LOADED1 | car loaded or not (first involved): Y=yes N=no blank=not applicable |
| TYPEQ | Type of train: 1=freight train, 2=passenger train, 3=commuter train, 4=work train, 5=single car, 6= cut of cars, 7= yard / switching, 8= light loco(s), 9= maintenance / inspection car |
| ENGHR | Number of hours engineers on duty: blank=not applicable |
| TYPTRK | Type of track: 1=main, 2=yard, 3=siding, 4=industry |
| CDTRHR | Number of hours conductors on duty: blank=not applicable |
| HEADEND1 | Number of head end locomotives |

The most important variables are applied to the association algorithm, which resulted in 58987 patterns. However, it is clear that going through all these patterns manually is not a viable option. Therefore, we used the scatter plot to visually see the rules and interactively choose the most significant ones based on their confidence value. The scatter plot of the confidence and the support for all rules is shown in Fig. 5. The plot consists of the support as x-axis and confidence as y-axis and each dot on the plot represents one of the obtained rules. We adjust the logarithm so that we can see only the patterns with confidence higher than 80%. Also, the dots are color coded so that the red dots indicate that the rule has high confidence value is important and needs to be further explored.

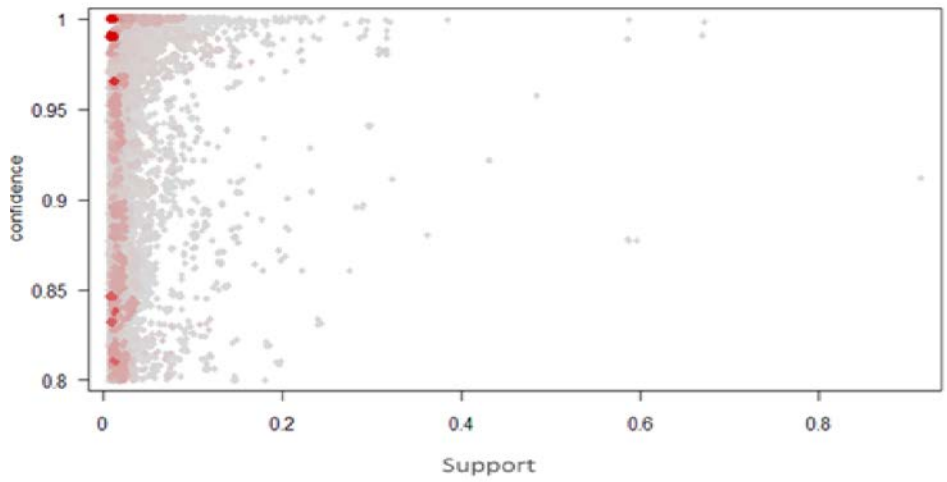


Fig. 5 - Scatter Plot

Table 2 displays an example of the four most important patterns and their support and confidence. These patterns are regarded as important because the confidence is above 80%. The first rule states that accidents due to human factors (H) often occur at rail yards (TYPEQ= yard / switching) when no grade crossing is involved (GXID=No) and the train engineers are not under the influence of drugs (DRUG=No). The algorithm also states that this pattern is 98.3% reliable

and applies to 12.4% of the input data. By analyzing the data manually, we found that among the 3782 accidents that are caused by human errors, 1397 accident occurred at the rail yards when train engineers tested negative on drugs and no GXID report was found. Therefore, the manual results are compatible with the automatic results obtained using the proposed algorithm. The second pattern states that the accidents caused by Miscellaneous factors (M) often occurs to passenger trains (TYPEQ=Passenger train) on a single main track (TRKNAME=Single main track) when train engineers are not on drugs (DRUG=No). It also states that this pattern applies to 10.5% of the input data and has 97.8% reliability. Manual analysis confirms that the highest number of accidents (34/35) due to Miscellaneous (M) factors occurred to passenger train on a single main track when train engineers tested negative for drugs. The third significant pattern in Table 2 implies that accidents caused by track factors (T) often occurs to freight trains (TYPEQ= Freight train) in state 48 (Texas) given no alcohol (ALCOHOL=No) or drugs (DRUG=No) are involved and no GXID (GXID=No) is involved. It also states that this pattern applies to 19.5% of the input data and has 94.4% reliability. This also agrees with the manual analysis which show that among the 639 accidents that happened to freight train in Texas, 200 accidents occurred due to track factors as illustrated in Fig. 6.

The last rule implies that most accidents caused by equipment factors (E) are occurring for freight trains (TYPEQ= Freight train) in state 48 (Texas) when the engineers are tested negative for drugs (DRUG=No). It also states that this pattern applies to 16.5% of the input data and has 91.2% reliability. This also agrees with the manual analysis, which shows that among the 639 accidents of freight train in Texas, 58 accidents occurred due to equipment factors as illustrated in Fig. 6.

Table. 2: Sample of the most important patterns

| Rule | Support | Confidence |
|---|---|---|
| {GXID=No, DRUG=No, TYPEQ= yard / switching} => {CAUSE=H} | 0.124 | 0.983 |
| {DRUG=No, TRKNAME=Single main track, TYPEQ=Passenger train} => {CAUSE=M} | 0.105 | 0.978 |
| {GXID=No, DRUG=No, ALCOHOL=No, TYPEQ= Freight train, State=48} => {CAUSE=T} | 0.195 | 0.944 |
| {DRUG=No, TYPEQ= Freight train, State=48 } => {CAUSE=E} | 0.165 | 0.912 |

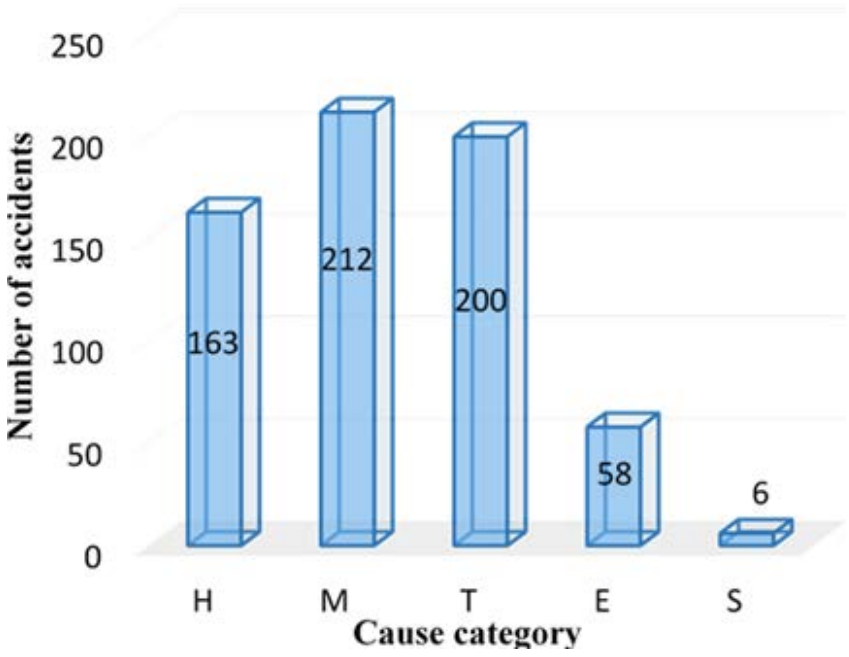


Fig. 6 - Freight train accidents in Texas between 2013 and 2016

Notice that the algorithm can be used to predict what cause category the accident should fall into. For instance, given that an accident occurred at rail yards (TYPEQ= yard / switching) when no grade crossing is involved (GXID=No) and the train engineers are not under the influence of drugs (DRUG=No), according to the first rule in Table 2, we can predict with 98.3% accuracy that the accident is caused by human factors.

Studying more closely the rules in Table 2 we can observe that the variables that appear in the rules are the ones that have the largest mean Gini score. The variable DRUG, for instance, is associated with every accident because the actual number of accidents involving DRUG is zero. GXID, on the other hand, has such a large mean Gini score since grade crossing accidents are quite common and the presence of a grade crossing is always a factor for a crossing related accident. Notice that all the variable that are regarded as important according to Table 1 have appeared in some rules generated from the association mining. However, some of those rules might have low confidence.

One of the main advantages of the proposed algorithm as compared to the manual analysis is the ability to detect and extract useful information from large-scale data with high computational speed and is scalable to very large datasets not feasible for manual analysis. Another key differentiator is that with the proposed approach is possible to detect the impact from weaker correlations among different parameters that may not be apparent using manual analysis.

# Chapter 4 Data Collection and Generation

Any of the intended activities involving Big Data depend on available data. We envisioned this data to be provided by Union Pacific and other railroads. However, while they were eager to support this effort, they simply could not share that data with outside parties, since it is highly sensitive and contains a plethora of operational insights. We were this faced with the prospect of having to procure the needed data a different way. The solution we pursued was to generate our own data sets. This had two key advantages:

1) It allowed us to control all relevant parameters in the generation of the data.
2) Because the data generation was fully under our control and we knew what the data contained, this also served as cross-validation of the Big Data analytics efforts that would utilize this data.

We therefore developed our own data generation tool, which is a full-fledged macro-train movement simulation across the United States. It simulates railcars, their components, train consists and cargo, their transit from a source to a destination location, failure events along the way, the impact of the train engineers driving habits, and much more. It also tracks component and railcar manufacturers and their product and service quality. Below is an image of the simulator in action.

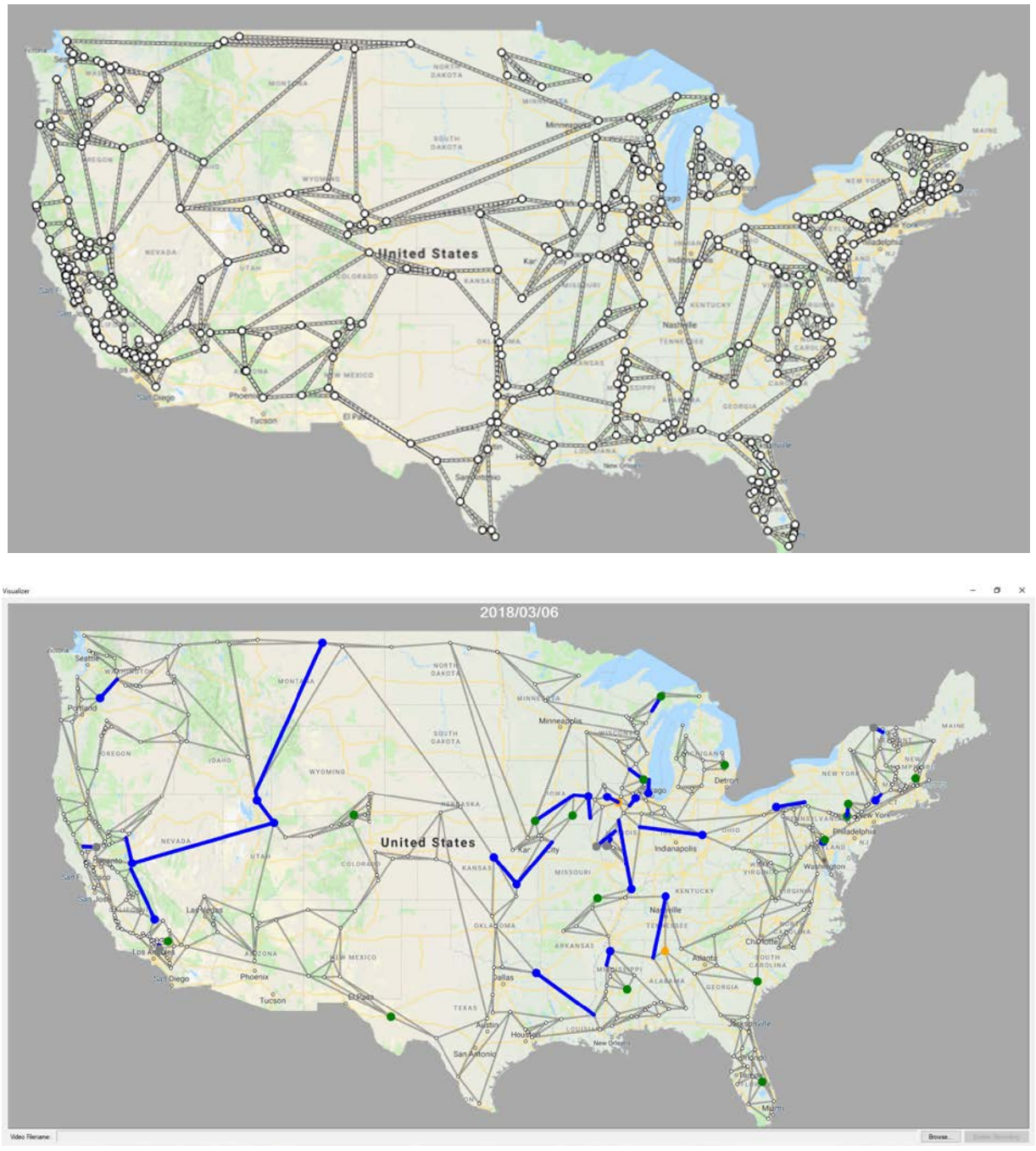


Fig. 7 - Screenshot from the Train Data Generator in action

In this particular simulation we can observe train movement (blue lines), train arrivals (green dots), and railcar failure events (yellow dot). The track network is abstracted as a direct

path between two train stations. Train station locations are retrieved from a set of 1000 GPS coordinates of real-world train stations in the United States. The simulator randomly selects a specified number of train stations and connects them in order to form the track network for the simulation instance.

All parameters used within the simulation is driven by random number distributions, fully configurable in a script file. This allows us to adjust any impacting factor for the simulation. Also considered by the simulator is the impact of terrain. For example, component wear and tear being worse when traversing mountains compared to the Midwestern plains. This is driven graphically through map files provided to the simulator.

The simulator also considers railcar maintenance. Servicing a railcar and its components leads to improvements in a component's quality, or its full replacement with a new component, in case the component quality has deteriorated too far. Railcar maintenance is scheduled with a fully controllable schedule for each railcar. This will allow us to evaluate different approaches on how to utilize Big Data Analytics and the forecasted component wear.

Finally, the simulator also tracks accumulated expenses for each railcar over its lifetime, including all maintenance and component replacements. It also considers, at a macro level, the cost of derailments. This allows us to express the economic impact of improving railcar maintenance scheduling.

The output of the simulator is a plethora of log files, such as information about each individual component's status, each railcar's complete history, maintenance records, and more. These data sets can then be processed using Big Data Analytics and compared to the initial conditions scripted into the data generator, in order to validate the accuracy of the conducted data analysis.

# Chapter 5 Data Processing

One of the key aspects in data analytics is to know what to focus on. With such a plethora of information available as in Big Data, it is easy to be deterred by the large number of input parameters and properties. Hence, our first focus was on exploring the use of Random Forest in gauging the importance of input parameters to the overall end result.

Random Forest is ideally suited for this task, because it produces as output an importance measure. We utilized it throughout our work. Random forests or random decision forests are an ensemble learning method for classification, regression and other tasks, that operate by constructing a multitude of decision trees at training time and outputting the class that is the mode of the classes (classification) or mean prediction (regression) of the individual trees. established that forests of trees splitting with oblique hyperplanes can gain accuracy as they grow without suffering from overtraining, as long as the forests are randomly restricted to be sensitive to only selected feature dimensions. A subsequent work along the same lines concluded that other splitting methods, as long as they are randomly forced to be insensitive to some feature dimensions, behave similarly. Note that this observation of a more complex classifier (a larger forest) getting more accurate nearly monotonically is in sharp contrast to the common belief that the complexity of a classifier can only grow to a certain level of accuracy before being hurt by overfitting.

The training algorithm for random forests applies the general technique of bootstrap aggregating, or bagging, to tree learners. Given a training set $X = x_1, ..., x_n$ with responses $Y = y_1, ..., y_n$, bagging repeatedly ($B$ times) selects a random sample with replacement of the training set and fits trees to these samples:

For $b = 1, ..., B$:

1. Sample, with replacement, $n$ training examples from $X$, $Y$; call these $X_b$, $Y_b$.
2. Train a classification or regression tree $f_b$ on $X_b$, $Y_b$.

After training, predictions for unseen samples $x'$ can be made by averaging the predictions from all the individual regression trees on $x'$.

$$\hat{f} = \frac{1}{B}\sum_{b=1}^{B} f_b\,(x')$$

This can also be derived by taking the majority vote in the case of classification trees.

$$\sigma = \sqrt{\frac{\sum_{b=1}^{B}(f_b(x')-\hat{f}\,)^2}{B-1}}$$

This bootstrapping procedure leads to better model performance because it decreases the variance of the model, without increasing the bias. This means that while the predictions of a single tree are highly sensitive to noise in its training set, the average of many trees is not, as long as the trees are not correlated. Simply training many trees on a single training set would give strongly correlated trees (or even the same tree many times, if the training algorithm is deterministic); bootstrap sampling is a way of de-correlating the trees by showing them different training sets. The above procedure describes the original bagging algorithm for trees. Random forests differ in only one way from this general scheme: they use a modified tree learning algorithm that selects, at each candidate split in the learning process, a random subset of the features. This process is sometimes called "feature bagging". The reason for doing this is the correlation of the trees in an ordinary bootstrap sample: if one or a few features are very strong predictors for the response variable (target output), these features will be selected in many of the B trees, causing them to become correlated. This can be analyzed as how bagging and random subspace projection contribute to accuracy gains under different conditions.

Typically, for a classification problem with p features, $\sqrt{p}$ (rounded down) features are used in each split. For regression problems the inventors recommend p/3 (rounded down) with a minimum node size of 5 as the default.

We can see that one of the most important factors overall was the grade crossing, indicating that a significant number of accidents occurred at highway-rail intersections. When we applied the same approach to our simulator data, we could see that one of the most important factors actually was the distance travelled and the manufacturer. This indicates that the reliability of manufacturing processes are a key focus, but also that normal operational tracking over a railcar's work orders plays a big role in predicting component failure. Also an important characteristic was the route travelled, indicating that the terrain impact in our simulator played an important role.

When focusing on the available railcar travel logs we could then analyze it to predict how much farther a railcar could travel before a component failure occurred. Due to the simplified nature in which our simulator produces wear and tear on components this could be extracted as a mathematical expression for a given railcar. In real-world applications, Big Data Analytics would be monitoring this property instead of making long-term forecasts using simple mathematical models. But the end result is the same: given the output of Big Data Analytics it was possible for each railcar to determine an approximate failure point, which can then directly be used to update maintenance scheduling. Furthermore, it also helped indicate what the most likely component contributing to the predicted failure will be, this directing maintenance efforts and streamlining the turnaround time. Overall, this has tremendous potential in helping the railroads reduce maintenance efforts while also increasing reliability and safety.

Different scenarios for analyzing breaks, axels, and bearing for different number of cars over multiple years have been simulated and analyzed. The results for each case is presented in the following Fig. 8 through 21.


Fig. 8 - US map with the mountains

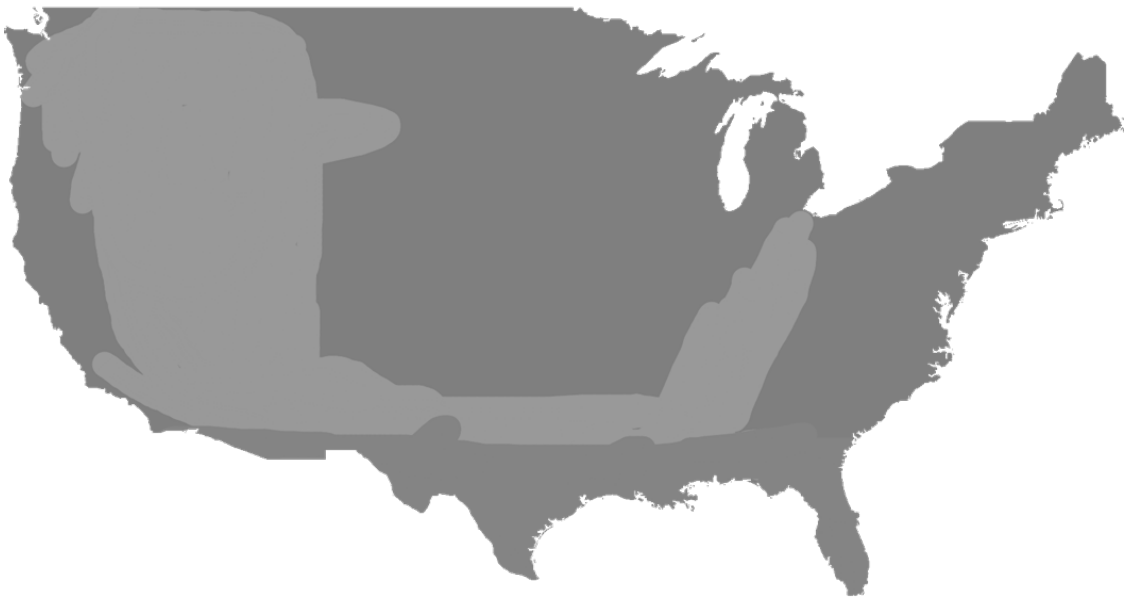

Fig. 9 - Damage based map

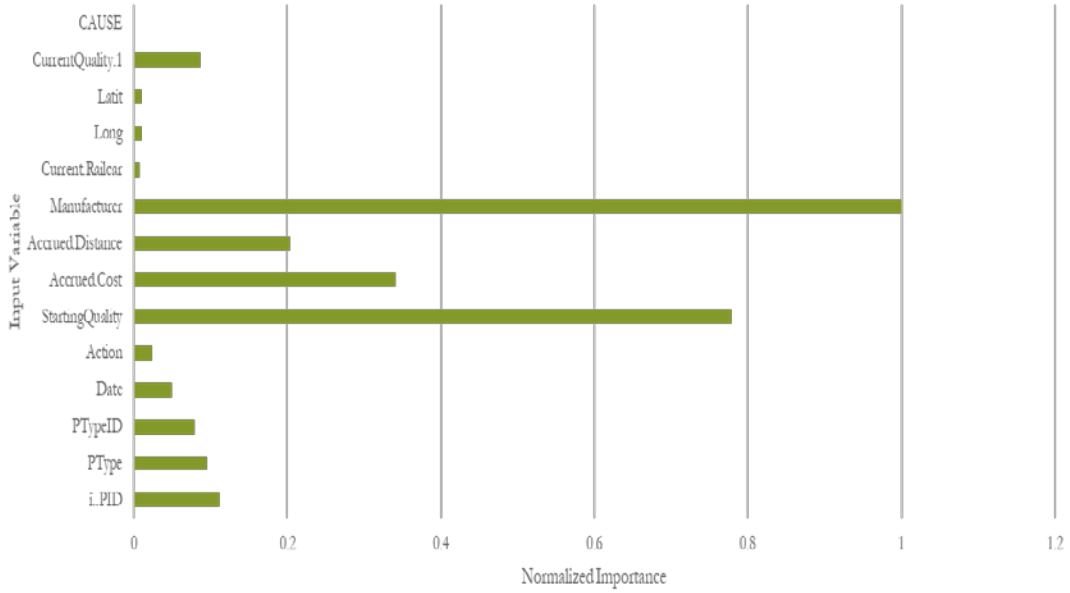


Fig. 10 - One car over period of 1 year

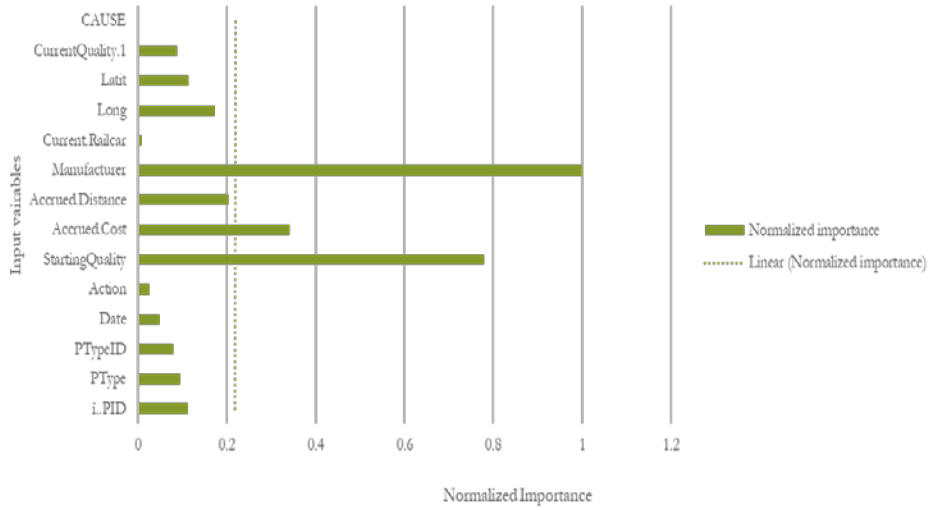


Fig. 11 - One car over the period of 10 years

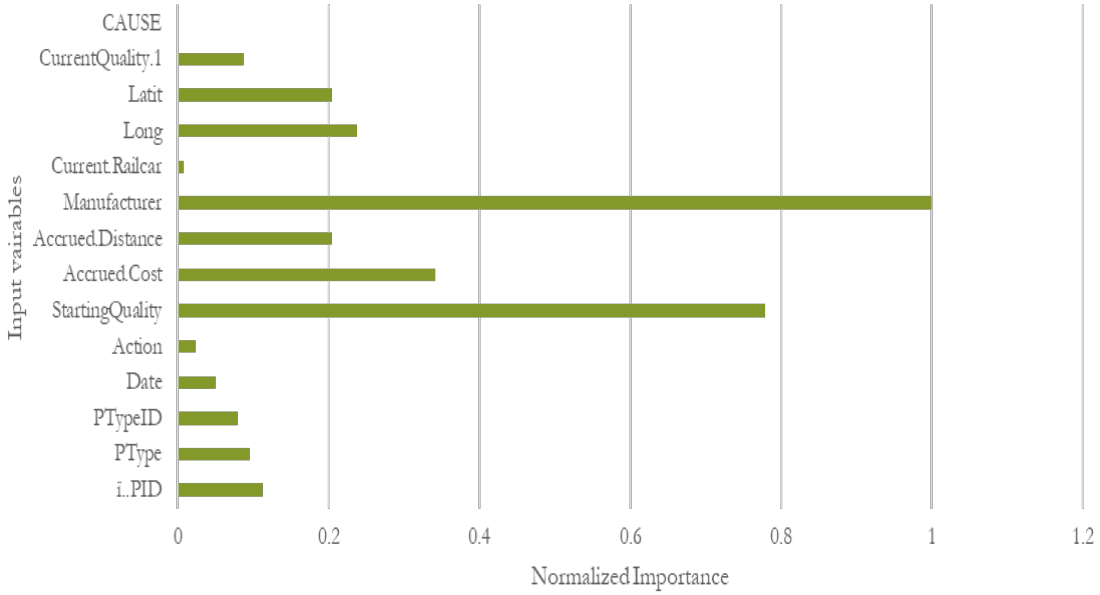


Fig. 12 - 5 cars over one year period

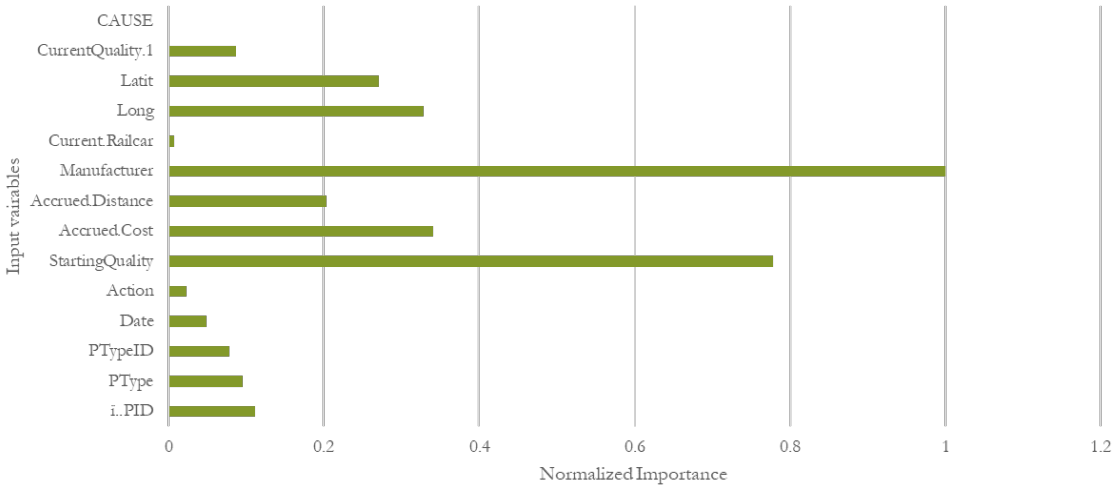


Fig. 13 - 5 cars over period of 5 years

- would the current quality of the brakes manufactured by manufacturer 34 changes as the train go through different regions on the US map


Fig. 14 –Scenario for the break quality

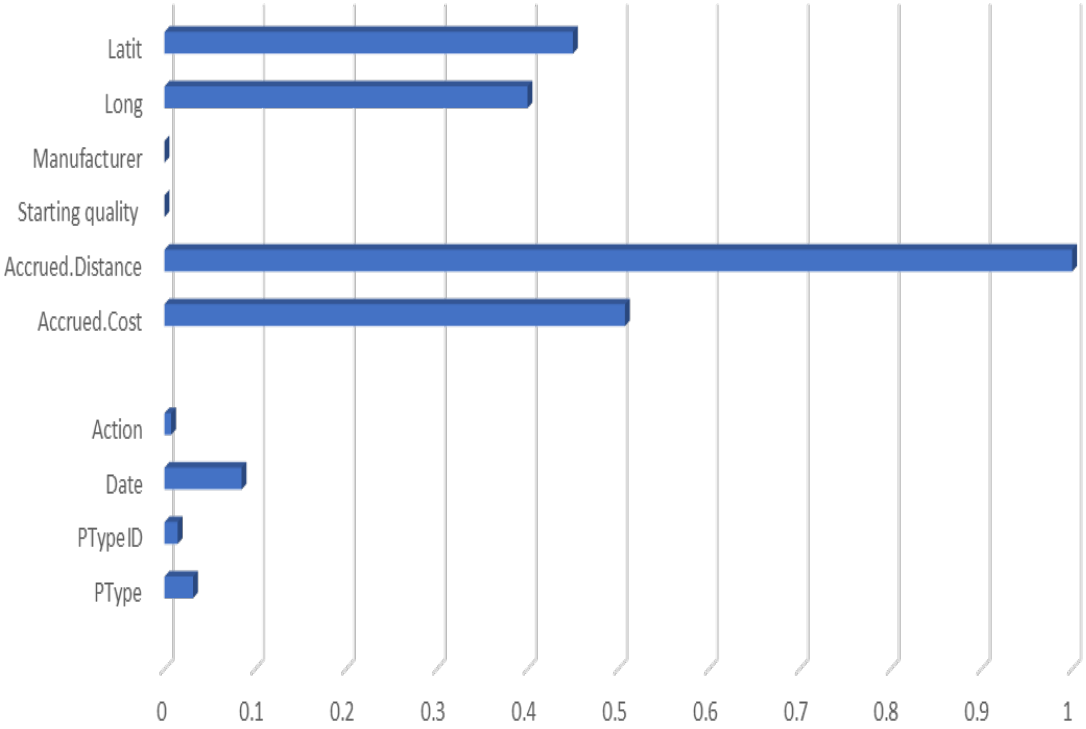


Fig. 15 - 5 cars over period of 10 years (Breaks)

- How would the current quality of the Axel manufactured by manufacturer 33 change as the train moves over different regions



Fig. 16 –Scenario for the Axel quality

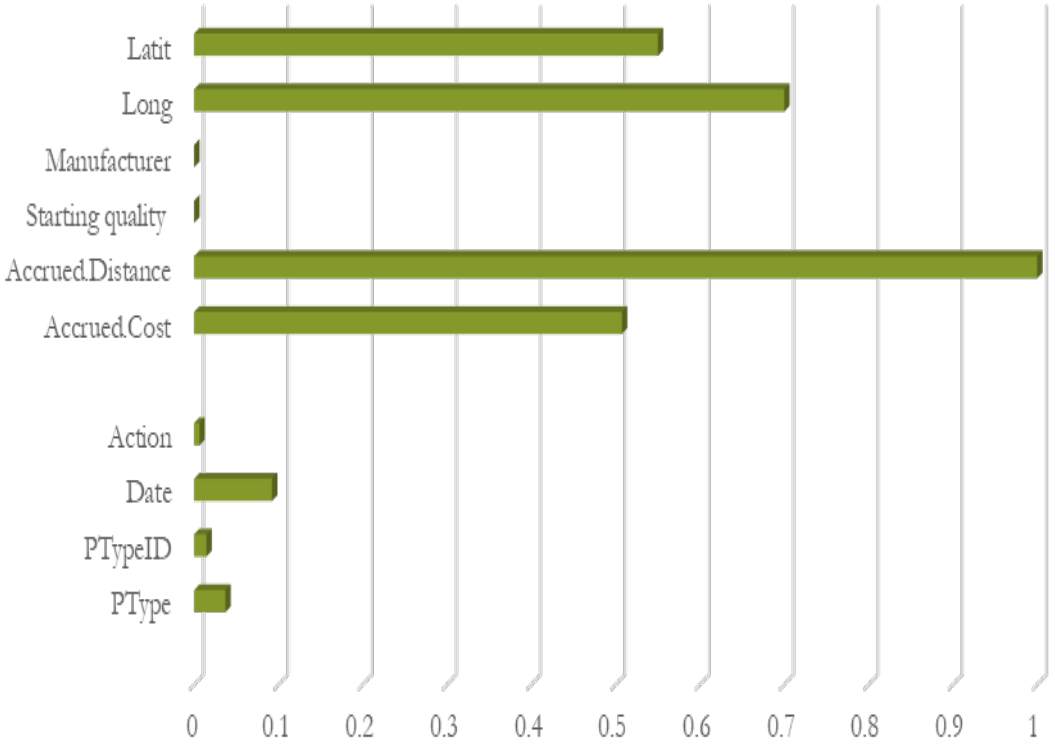


Fig. 17 - 5 cars over period of 10 years (Axel)

- How would the current quality of the bearings manufactured by manufacturer 35 change as the train moves over different regions


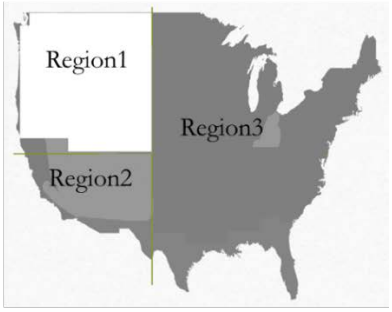


Fig. 18 - Scenario for the Bearing quality

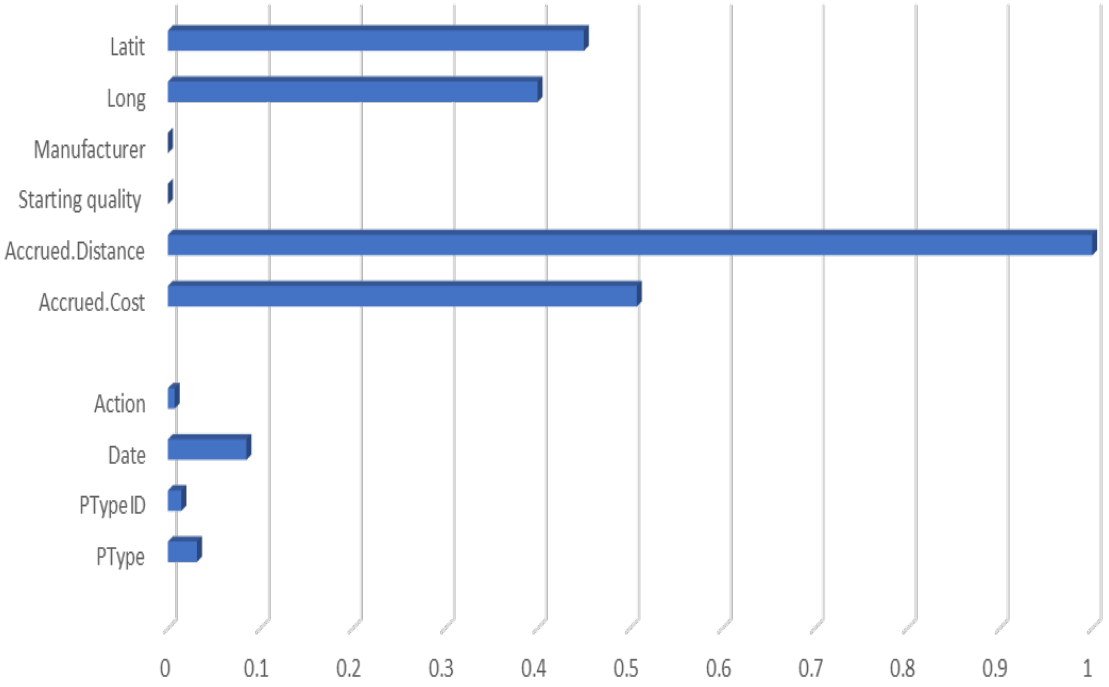


Fig. 19 - Two cars over period of 15 years

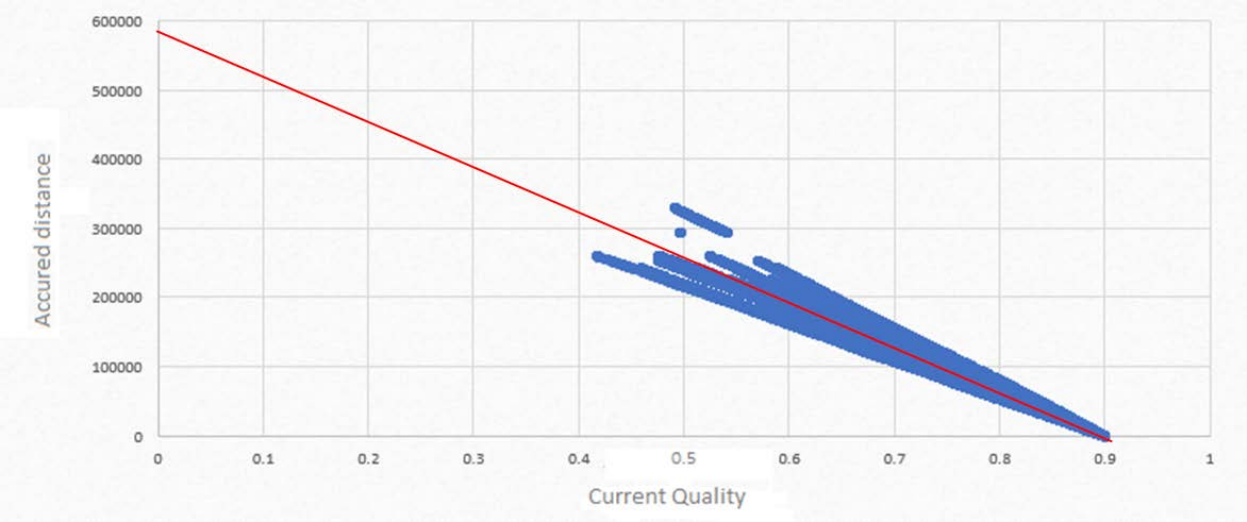


Fig. 20 - 2 cars over period of 20 years

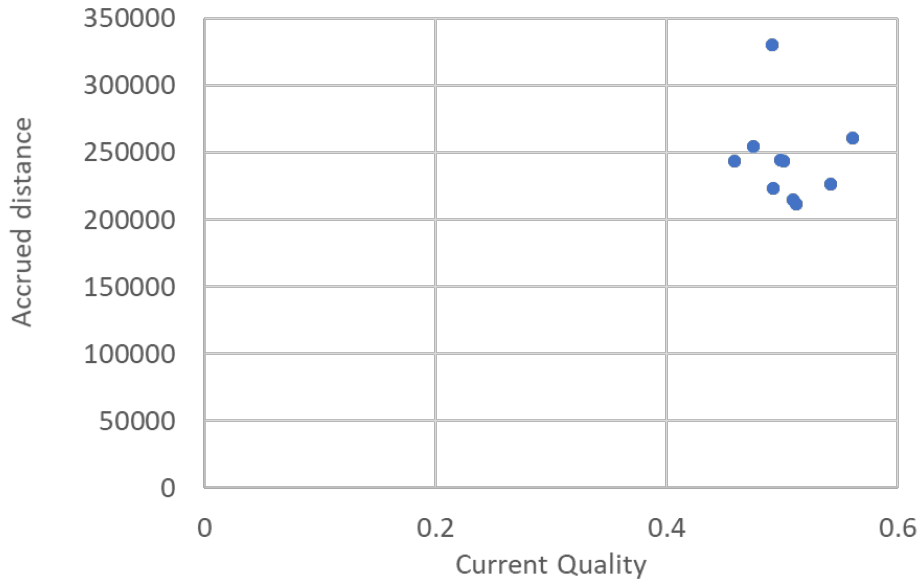


Fig. 21 - Quality versus distance for failed wheels for 2 cars over period of 20 years

# Chapter 6 Summary and Conclusions

## 6.1 Summary

We proposed a comprehensive Big Data algorithm that utilizes the importance feature from RF algorithms and the pattern detection ability from the association mining algorithms in combination to reduce computational complexity while retaining all the insights available from the collective data set. The developed algorithm was applied to the FRA accidents/incidents data as an evaluation tool for its efficacy and has shown results similar to the results that were obtained using manual analysis, thus validating its accuracy. Our work shows that Big Data analytics applies to maintenance and operational data can reliably identify accident categories and cause factors, and thus assist with improving the productivity, reliability, and safety of the rail operations.

## 6.2 Future Work

This was a truly exciting and engaging project, focused on a topic of vital importance for the freight railroad industry in North America – leveraging all of their collected data and analyzing it to obtain new insights into component and railcar degradation that can lead to derailments if not detected or repaired in time. But maintenance scheduling thus far is driven based on best-practice intervals, without insights from such data, and thus either misses approaching component failure or is performed too frequently and thus unnecessary expense. By leveraging our insights from this project, which shows that Big Data can make a significant impact on maintenance schedule optimization, we can help improve rail safety and reduce operational expenses.

## 6.3 Publications Resulting from Research

During this project we have thus far published one conference paper, and are in the process of completing the writing on two more papers. The published conference paper is titled “Novel Insights For Railroad Maintenance Using Big Data Analytics”, by N. Albakay, M. Hempel, and H. Sharif, presented at and published in the conference proceedings of the 2018 ASME Joint Rail Conference, held April 18-21, 2018 in Pittsburgh, PA, USA [3].

References


1. Federal Railroad Administration, Office of Safety Analysis, [online], http://safetydata.fra.dot.gov/OfficeofSafety/default.aspx

2. Federal Railroad Administration(FRA): https://www.fra.dot.gov/Page/P0001

3. M. Zarembski, "Some examples of big data in railroad engineering", 2014 IEEE International Conference on Big Data (Big Data), 2014.

4. A. Zarembski, "Integration of multiple inspection system data to identify potentially unsafe track rail conditions: data collection, consolidation and preparation", Report Prepared for US Federal Railroad Administration, 2014.

5. Federal Railroad Administration(FRA): https://www.fra.dot.gov/Page/P0001

6. Y. Xu, "Research and implementation of improved random forest algorithm based on Spark", IEEE 2nd International Conference on Big Data Analysis (ICBDA), 2017.

7. W. Lin, Z. Wu, L. Lin, A. Wen and J. Li, "An ensemble Random Forest Algorithm for insurance big data analysis", IEEE Access, 2017.

8. A. Behnamian, K. Millard, S. N. Banks, L. White, M. Richardson and J. Pasher, "A systematic approach for variable selection with Random Forests: achieving stable variable importance values", IEEE Geoscience and Remote Sensing Letters, 2017.

9. Y. Motai, "Kernel association for classification and prediction: a survey", IEEE Transactions on Neural Networks and Learning Systems, 2015.

10. H. Yan and H. Hu, "A Study on association algorithm of smart campus mining platform based on big data", International Conference on Intelligent Transportation, Big Data & Smart City (ICITBS), 2017.

11. K. Max, and K. Johnson, "Applied predictive modeling", Vol. 26, Springer, 2013.

12. RStudio homepage, [Online] https://www.rstudio.com/

13. Federal Railroad Administration Office of Safety Analysis, [Online]http://safetydata.fra.dot.gov/OfficeofSafety/Default.aspx