TTCI TECHNOLOGY DIGEST

OCTOBER 2019 TD-19-021

The work described in this document was performed by Transportation Technology Center, Inc., a wholly owned subsidiary of the Association of American Railroads.

Rail Fatigue Defect Prediction Using Machine Learning: A Preliminary Study

Ananyo Banerjee (TTCI), Xiang Liu (Rutgers University)

Key Findings:

- Preliminary analysis shows that the developed machine learning algorithm has a reasonable fit to the empirical data.
- Annual traffic density, annual number of car passes, rail age, grade, rail size, curvature, and other factors were shown to affect rail fatigue defect probability.
- The machine algorithm was implemented into a decision support tool that can be used to automate rail defect prediction based on provided input information.

Transportation Technology Center, Inc. (TTCI) and Rutgers University explored the potential use of machine learning to better predict rail fatigue defects. This also could aid in prioritizing inspection and maintenance efforts. Early results of the algorithms developed for this project are promising.[1] Based on network-level traffic, maintenance, infrastructure, and inspection data from one Class I railroad, a pilot study was conducted to explore the feasibility of rail fatigue defect prediction leveraging big data analytics. Three advanced machine learning algorithms were developed, evaluated, and compared. An initial step in using machine learning for rail defect prediction progress indicates that appropriate machine learning algorithms may provide a reasonable model fit. Annual traffic density, rail age, annual number of car passes, grade, degree of curvature, and rail size are among important factors for rail fatigue defect prediction.

A working model was developed for predicting the probability of rail fatigue defect in a particular track condition. This research aims to advance a previous model called "HALTRACK," developed in the 1990s by Massachusetts Institute of Technology in collaboration with the researchers at the AAR's Research and Test Department, a predecessor of TTCI.[2] HALTRACK was calibrated based on hundreds of simulations in PHOENIX, a mechanistic model developed by the Association of American Railroads (AAR).

A recent HALTRACK evaluation project by TTCI and Rutgers University[3] highlighted the need to update the capabilities to predict rail failures based on current rail infrastructure and operational data. Further, a research need was identified to account for more variables related to rail fatigue defect occurrence. In particular, as railroad data accessibility continues to grow, it raises new interest to leverage big data analytics across multiple network-level databases for more accurate rail defect prediction. These research needs motivated the development of a preliminary machine learning model as described in this *Technology Digest*.

HALTRACK was unable to handle present volumes of data due to its limited capabilities and dependence upon simulated data rather than real data collected from railroads. However, additional research is needed to account for factors that were not incorporated in the current model. The model for this study is based on

one Class I railroad's data and can be expanded to a general model having multiple databases from different Class I railroads. Also, machine learning generally depicts correlation instead of causality. The latter would be a challenging but useful direction in the next step of this research. An understanding of rail defect occurrence is useful to identify high-risk spots; and thus, a better capability to prioritize inspection and maintenance. There is limited prior research to predict rail defect probability accounting for track-related characteristics (e.g., curvature, grade, presence of turnout), traffic-related characteristics (e.g., gross tonnage, number of wheel passes), rail defect and track geometry defect inspection histories, and maintenance activities (e.g., grinding, ballast cleaning) based on network-level data.

Data Collection and Integration

Four basic types of data were collected from 2011 to 2016 over a network of more than 20,000 miles of track. Information included inspection data such as detected rail defects, all types of track geometry defects and vehicle track interaction exception data; maintenance activity data such as rail grinding and ballast cleaning activities; track layout data such as curvature, grade, curve, rail age; and traffic exposure data such as annual tonnage and total number of car passes. A total of 12 raw databases were provided by the collaborative Class I railroad. The first step was to cross link these data sources based on the location information (e.g., route indicator, beginning and ending mileposts). In discussion with the data provider and a review of the literature, the entire network was classified into thousands of rail segments (varying length) according to track and traffic characteristics. A data screening process found that the variation of the same factor on the same segment was trivial; therefore, each segment could be considered as an approximately homogeneous unit. On each rail segment, there is information regarding prior defect history, traffic tonnage, track characteristics and other information as shown in Figure 1.

If a segment had a defect in a given year, the response variable (defect indicator) is equal to 1. Otherwise, the response variable is zero. In this way, the rail defect prediction problem can be viewed as a binary classification problem in machine learning field.

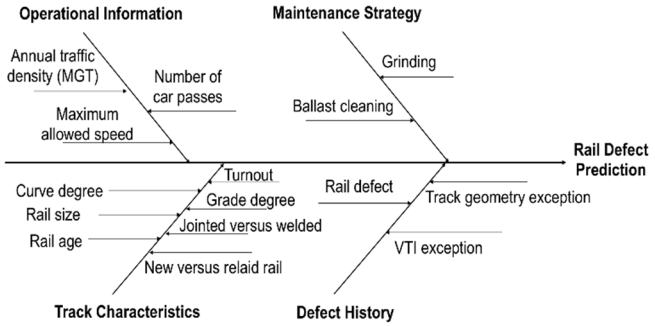

Figure 1. Factors included in the model (partial)

The binary classification is reasonable due to the fact that the same rail segment experiencing two or more defects is extremely low (>0.2 percent) in the dataset. Among all types of rail defects, this research focused on fatigue defects forming in the head of the rail (including detail fractures, transverse and compound fissures, horizontal splits, vertical splits), which comprised over one-third of all types of rail defects and the majority of rail-caused derailments on the studied railroad. This model may be modified to account for other types of rail defects in future analyses. Weld defects were not considered in this study since the research focus was only on defects originating in the rails due to fatigue loading. The analysis includes 554,878 rail segments (note: left rail and right rail is differentiated in recognition of their different responses to loading on the curved track). In this dataset, only 1.5 percent of segments experienced a rail defect in the study period. This indicates that rail defect prediction is a challenging rare event machine learning problem (imbalanced data mining).

Model Development

Machine learning models allow analysts to uncover hidden insights through learning from historical relationships and trends in the data and make predictions. Rail defect prediction can be described as a classification problem. The output would be either that there is a defect (coded as 1) or no defect (coded as 0). There are various machine learning algorithms that can be used for this type of problem. In this research, three algorithms — Extreme Gradient Boosting (XGBoost),[4] Random Forests (RF), and Logistic Regression (LR) — were used. XGBoost provides a method that solves many data science problems in a fast and accurate way. Random forests, which was developed

in 1990s, is a method for classification, regression and other tasks. The RF method is built upon a multitude of decision trees. LR is a classical regression technique, which describes the parametric relationship between input variables and binary output variables. These three algorithms represent three common types of machine learning methods, which are boosting, ensemble learning and single classifier, respectively. Based on the data used in this study, it was found that XGBoost outperformed the other two alternatives. Therefore, the following discussions will focus on the results using XGBoost. This research represents the first application of XGBoost to track big data analytics.

Model Validation and Comparison

The dataset was divided into five subsets of equal size in a process called five-fold cross validation. That ensured that each subset has the same proportion of positive (event) and negative (no event) examples using stratified sampling. For all possible choices of four subsets, the union of four subsets was used for model training, and the induced classifier was tested on the remaining subset. In this way, 80 percent of data was used for model development and the remaining 20 percent of the data was used for blind prediction. Each observation has the equal opportunity to be included in calibrating and testing data, minimizing sampling errors.

The confusion matrix was used to present the results of classification algorithms. True positives (TP) record the actual positives that are correctly classified. Similarly, false positives (FP) record the actual negatives that are incorrectly classified as positives. In this research, the positive represents the occurrence of rail fatigue defect, and the negative means no defect on the rail segment in a given year. The four values, TP (true positives), FN (false negatives), FP (false positives), TN (true negatives), are used to calculate sensitivity (Equation 1) and specificity (Equation 2).

$$\text{Sensitivity} = TP / (TP + FN) \qquad (1)$$

$$\text{Specificity} = TN / (FP + TN) \qquad (2)$$

Based on sensitivity and specificity, the receiver operating characteristic (ROC) curve is used to compare the performance of alternative classification algorithms as shown in Figure 2. In the ROC curve, the x-axis is 1-specificity, which means the false positive rate. The y-axis is the sensitivity, which means the true positive rate. The area under the ROC curve is called area under curve (AUC), which is widely used to compare model fit. A well-accepted rule of thumb is that if the AUC is over 0.8, the model may have a reasonable fit. The higher the AUC, the better the model fit.[5] The optimal probability threshold can be determined in order to binarize the predicted probability to best fit the empirical data. In the preliminary research, the AUC using XGBoost is 0.84, indicating an overall reasonable goodness of fit.

While inferring the importance of each input variable on the output (rail fatigue defect occurrence), it was found that annual traffic density, rail age, number of car passes (repetitive wheel loads on the rail), grade, rail size, degree of curvature, and rail size are among the most influencing variables for rail fatigue defect prediction.

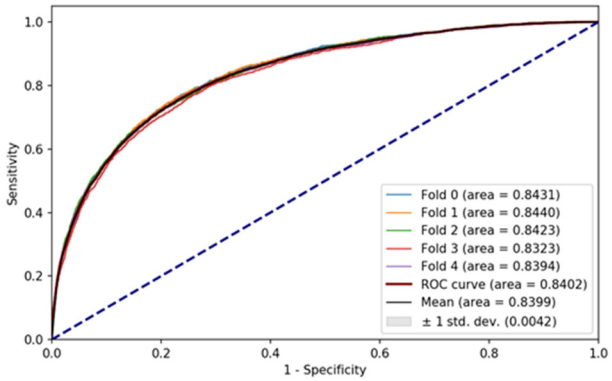

Figure 2. Receiver operating characteristic (ROC) curve

Rail Fatigue Defect Prediction Tool

A computer tool was created for the implementation of the machine learning algorithm. The tool takes the input information and predicts the rail fatigue defect occurrence probability for the specified segment. Table 1 shows the input variables for using the tool. The default value is the mean value of the distribution of each input variable in the dataset used for this research. Using the default values in Table 1 as an example, a quarter-mile rail section has an estimated fatigue defect probability of 0.13 in 1 year. If rail age increases to 80 years, the rail defect probability is expected to increase to 0.20. Similarly, if the annual traffic density on the track increases from 20 MGT to 60 MGT;

the rail fatigue defect probability would increase to 0.26 with all other factors being equal. It is to be noted that the machine learning model simultaneously accounts for multiple input variables in a complex way; therefore, the marginal effect of each variable is dependent on the values for other variables.

Table 1. Inputs for the prediction tool

Input	Input Range	Default Value
Rail age (years)	0-80	24
Segment length (miles)	0-10	0.25
Annual traffic density (MGT)	0-100	20
Annual number of car passes (in thousands)	0-700	250
Max. allowed speed (mph)	0-60	35
Curve (degrees)	0-20	0 for tangent; 3 for curved track
Grade (percent)	0-4	0.4
Rail position	Tangent rail, high rail, low rail	Tangent rail
Rail size (pounds/yard)	0-155	132
Number of turnouts/mile	0-2	0.5
Rail quality index	New rail, re-laid rail	New rail
Number of ballast cleaning/year	0-2	0
No. of grinding passes/year	0-4	1
No. of prior defects (all types) per mile/year	0-10	0.3
No. of prior vehicle track interaction (VTI) exceptions per mile/year	0-15	0.5
No. of prior track geometry defects per mile/year	0-10	2

CONCLUSION

This research used machine learning algorithms to predict rail fatigue defects based on big data across many railroad databases. The algorithm accounts for a variety of factors related to track infrastructure, maintenance activities, traffic volume and operations. The preliminary analysis shows that the developed machine learning algorithm has a reasonable fit to the empirical data. Annual traffic density, annual number of car passes, rail age, grade, rail size, curvature, and other factors were shown to affect rail fatigue defect probability. The machine algorithm was implemented into a decision support tool that can be used to automate rail defect prediction based on provided input information. The tool, upon being further developed, could be used to identify the locations with more chances of developing rail defects, thereby providing information to prioritize infrastructure inspection and maintenance.

FUTURE RESEARCH

The dataset used in this study is not comprehensive, nor exhaustive. Future research can be conducted to collect additional variables (e.g., rail wear, substructure condition) to examine whether the model can be further improved. Collecting data from other railroads can enlarge the sample size and better represent nationwide railroad operations. Machine learning generally depicts correlation, but not necessarily the causality between output and input variables. Future research is needed to better understand the causal impact of certain factors on rail defect risk, and how the changes of certain practices (e.g., maintenance frequencies) may change the rail defect occurrence probability.

ACKNOWLEDGEMENTS

The authors would like to thank Aihong Wen, Jennifer Hollar, Mark Dingler, Daniel Hampton, Kiomars Nassiri Kahnamooee, Ed Tubbs, Lisa Evans-Boley, Stephen Love and William Lahnen for providing data for this research.

References

1. Liu, X., Saat M. R., Barkan, C.P.L. 2012. Analysis of causes of major train derailment and their effect on accident rates. *Transportation Research Record: Journal of the Transportation Research Boa*rd, 2289, pp.154-163.
2. Martland, C.D., McGovern, M., and Shyr. F.Y. 1996. HALTRACK 96 Analysis of the effects of heavy axle loads on track: users' guide. MIT, Cambridge, MA.
3. Davis, D., Banerjee, A., Liu, X. 2016. Rail defect prediction model evaluation. *Technology Digest* TD-16-043. AAR/TTCI, Pueblo, CO.
4. Chen, T., Guestrin, C. 2016. XGBoost: A scalable tree boosting system. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 785-794.
5. Zhou. Z.H. 2012. Ensemble methods: foundations and algorithms. Chapman and Hall/CRC.

For comments or questions about this publication, contact Ananyo_Banerjee@aar.com