Transportation Research Record

Machine Learning Analysis of Railroad Freight Train Derailment Severity

--Manuscript Draft--

Full Title:	Machine Learning Analysis of Railroad Freight Train Derailment Severity
Abstract:	Freight train derailment rates have declined considerably over the past decade; however, when they do occur, they can damage infrastructure and rolling stock, interfere with transportation, and may cause casualties and releases of dangerous goods. Derailment analysis and prevention has long been a high priority for the rail industry and policy makers. Therefore, it is imperative to investigate both train derailment severity and consequences. Most previous research has focused on analysis of factors affecting the probability of train derailments, but less research has been conducted on understanding factors affecting derailment severity estimation. This paper aims to identify variables affecting derailment severity and develop a new methodology for estimation. Extensive, detailed data from the U.S. Department of Transportation Federal Railroad Administration (FRA) were used to construct and compare several machine learning models for train derailment severity prediction, including multiple regression, KNN, regularized regression (ridge and LASSO), random forest, and gradient boosting. The results can be used to provide insights for the industry to understand factors affecting derailment severity and enable better informed approaches to risk reduction.
Manuscript Classifications:	Rail; Railroad Operational Safety AR070; Train; System Safety; General
Manuscript Number:	
Article Type:	Presentation and Publication
Order of Authors:	Brandon Z. Wang
	Jaemin Kim
	Christopher P.L. Barkan, Ph.D.

Machine Learning Analysis of Railroad Freight Train Derailment Severity

Brandon Z. Wang
Graduate Research Assistant
Rail Transportation and Engineering Center – RailTEC
Department of Civil and Environmental Engineering
University of Illinois at Urbana-Champaign, Urbana, IL 61801
Email: zwang144@illinois.edu

Jaemin Kim
Graduate Research Assistant
Rail Transportation and Engineering Center – RailTEC
Department of Civil and Environmental Engineering
University of Illinois at Urbana-Champaign, Urbana, IL 61801
Email: jkim557@illinois.edu

Christopher P.L. Barkan, Ph.D.
Professor
Rail Transportation and Engineering Center – RailTEC
Department of Civil and Environmental Engineering
University of Illinois at Urbana-Champaign, Urbana, IL 61801
Email: cbarkan@illinois.edu

Word Count: 3066 words + 1 table (250 words per table) = 3,316 words

Submitted August 1, 2019

ABSTRACT

Freight train derailment rates have declined considerably over the past decade; however, when they do occur, they can damage infrastructure and rolling stock, interfere with transportation, and may cause casualties and releases of dangerous goods. Derailment analysis and prevention has long been a high priority for the rail industry and policy makers. Therefore, it is imperative to investigate both train derailment severity and consequences. Most previous research has focused on analysis of factors affecting the probability of train derailments, but less research has been conducted on understanding factors affecting derailment severity estimation. This paper aims to identify variables affecting derailment severity and develop a new methodology for estimation. Extensive, detailed data from the U.S. Department of Transportation Federal Railroad Administration (FRA) were used to construct and compare several machine learning models for train derailment severity prediction, including multiple regression, KNN, regularized regression (ridge and LASSO), random forest, and gradient boosting. The results can be used to provide insights for the industry to understand factors affecting derailment severity and enable better informed approaches to risk reduction.

Keywords: Freight train, derailment, accident, severity, regression, machine learning, prediction, risk

INTRODUCTION AND LITERATURE REVIEW

Freight train derailment rates have declined considerably over the past decade; however, when they do occur, they can damage infrastructure and rolling stock, interfere with operating schedules, and may cause casualties and release hazardous materials. Therefore, derailment prevention is a high priority for the rail industry and government regulators, and understanding the most important factors contributing to derailment severity is critical to effective allocation of safety and risk management resources. Most previous research has focused on analysis of factors affecting the probability of train derailments, whereas relatively less has focused on analysis of factors affecting derailment severity. Although derailments have a low probability of occurrence, they have the potential for severe consequences. Quantifying derailment severity and identifying the associated factors can provide insights for the rail industry and government to inform and prioritize safety improvement strategies.

Two general approaches to estimating derailment severity have been developed in the past: computer simulation and statistical analysis. Simulation models are based on nonlinear wheel-rail interaction models. Yang et al (1, 2) developed a simulation model and identified position of the first car involved and derailment speed as factors affecting severity (in this paper, first derailed vehicle (FDV) will be used to refer to the position of the first derailed car or locomotive in a train). Coppens et al. (3) and Birk et al. (4) developed a derailment accident computer simulation model (DERACS), that predicts the number of cars derailed based on inputs such as train speed, number of cars in the consist, ground friction, and braking rate.

A second method for derailment severity modelling is statistical analysis. Nayak et al. (5) first developed analytical equations for derailment severity estimation using Federal Railroad Administration (FRA) data on track class, accident cause, and train speed. Saccomanno and El-Hage (6, 7) developed a truncated geometric model to estimate the relationship between number of cars derailed and derailment speed. Studies have also found that residual train length affects the number of cars derailed (8, 9). Residual length or residual first derailed vehicle (RFDV) is defined as the number of cars or locomotives behind the first derailed vehicle. Liu et al. (10) used a negative binomial regression model to estimate the conditional mean of train derailment severity and found that the proportion of loaded cars also affects derailment severity.

The number of cars derailed has been used as a metric for derailment severity (5, 6, 10-12). In this context, the term "cars" indicates railcars, locomotives, and cabooses. The FRA accident database provides another metric that can be used to measure severity: the cost of damage to track and equipment. A variation on this is the monetary damage ratio, where the total monetary damage is divided by the FRA incident reporting threshold, which is adjusted periodically to account for inflation (13). Other alternatives include multivariate applications with multiple response variables, such as using both number of cars derailed and total monetary damage together. However, if they are highly correlated, it is more difficult to identify the differences between the dependent variables.

The following response variable distributions were compared: cars derailed, total monetary damage, and damage ratio. The skewness for cars derailed, total monetary damage, and damage ratio was 2.36, 3.17, 2.91, respectively. Due to their skewness, a log transformation was used to compare the three variables and their respective skewness was 0.03, -0.27, and 0.23. Number of cars derailed had the lowest skew as a response variable. In addition, recorded monetary damage can be influenced by other factors that do not reflect severity or energy of the derailment; including the cost difference between locomotives and freight cars and the cost difference between repairing open track versus turnouts and other special trackwork (10, 14). Therefore, in this paper number of cars derailed is used as the derailment severity metric because it is believed to be a reasonable measure of the forces involved in an accident.

RESEARCH OBJECTIVES

There are three principal objectives for the research described in this paper. First, the major covariates affecting derailment severity will be investigated using the most current data. Next, a supervised machine learning approach for estimating derailment severity will be introduced. Lastly, those models will be used to investigate features affecting derailment severity.

DATA AND METHODOLOGY

The database used in this research is the U.S. Department of Transportation (DOT) FRA Rail Equipment Accident or Incident (REA) database (15). The FRA REA database contains all accidents or incidents that exceed a specified monetary threshold for damages to infrastructure and equipment (16). The FRA periodically adjusts the reporting threshold to account for inflation; over the period covered in this study it increased from $7,700 in 2006, to $10,700 in 2017 (13). This database contains information for each accident, including features such as cause, type, date, location, accident speed, number of cars derailed, and numerous other details (16). Class 1 railroad mainline freight train derailments from 2005 to 2017 were used in this paper.

INVESTIGATION OF FACTORS AFFECTING DERAILMENT SEVERITY

As discussed in the introduction, several features contribute to freight train derailment severity, including speed, FDV, RFDV, and train length. The first part of this research was to examine these features and their effect on derailment severity. Figure 1 shows the relationship between average severity and each of these four features. Residual train length and derailment speed display a strong linear relationship and have a high r-squared value. However, one should be cautious when conducting regression analysis using binned data as it has the potential to lose statistical power, or fail to detect non-linear relationships.

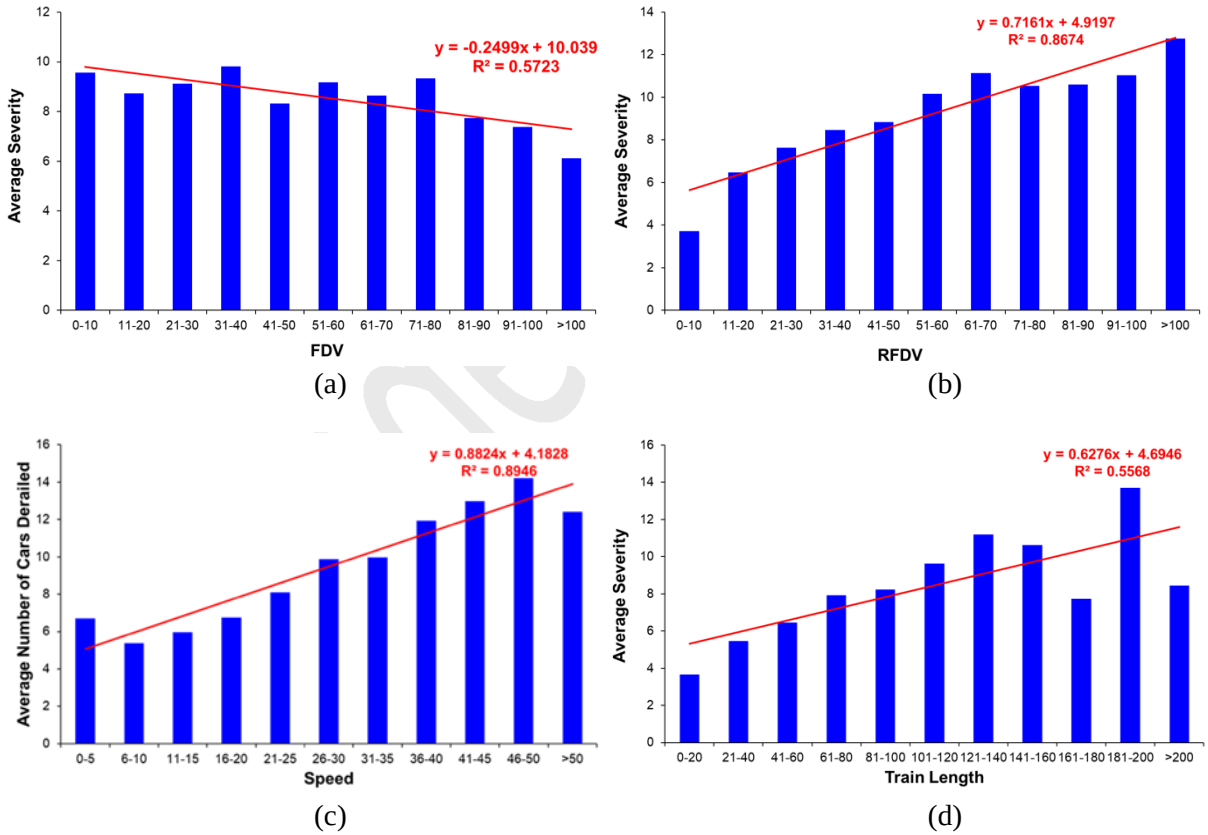

Figure 1 (a) Average Severity vs RFDV (b) Average Severity vs FDV (c) Average Severity vs Derailment Speed (d) Average Severity vs Train Length (Class 1 mainline freight train derailments, 2005 – 2017)

Next, statistical modeling was used to construct a multiple regression model with these features:

$$Severity = 0.178\ speed + 0.077\ length - 0.066\ FDV - 0.213 \quad (1)$$

The multiple regression model identified speed, length, and FDV as significant variables; however, the model had a low R-squared value of 0.30.. The multiple regression results suggested that the relationship between some variables and the response variable was non-linear. Therefore, the next part of the analysis focused on machine learning techniques to examine all the factors and: identify the ones most strongly affecting severity, compare different models in terms of prediction accuracy, and identify and characterize any non-linear relationships between the features and severity.

MODEL DEVELOPMENT

Framework and data pre-processing

As discussed above, two principal methods of derailment severity estimation have been used in previous research. A new method introduced in this paper theuse of machine learning that focuses on feature identification and prediction. Machine learning is an algorithmic approach in which the model learns from the data instead of focusing on covariates. In contrast to traditional statistical modeling, a machine learning model does not model the data generating process but instead focuses on how the data affecting the response variable (17). Machine learning optimizes predictive performance by learning from the data. The difference between statistical and machine learning models is their emphasis. Statistical models emphasize inference and machine learning models emphasize prediction. The commonly used machine learning methods include supervised learning, unsupervised learning, and reinforcement learning.

Supervised learning means the data contains a label or response variable, which can be used to judge the algorithm's performance. On the other hand, unsupervised learning attempts to find relationships between data points and there are no response variables; some common ones are clustering and anomaly detection. Reinforcement learning has artificial intelligence agents working on optimized path or strategies for a specific task, such as moves in a chess game. Supervised learning is used in this paper because of the labelled response variable and it can also be categorized into two methods: regression and classification. Regression is used when the response variable is continuous, and classification is used when the response variable needs to be classified into different classes. This paper will implement supervised learning that involves a known response variable with labeled data.

The general machine learning pipeline follows four steps: data pre-processing, feature engineering, model selection and parameterization, and prediction. This research included development of a raw data clean up, data transformation, and feature engineering procedure for the FRA REA database for derailment severity prediction. Five machine learning models were implemented and used for prediction, the features contributing the most to derailment severity were also identified.

The first step involved data cleaning and pre-processing to capture any missing, skewed, or incorrectly formatted data. A process to import and clean raw data from the FRA REA database was developed, and features with identical information were removed including extra features for year, month, and date. Certain features that were all blank or zero were also removed, including dummy fields, empty narrative, and alcohol or drug use. Some features such as traffic density and monetary damage included unusually large values that were considered outliers. This was resolved by investigating the frequency distribution and official reports including from the National Transportation Safety Board. Due to the possibility of certain large accidents having exceptionally high damages, a data point was only removed when an official report indicated it was anomalous.

Feature engineering is the process of transforming data into representable feature vectors (18). This included creating new variables using domain knowledge, such as signal information, region, and train consist characteristics. Method of operation indicates whether a section of track has electric track circuits and wayside signals or not (19). Prior to May 31, 2011, the FRA recorded method of operation in 12 categories; after that date, FRA (19) collapsed the categories into just two, either signalled or non-signaled, consequently, method of operation was treated as a binary categorical variable in this research. Categorical variables were also transformed into numeric values using one-hot encoding (20).

Model selection and parametrization

Five different supervised machine learning algorithms were used to find the best fitting model for the cleaned and pre-processed derailment data. The models implemented are K-nearest neighbors (KNN), ridge regression, Least Absolute Shrinkage and Selection Operator (LASSO) regression, random forest, and Extreme Gradient Boosting (xgboost). The dataset was randomly partitioned into a 70% training set and a 30% test set. The training set was used to fit the models and the test set used to validate the prediction accuracy (21). When a model produces good results with the training set but poor results with the test set, it is overfitting. If a model has a high error rate on the training set and cannot generalize to the test set, it is underfitting. Each of the five methods were tested to see which one had the highest predictive power. The feature importance was also examined using the most effective model.

The baseline model for this research is nearest neighbour. KNN is a nonparametric method that identifies the k point(s) closest to x with the averages of nearby observations. A distance metric is used to calculate nearby points. The prediction at a given target point x is shown in equation 2.

$$\hat{y} = \frac{1}{k} \sum_{x_i \in N_k(x)} y_i \tag{2}$$

Where:

k = number of samples

$N_k(x)$= k samples from training data closest to x, based on distance metric

There are different distance metrics including Euclidean distance, Manhattan distance, Cosine distance, and Mahalanobis distance (20). Manhattan distance is the most commonly used and is the distance between two vectors. Mahalanobis distance (d) is defined below and takes covariance into account, and considers the correlations of the data.

$$d(\vec{x}, \vec{y}) = \sqrt{(\vec{x} - \vec{y})^T S^{-1} (\vec{x} - \vec{y})} \tag{3}$$

Where:

S = covariance matrix of vector x and vector y

The input parameter, k is the neighborhood size and can be adjusted for different purposes. Usually, when k is small, the model has a high variance but lower bias. On the other hand, a large k has lower variance but much more bias. k was initialized as the square root of the total number of features for regression models and a value of 58 was used after parameter tuning.

Ridge regression is a shrinkage method that regularizes the coefficient estimates of a linear model (20). A shrinkage penalty (l2 norm) is incorporated to reduce variance and multicollinearity by regularizing the coefficients. Shrinkage may introduce small bias but the reduction in variance leads to an overall smaller error. The tuning parameter lambda was selected using cross-validation on the training set, and the lambda that produced the lowest error on the training set was used.

$$\hat{\beta} = argmin_\beta \sum_i (y_i - {x_i}^T \beta)^2 + \lambda \sum_{j=1}^{p} {\beta_j}^2 \tag{4}$$

Where:

$\widehat{\beta}$ = ridge estimator

λ = shrinkage parameter

p = number of independent variables

X = design matrix with dimension n x p

${x_i}^T \beta$ = linear predictor

LASSO regression, also known as L1 penalty, is a similar method to ridge regression that performs variable selection and regularization to enhance the prediction accuracy and interpretability of the model. However, while ridge regression shrinks the coefficients towards zero and does not result in exclusion of any coefficients, lasso regression shrinks selected variables to zero and excludes them from the model. This adds more interpretability to the final model. The L-1 regularization is an added penalty for having larger weights for coefficients. Similar to ridge regression, the lambda minimum was found using cross-validation.

$$\hat{\beta} = argmin_{\beta} \sum_{i} (y_i - {x_i}^T \beta)^2 + \lambda \sum_{j=1}^{p} |\beta_j| \quad (5)$$

Where:

$\widehat{\beta}$ = lasso estimator
λ = shrinkage parameter
p = number of independent variables
X= design matrix with dimension n x p
${x_i}^T \beta$ = linear predictor

Random forest is an aggregation of a collection of tree-structured models (i.e. decision trees) (22). It is an ensemble learning method that can be used for regression and classification. At each node, the algorithm randomly selects a subset of features from the full set of features and each tree is built on a subset of observations. This reduces correlation between trees and avoids overfitting. Random forest has been found to be robust to outliers, easily parallelized, and consistently performs well (22). For regression, the number of variables randomly selected at each split is based on the total number of variables divided by three. The number of trees selected was 400, to ensure each input row gets predicted more than once.

Extreme Gradient Boosting (xgboost) is a gradient boosting framework introduced by Friedman (23). It can be applied to many machine learning methods for both regression and classification (20). It is an ensemble of many weaker and smaller models such as decision trees. Xgboost is one implementation of a gradient boosting technique and uses regularization to control overfitting. Cross-validation was used to select the optimal number of iterations for the model.

RESULTS

Performance metric and feature importance

For regression models, there are several evaluation metrics including mean absolute error (MAE), mean squared error (MSE), and root mean squared error (RMSE). MAE measures the absolute magnitude of the errors and is useful for certain cases; however, it should be used with caution due to the absolute value. MSE calculates the mean squared difference between predicted values and actual values. A single prediction with large error would contribute greatly to the MSE, so RMSE is used instead. It is a quadratic metric that is the square root of the MSE and provides better interpretability and a more suitable measure of goodness of fit.

$$RMSE = \sqrt{\frac{1}{n}\sum_{j=1}^{n}(y_j - \hat{y}_j)^2} \quad (6)$$

Where,

y_j = observed

$\hat{y}_j$ = predicted

The RMSE was calculated for each model (Table 1) and random forest and xgboost outperformed the other three models by the largest margin. Although those two models outperformed the others, the first three models provided greater interpretability.

TABLE 1 Comparison of Model Results

Model	RMSE
KNN	1.411
Ridge	1.106
LASSO	1.074
Random Forest	0.956
xgboost	0.950

Feature or variable importance describe how much each feature contributes to the prediction accuracy (24). Different models select different covariates, and this is known as the "Rashomon" effect (17). Random forest is one of several approaches that provide model-agnostic feature importance and it was used for this research. In random forest models, feature importance is measured by a mean reduction in impurity (classification) or mean reduction in residual sum of squares (regression) (22).

The default random forest feature importance in R and Python show a bias towards correlated predictor variables (25). This is due to the increase in splits from continuous and high cardinality variables, making them more likely to be selected as the important variables. Strobl et al. suggested using permutation importance over the default method. Permutation importance is a method that permutes values for each feature and examines whether it changes the prediction error. This can decide if a variable has a positive effect on the prediction accuracy. Lastly, if high collinearity exists, conditional permutation importance can be used (25).

Figure 2 indicates that the following features contribute most to derailment severity: residual FDV, position of causing car, track-related causes, derailment speed, maximum speed, and gross tonnage. Some other features identified include mechanical and human factor causes.

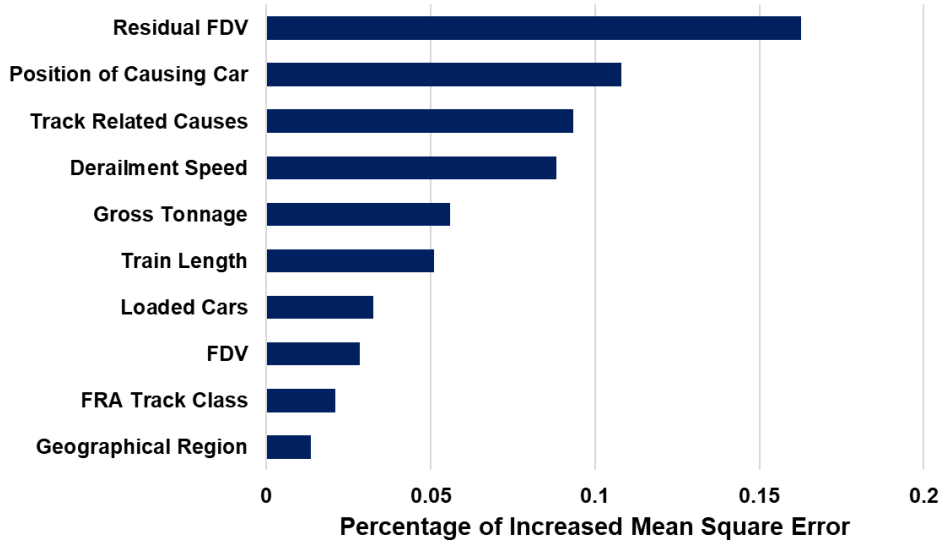

Figure 2 Feature Importance

CONCLUSIONS AND DISCUSSION

This paper investigated the U.S. Class I railroad mainline freight train derailment severity using the FRA REA database. Three derailment severity metrics were compared, and number of cars derailed was determined as the most suitable for derailment severity estimation. Four features affecting derailment severity were analyzed and a multiple regression was conducted using those features.

A methodology for applied supervised machine learning algorithms to predict derailment severity was introduced. Machine learning techniques have been underutilized in railroad safety research. This paper provides a complete data pre-processing and model building pipeline for freight train derailment analysis. Traditional statistical analysis provides good interpretability, but with certain costs, mainly the limited feasible number of features, model assumptions, and robustness. The applied machine learning framework described here allows researchers to work with more features and to better understand and evaluate their effect on derailment risk assessment and management.

This proactive approach identifies critical features affecting derailment severity using these algorithms. The permutation-based feature selection method used allows accurate identification of the most important features affecting derailment severity. This will provide decision makers with a means of more efficiently allocating resources to prevent derailments with the greatest risk. For example, track related causes have emerged as a leading factor affecting severity. Practitioners aiming to reduce derailment severity on a network could consider focusing proportionately more resources on inspection or maintenance activities to prevent these types of accidents. Another aspect of this are two new features found to affect severity, loaded cars and geographical region. Factors such as these are intrinsic to transportation services and cannot be avoided. The methodology described here enables them to be analyzed as part of the overall route planning and network risk management process.

ACKNOWLEDGMENTS

Support for this research was provided by the Association of American Railroads, BNSF Railway, and the National University Rail Center, a US DOT OST Tier 1 University Transportation Center. This paper is solely the work of the authors and does not necessarily reflect the opinions of the sponsors.

AUTHOR CONTRIBUTIONS

The authors confirm contribution to the paper as follows: study conception and design: B. Wang, C. Barkan; data collection: B. Wang; analysis and interpretation of results: B. Wang, J. Kim, C. Barkan; draft manuscript preparation: B. Wang. J. Kim, C. Barkan. All authors reviewed the results and approved the final version of the manuscript.

REFERENCES

1. Yang, T.H., Manos, W.P., Johnstone, B. *A Study Continuation of Derailment Behavior Final Report (Phase 08 Report on Computer Derailment Study).* RPI/AAR Report RA-08-1-12 (R-135). Railroad Tank Car Safety Research and Test Project, Association of American Railroads, 1972.

2. Yang, T.H., Manos, W.P., Johnstone, B. Dynamic Analysis of Train Derailments. Rail Transportation Proceedings, The American Society of Mechanical Engineers, New York, 1973. 72-WA/RT-6, p.8.

3. Coppens, A.J., Wong, J.D.E., Bibby, A., Birk, A.M. and Anderson, R.J. *Development of a Derailment Accident Computer Simulation Model*. Transport Canada Report No. TP 9254E. Transportation Development Centre and Transport of Dangerous Goods, Ottawa, Canada, 1988.

4. Birk, A.M., Anderson, R.J. and Coopens, A.J. A Computer Simulation of a Derailment Accident, Parts I & II. *Journal of Hazardous Materials*, 1990. Volume: 25, pp. 121-147 & 149-165.

5. Nayak, P.R., D.B. Rosenfield, and J.H. Hagopian. *Event Probabilities and Impact Zones for Hazardous Materials Accidents on Railroads*. Report DOT/FRA/ORD-83/20, Federal Railroad Administration, U.S. Department of Transportation, 1983.

6. Saccomanno, F.F., and S.M. El-Hage. Minimizing Derailments of Railcars Carrying Dangerous Commodities through Effective Marshaling Strategies. *Transportation Research Record: Journal of the Transportation Research Board*, 1989. Volume:1245, pp.34-51.

7. Saccomanno, F.F., and S.M. El-Hage. Establishing Derailment Profile by Position for Corridor Shipments of Dangerous Goods. *Canadian Journal of Civil Engineering*, 1991. Volume: 18, pp 67-75.

8. Anderson, R.T., and C.P.L. Barkan. Derailment Probability Analyses and Modeling of Mainline Freight Trains. Proceedings of the 8th International Heavy Haul Railway Conference, International Heavy Haul Association, Rio de Janeiro, Brazil, 2005.

9. Bagheri, M. Risk-Based Model for Effective Marshalling of Dangerous GoodsRailway Cars. Doctorate Thesis. University of Waterloo, Ontario. 2009.

10. Liu, X., M.R. Saat, X. Qin, and C.P.L. Barkan. Analysis of U.S. Freight-Train Derailment Severity using Zero-Truncated Negative Binomial Regression and Quantile Regression. *Accident Analysis and Prevention*, 2013. Volume: 59, pp.97-93.

11. Barkan, C.P.L., C.T. Dick, and R.T. Anderson. Analysis of Railroad Derailment Factors Affecting Hazardous Materials Transportation Risk. *Transportation Research Record: Journal of the Transportation Research Board*, 2003. Volume: 1825, pp. 64-74.

12. Anderson, R.T. and C.P.L. Barkan. Railroad Accident Rates for Use in Transportation Risk Analysis. *Transportation Research Record: Journal of the Transportation Research Board*, 2004. Volume:1863, pp. 88-98.

13. Federal Railroad Administration. *Railroad Equipment Accident/Incident Reporting Threshold 1957-2017*. U.S. Department of Transportation, 2017.

14. Dick, C.T. Multivariate Statistical Model for Predicting Occurrence and Location of Broken Rails. *Transportation Research Record: Journal of the Transportation Research Board*, 2003. Volume: 1825, pp. 48-55.

15. *FRA Office of Safety Analysis Accident Database*. Federal Railroad Administration, U.S. Department of Transportation. https://safetydata.fra.dot.gov/OfficeofSafety/publicsite/on_the_fly_download.aspx. Accessed Jan 20, 2019.

16. Federal Railroad Administration. *Guide for Preparing Accident/Incident Reports*. U.S. Department of Transportation, 2011.

17. Breiman, L. Statistical Modeling: The Two Cultures. *Statistical Science*, 2001. Volume: 16, No.3, pp. 199 - 215.

18. *Machine Learning Crash Course*. Google Developers, Google. https://developers.google.com/machine-learning/crash-course/. Accessed Jan 30, 2019.

19. *Code of Federal Regulations, Title 49, Parts 236*. Federal Railroad Administration, U.S. Department of Transportation, 2011.

20. James, G., D. Witten, T. Hastie, and R. Tibshirani. *An Introduction to Statistical Learning: with Applications in R*. Springer, New York, NY, 2013.

21. Hastie, T., R. Tibshirani, and J. Friedman. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction, Second Edition*. Springer, New York, NY, 2009.

22. Breiman, L. Random Forests. *Machine Learning*, 2001. Volume: 45, No.1, pp. 5–32.

23. Friedman, J.H. Greedy Function Approximation: A Gradient Boosting Machine. *Annals of Statistics*, 2001. pp. 1189–1232.

24. Fisher, Aaron, C. Rudin, and F. Dominici. J.H. Model Class Reliance: Variable Importance Measures for any Machine Learning Model Class, from the "Rashomon" Perspective, 2018. http://arxiv.org/abs/1801.01489.

25. Strobl, Carolin, A-L. Boulesteix, T. Kneib, T. Augustin, and A. Zeileis. Conditional Variable Importance for Random Forests. *BMC Bioinformatics*, *2008.* Volume: 9: 307.