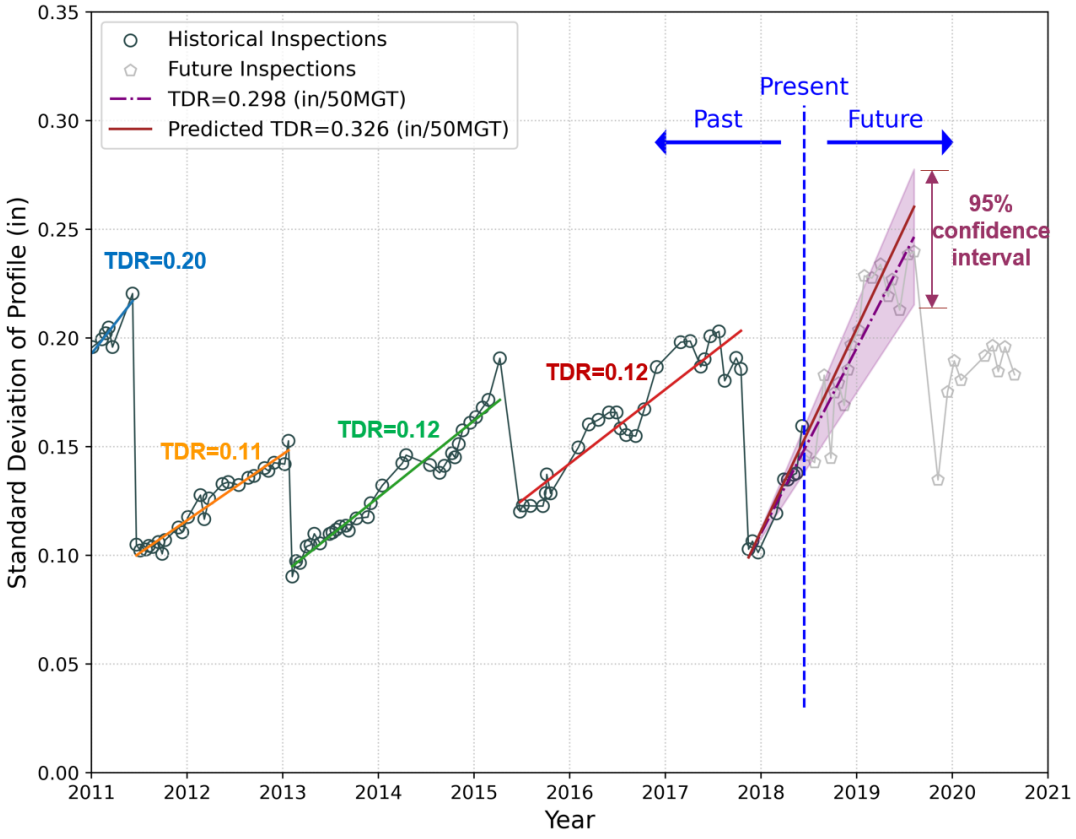
U.S. Department
of Transportation
Federal Railroad
Administration

Office of Research,
Development and Technology
Washington, DC 20590

Machine Learning-based Analysis for Railway Track Maintenance



NOTICE

This document is disseminated under the sponsorship of the Department of Transportation in the interest of information exchange. The United States Government assumes no liability for its contents or use thereof. Any opinions, findings and conclusions, or recommendations expressed in this material do not necessarily reflect the views or policies of the United States Government, nor does mention of trade names, commercial products, or organizations imply endorsement by the United States Government. The United States Government assumes no liability for the content or use of the material contained in this document.

NOTICE

The United States Government does not endorse products or manufacturers. Trade or manufacturers' names appear herein solely because they are considered essential to the objective of this report.

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 0704-0188

The public reporting burden for this collection of information is estimated to average 1 hour per response, including the time for reviewing instructions, searching existing data sources, gathering and maintaining the data needed, and completing and reviewing the collection of information. Send comments regarding this burden estimate or any other aspect of this collection of information, including suggestions for reducing the burden, to Department of Defense, Washington Headquarters Services, Directorate for Information Operations and Reports (0704-0188), 1215 Jefferson Davis Highway, Suite 1204, Arlington, VA 22202-4302. Respondents should be aware that notwithstanding any other provision of law, no person shall be subject to any penalty for failing to comply with a collection of information if it does not display a currently valid OMB control number.
PLEASE DO NOT RETURN YOUR FORM TO THE ABOVE ADDRESS.

1. REPORT DATE *(DD-MM-YYYY)*	**2. REPORT TYPE**	**3. DATES COVERED** *(From - To)*
30 September 2026	Technical Report	July 2022 – July 2024

4. TITLE AND SUBTITLE

Machine Learning-based Analysis for Railway Track Maintenance

5a. CONTRACT NUMBER
FR22RPD31000000010

5b. GRANT NUMBER

5c. PROGRAM ELEMENT NUMBER

6. AUTHOR(S)

Saeed Goodarzi: ORCiD #0000-0003-0250-6948
Carlton L. Ho: ORCiD #0000-0002-1510-4369
Kevin Kashani: ORCiD #0009-0007-9772-4951
James P. Hyslip: ORCiD #0009-0005-4418-159X

5d. PROJECT NUMBER

5e. TASK NUMBER

5f. WORK UNIT NUMBER

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)

University of Massachusetts Amherst
140 Natural Resources Way
Amherst, MA 01003

Loram Technologies, Inc.
111 Cooperative Way #400
Georgetown, TX 78626

8. PERFORMING ORGANIZATION REPORT NUMBER

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)

U.S. Department of Transportation
Federal Railroad Administration
Office of Research, Development, and Technology
1200 New Jersey Avenue, SE
Washington, DC 20590

10. SPONSOR/MONITOR'S ACRONYM(S)

11. SPONSOR/MONITOR'S REPORT NUMBER(S)
DOT/FRA/ORD-26/28

12. DISTRIBUTION/AVAILABILITY STATEMENT

This document is available to the public through the FRA website.

13. SUPPLEMENTARY NOTES

COR: Hugh Thompson

14. ABSTRACT

Understanding and mitigating track degradation rates (TDRs) is crucial for ensuring the safety and efficiency of railway systems, as well as for controlling maintenance costs and preserving ride quality. The Federal Railroad Administration (FRA) sponsored the University of Massachusetts Amherst and Loram Technologies, Inc. to extensively analyze TDR over 280 miles of passenger revenue track using data from 2011 to 2021. Using a large dataset was pivotal, as it enhanced the robustness of the analysis, allowing for more accurate and reliable conclusions. The researchers used ground penetrating radar (GPR) and light detection and ranging (LiDAR) along with geometry data to examine subsurface and drainage conditions of the tracks. The project employed two distinct methods to estimate TDR, each offering unique insights and applicability based on available historical data and desired precision. These complementary approaches provide the industry with versatile tools for estimating TDR and improving maintenance scheduling, emphasizing their reliability and adaptability.

15. SUBJECT TERMS

Railroad, machine learning, large data, drainage condition, GPR, LiDAR

16. SECURITY CLASSIFICATION OF:			**17. LIMITATION OF ABSTRACT**	**18. NUMBER OF PAGES**	**19a. NAME OF RESPONSIBLE PERSON**
a. REPORT	**b. ABSTRACT**	**c. THIS PAGE**		54	**Hugh Thompson**
					19b. TELEPHONE NUMBER *(Include area code)*

Acknowledgments

The authors extend their gratitude to Hugh Thompson from the Federal Railroad Administration and Ted Sussmann from the Volpe National Transportation Systems Center for their invaluable support and guidance throughout this project. They also thank Mike Trosino and Amanda Kessler from Amtrak for their collaborative support in providing the data for the analysis in this research.

Contents

Executive Summary 1

1. Introduction 2

1.1 Background 3
1.2 Objectives 3
1.3 Overall Approach 4
1.4 Scope 4
1.5 Organization of the Report 4

2. Stochastic Approach for Estimating Probability of Track Degradation Rate 5

2.1 TDR 5
2.1.1 Geometry Data 5
2.1.2 Data Preprocessing 5
2.1.3 Calculating TDR 5

2.2 Categorizing the Dataset 7
2.2.1 Categorizing Based on Subsurface Condition 8
2.2.2 Categorizing Based on Cross Section and Drainage Conditions 11

2.3 Stochastic Models 15
2.3.1 Fitting Probability Distributions 17
2.3.2 Effect of Subsurface Conditions 20
2.3.3 Effect of Drainage and Cross-section Conditions 21

3. Predicting Value of TDR Using ML Models 23

3.1 Extracted Features in a Block 23
3.2 Effect of Ditch Properties on TDR 24
3.2.1 Effect of Depth of Ditch on TDR 25
3.2.2 Effect of Distance of Ditch on TDR 26
3.2.3 Effect of Lateral Ditch Condition on TDR 26
3.2.4 Effect of Longitudinal Ditch Condition on TDR 28

3.3 Machine Learning Models 30
3.3.1 Random Forest 30
3.3.2 Support Vector Regression 30
3.3.3 XGBoost 31
3.3.4 Artificial Neural Network 32

3.4 Prediction Models 32
3.4.1 Performance Metrics 32
3.4.2 Splitting Dataset 33
3.4.3 Fine-tuning Hyperparameters 33

3.4.4 Results of Predicting TDR 34
3.4.5 Feature Importance 37
4. Conclusion and Publications 39
4.1 Major Conclusions 39
4.2 Journal Reports and Conference Papers 40
4.3 Recommendations for Future Research 40
5. References 41
Appendix A. The Exceedance Probability (i.e., CCDF) for Intervals of TDR in Each Category 44
Abbreviations and Acronyms 46

Illustrations

Figure 1. An example of studied profile geometry data of a block: (a) SD of profile for different geometry inspections vs. time and dates of collecting GPR, (b) fitted linear regression models to calculate track degradation rates. 6

Figure 2. Example BFI data block. The final value of BFI in a block strongly depends on the summarizing approach. Blue: BFI=0, red BFI=45. 8

Figure 3. The correlation plot using Pearson equation; violet: Pearson Correlation=0, yellow: Pearson correlation=45. 9

Figure 4. Box plot of track degradation rate with respect to subsurface conditions; (a) short period degradation; (b) long period degradation. 10

Figure 5. Flowchart for evaluating drainage and cross section condition from LiDAR data. 12

Figure 6. Applications of the proposed algorithm for determining cross-section and drainage condition of tracks; (a) typical cut cross-section (poor drainage condition) for all three tracks; (b) typical AGO cross section (moderate drainage condition); (c) typical fill cross section (good drainage condition); (d) cut cross-section for right terrain and AGO cross-section for left terrain due to existence of a good ditch. 14

Figure 7. Box plot of track degradation rate with respect to drainage conditions. 15

Figure 8. Flowchart diagram of approach for stochastic analysis of track degradation rate. 16

Figure 9. Histograms and the best fitted distributions of categories from clean ballast (G1) to highly fouled ballast (G4) and AGO cross-section. 18

Figure 10. CDF of actual data and the best fitted distribution of categories from clean ballast (G1) to highly fouled ballast (G4) and AGO cross-section. 19

Figure 11. Q-Q and P-P plots of category “G3 and AGO,” which has the lowest K-S p-value among categories. 20

Figure 12. Effect of subsurface conditions on track degradation rates. 21

Figure 13. Effect of drainage and cross-section conditions on track degradation rates. 22

Figure 14. An example of SD vs time for a block. The goal of prediction models is to predict the TDR in current cycle, which is the slope of dotted purple line. 23

Figure 15. Cross section of a block as an example to show depth and distance of ditch; Distance of ditch is minimum value between D1 and D2. 24

Figure 16. Effect of ditch depth on TDR. 25

Figure 17. Effect of ditch distance from center of track on TDR. 26

Figure 18. The boxplot of TDR considering both distance and depth of ditch; the effect of ditch distance is more significant when the ditch is shallow. 27

Figure 19. Comparing TDR for good and poor lateral drainage quality across different percentiles; the ratio of TDR for poor and good lateral drainage is shown by maroon “x” mark and is around 1.55 for different percentiles. 27

Figure 20. Schematic representation of longitudinal condition of ditch; in poor longitudinal drainage, water accumulates in the ditch, while in good longitudinal drainage, water flows freely. 28

Figure 21. Topography contour lines at a milepost; the ditch location is automatically determined using LiDAR data and depicted by a green line. Note that this figure is not a geospatial map, but rather an indication of elevations at a distance from the centerline of track. It is also important that this figure is independent of the variations in alignment of the track.... 29

Figure 22. The effect of longitudinal drainage condition on TDR. 29

Figure 23. Fine-tuning the number of trees and maximum depth for the XGBoost model. 34

Figure 24. Scatter plot of predicted TDR vs. actual TDR values using XGBoost prediction model..... 35

Figure 25. Concentration of data points for predicted TDR vs actual TDR in the test dataset, utilizing the XGBoost prediction model. 36

Figure 26. Histograms of residuals for both the training and testing datasets. 37

Figure 27. The application of the proposed ML model on a block as an example. 37

Figure 28. The importance of features for predicting TDR using the XGBoost model. 38

Tables

Table 1. Categorizing dataset based on subsurface condition of track 10

Table 2. Statistical values of track degradation rate in each category. The distribution is represented with the minimum value, median value, maximum value, standard deviation, mean value skewness and kurtosis. 16

Table 3. Goodness-of-fit values and parameters of the best fitted distribution on each category. 17

Table 4. List of Extracted Features for Predicting TDR. 24

Table 5. Statistical performance of developed models for predicting TDR. 34

Executive Summary

Understanding and mitigating track degradation rates (TDRs) is crucial for ensuring the safety and efficiency of railway systems, as well as for controlling maintenance costs and preserving ride quality. The Federal Railroad Administration (FRA) sponsored the University of Massachusetts Amherst and Loram Technologies, Inc. to extensively analyze TDR data for over 280 miles of passenger revenue track using data from 2011 to 2021. Using a large dataset was pivotal, as it enhanced the robustness of the analysis, allowing for more accurate and reliable conclusions. The researchers used ground penetrating radar (GPR) and light detection and ranging (LiDAR) along with geometry data to examine subsurface and drainage conditions of the tracks. The project employed two distinct methods to estimate TDR, each offering unique insights and applicability based on available historical data and desired precision. These complementary approaches provide the industry with versatile tools for estimating TDR and improving maintenance scheduling, emphasizing their reliability and adaptability.

The first method involved segmenting 35,000 track blocks into 12 categories based on ballast fouling index and external drainage conditions. The research team fitted 65 probability distributions (e.g. Lognormal, F-distribution, and Dagum) to each category and determined the best-fitted distribution for each using goodness-of-fit tests like Kolmogorov-Smirnov (K-S) and Cramér Von-Mises (CvM). This stochastic approach allowed the team to estimate the probability that actual TDR exceeded an allowable value. Due to the large number of blocks, researchers proposed an algorithm to evaluate the external drainage condition of the track using LiDAR data. The results of the stochastic analysis indicate that the probability of experiencing high TDR in highly fouled ballast is 10 times greater than in clean ballast, highlighting the detrimental effects of fouling on track behavior.

The second approach is more accurate than the first approach, but requires more historical data. In this method, researchers use machine learning (ML) models to predict the value of TDR directly, rather than estimating the exceedance probability over an allowable value. This approach involves extracting 12 features for TDR prediction, including track subsurface condition, ditch properties, and geometry history. The project team employed four ML models – Random Forest, Support Vector Regressor (SVR), XGBoost, and Artificial Neural Network (ANN) – to predict TDR values. The results indicated that XGBoost slightly outperformed the others, successfully predicting TDR with a mean absolute error (MAE) of 0.019 and a root mean squared error (RMSE) of 0.029 in/50 MGT.

The research findings highlighted the critical influence of lateral drainage conditions on TDR. Tracks with poor lateral drainage, characterized by ditches shallower than 2.5 ft (76 cm) and located over 15 ft (4.57 m) away, experienced TDRs approximately 55 percent higher than those with optimal drainage conditions. While longitudinal ditch properties also impacted TDR, their effect was less significant compared to lateral drainage. These insights underscore the importance of maintaining ditches in good condition, especially considering the detrimental effects of high TDR on track longevity and safety.

1. Introduction

The Federal Railroad Administration (FRA) sponsored the University of Massachusetts Amherst and Loram Technologies, Inc. to extensively analyze track degradation rates (TDR) using a large dataset of geometry data from 2011 to 2021 with a cumulative length of 287 miles of passenger tracks. This research focused on employing probabilistic and machine learning (ML) approaches to evaluate the need for railroad maintenance based on track conditions. Researchers used a numerical approach, constructed from regularly measured data associated with track geometry, ground penetrating radar (GPR) and light detection and ranging (LiDAR). The model assesses the correlations in probabilistic terms; thus, the means to properly assess functional uncertainties are a critical part of the proposed approach. As probabilistic relationships rely on collected data, the size of datasets correlates with the accuracy of probabilistic relationships and, thus, ML models.

To safely operate, railroads must maintain all their assets. The efficiency of ballasted track systems depends on having a means to implement remedial measures (e.g., maintenance, repair, and replacement) to ensure safe operation and allow for optimal traffic. Railway track maintenance can be categorized as either corrective maintenance or preventive maintenance [1]. Corrective maintenance rectifies existing defects that result in track closure or speed restriction. This type of maintenance has high cost due to sudden failure and system recovery. Preventive maintenance predicts defects and maintains them before they surpass safety thresholds. Preventive maintenance increases track safety by minimizing defects, and reduces the potential for track closure or speed limit [2].

Railroads often use geometry cars to collect a wide variety of data, including measurements of the car dynamics and track geometry. Less frequently, these cars evaluate the condition of the track system using GPR and LiDAR. These data can be used to develop a way to evaluate maintenance needs based upon a probabilistic analysis.

The project team developed two innovative methods to accurately estimate TDR. The first approach involved a detailed analysis of 35,000 track blocks, categorized into 12 distinct groups based on critical factors such as the ballast fouling index (BFI) and external drainage conditions, each representing a potential impact on TDR. For each of these groups, researchers applied a comprehensive set of 65 probability distributions (e.g., Lognormal, F-distribution, and Dagum) and determined the most appropriate distribution for each category through goodness-of-fit tests, such as the Kolmogorov-Smirnov (K-S) and Cramér von-Mises (CvM) tests.

The second method was more data-driven, using advanced ML models to directly predict the TDR value. This method relied on a complex array of 12 features for TDR prediction, including diverse parameters like track subsurface condition, ditch properties, and historical geometry data. Researchers employed four advanced ML models in this analysis: Random Forest, Support Vector Regressor (SVR), XGBoost, and Artificial Neural Network (ANN).

This dual-methodology approach highlighted the varied ways TDR can be assessed and predicted, and underscored the importance of selecting the right tools and analyses based on specific track conditions and available data. By offering these two distinct yet complementary approaches, the project makes significant contributions to the field of railway track maintenance, paving the way for more informed, data-driven decision-making processes that can significantly reduce maintenance costs and improve overall track performance and safety.

1.1 Background

Railroads have been incorporating a variety of condition inspection technologies for decades. Transforming this wealth of data into actionable insights is important for ensuring the safest and most efficient operation of railroads. Effective predictive models, which consider the interplay between condition data and performance metrics, can significantly boost confidence in decision-making.

Ensuring the safety of train operations is of utmost importance, and accurately forecasting potential issues in specific track sections is crucial for enhancing this safety. The mechanical behavior of railway tracks is influenced by numerous factors, including the BFI, million gross tons (MGT), drainage conditions, and historical track data. An optimal predictive model should incorporate these and other relevant factors to improve the precision of future track condition forecasts. These predictive models allow for more focused maintenance efforts, targeting only the track sections that require attention.

Drainage is a critical element in railway infrastructure design, significantly influencing track conditions and associated maintenance costs [3]. It includes two main aspects: (i) internal drainage, designed to efficiently remove water from the track structure, and (ii) external drainage, which manages surrounding water sources to prevent track inundation. Research by Kashani et al. [4] using triaxial tests revealed a notable reduction in ballast strength with increased water content, particularly in scenarios with a high BFI. This correlation between water content and diminished strength was similarly observed in studies on abraded ballast [5]. Further, Mamou et al. [6] investigated the effect of drainage on track foundations through laboratory-based hollow cylinder tests, finding that blocking drainage significantly lowers the threshold of cyclic shear stress required to cause failure.

Although drainage directly influences track degradation, its effects have not been extensively incorporated into data-driven, ML models. This may be due to the challenges in quantitatively evaluating drainage quality across different railway track sections. This research bridges the gap by assessing the drainage conditions of railway sections using LiDAR data, considering the specific topography of each section.

Maintaining ballasted track systems requires maintenance, repair, and replacement to ensure safe operation and allow for optimal traffic. The need to conduct track work is typically identified on an ad hoc basis, either by geometry car measurements or by visual observations. Chrismer and Selig [7] proposed a method to estimate the need for ballast-related maintenance based upon track settlement and roughness. This report presents a three-step strategy to develop a probabilistic approach to assess the state of good repair of railroad ballast with respect to fouling. The proposed probabilistic approach is based upon the vast quantity of data collected by railroads on track condition.

1.2 Objectives

The primary objective of this research was to develop advanced ML models that can effectively determine the need for maintenance operations on railway tracks. The team used extensive datasets (e.g., track geometry, GPR, and LiDAR data collected from railway systems) to train these models to forecast potential, near-future defects in specific track mileages. Furthermore, the project team designed the models to provide detailed insights into how various factors, such as properties of track ditches, impact TDR. Such insights are important for making informed

decisions, allowing for more strategic and efficient maintenance planning, ultimately leading to safer and more reliable railway operations.

1.3 Overall Approach

The overall approach of this research involved the development of ML models to accurately predict the condition of railway tracks. To achieve this, the study began with a thorough investigation and preprocessing of a substantial dataset, encompassing 35,000 track blocks. The research team then divided these blocks based on their subsurface conditions and external drainage characteristics. For each category, the team estimated the probability of TDR using an optimal probability distribution function. In a parallel approach, researchers trained four distinct ML models (i.e., Random Forest, SVR, XGBoost, and ANN) to predict the TDR values. The final phase of the research involved applying these proposed models for in-depth analysis and interpretation. This stage was important for understanding how various factors influence TDR, providing valuable insights that can inform future track maintenance strategies and decisions.

1.4 Scope

This research is centered around the development and application of ML models to precisely forecast the condition of railway tracks, with a specific focus on passenger rail systems. The team examined the influence of various factors, particularly the properties of track ditches, on TDR. The project team used four ML models to identify potential locations with geometric defects. These models integrate and analyze diverse parameters, including drainage quality and subsurface conditions, to offer comprehensive predictions. The team worked to enhance the predictive accuracy of track conditions and examine how different elements affect the integrity and maintenance needs of passenger rail tracks.

1.5 Organization of the Report

This report is organized into four sections. Section 2 outlines the technical aspect of fitting probability distributions for the estimation of TDR, and includes an analysis of the impact of high BFI on TDR. Section 3 discusses the extraction of 12 key features and the training of ML models for predicting TDR, and explores the influence of ditch properties on TDR. Finally, Section 4 concludes the report, summarizing the major findings of the research and its implications.

2. Stochastic Approach for Estimating Probability of Track Degradation Rate

Parameters that influence TDRs must be identified to accurately estimate their probability. While external drainage condition is a known factor affecting TDR, previous research has not used LiDAR to evaluate these drainage conditions and incorporate them into TDR estimation. To bridge this gap, this section proposes a novel framework that automatically assesses the lateral cross-section and drainage condition of tracks using LiDAR data. Subsurface condition of the track, derived from GPR data, is considered as another influential factor. The expected TDR is then estimated based on a combination of both subsurface and external drainage conditions.

2.1 TDR

2.1.1 Geometry Data

In the railway industry, track geometry data are regularly collected to detect problematic sections of the track, evaluate the rate of deterioration, and schedule maintenance intervention. The safety of track is ensured by keeping geometry defects under certain thresholds specified in compliance manuals [8]. The large dataset, accumulated over years, provides an opportunity to extract valuable knowledge regarding the mechanical behavior of railway track.

The research team analyzed 146 geometry inspections that were conducted between 2011 and 2021. The track was divided into blocks, each with a length of 80 ft (24.38 m). As geometry inspections provide data for every foot of track, there were 80 datapoints in each block. In this study, the team used the 62-ft profile mid-chord offset channel, which is considered the best indicator of subsurface problems [9].

2.1.2 Data Preprocessing

Data preprocessing is essential in analyses using large datasets due to noise and variations in the data from uncalibrated sensors, various weather conditions, and unaligned inspections. Geometry cars use GPS for the geographical location of data, but datasets from different inspections can be unaligned by a few feet or more. The project team used a cross-correlation algorithm to determine the amount of positional errors between the inspections, shifting the measurements from different dates based on the positional errors for data alignment [9].

The dynamic loads of trains in assets such as switches, under-grade bridges, and during interlocking is higher than on open tracks [10]. Therefore, the settlement and overall behavior of track at these asset locations are different from open tracks. To achieve accurate estimations of TDR for open tracks, researchers removed blocks with assets from the dataset.

2.1.3 Calculating TDR

TDR refers to the changes in geometry indices over time or traffic. The railway industry employs geometry indices such as standard deviation (SD),[1] track roughness, or track quality index to represent the condition of tracks [9]. In this study, the team used SD because it has the best

[1] SD is the amount of variation of a set of geometry defects from the mean value of a block.

correlation with track settlement [9]. Figure 1 (a) shows the SD of profile vs. time for an example block (solid and dotted lines represent short and long period degradation, respectively). As track degradation rate can be more accurately estimated based on subsurface conditions, it should be calculated around the dates of collecting GPR data. The GPR for this block was collected in November 2012, October 2014, and September 2019.

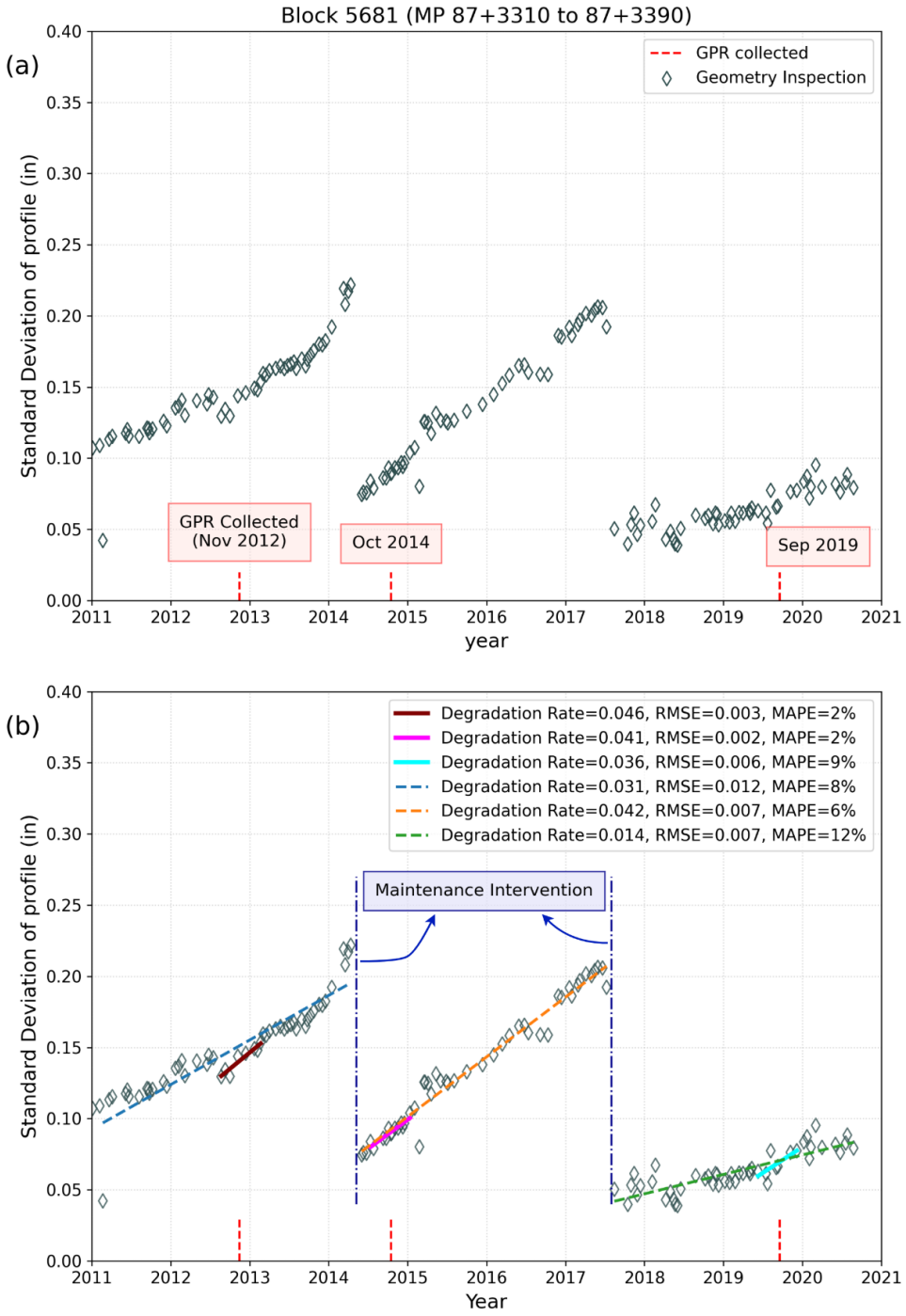


Figure 1. An example of studied profile geometry data of a block: (a) SD of profile for different geometry inspections vs. time and dates of collecting GPR, (b) fitted linear regression models to calculate track degradation rates.

The research team employed two different methods for calculating TDR: short period degradation and long period degradation. For short period degradation, the team selected periodic track geometry car inspections that had occurred three months before and after the date of collecting GPR (e.g., geometry inspections from August 2012 to February 2013 were selected for GPR data in November 2012). A linear regression model was fitted to the selected inspections. The slope of the fitted line represents short period degradation at the date of collecting GPR, shown by solid lines in Figure 1 (b). The long period degradation method requires detection of previous maintenance interventions to filter out maintenance effects on geometry behavior. Correctly detecting previous maintenance interventions is a complex task due to high variations and noise in geometry data. To cope with these variations, a combination of four conditions was considered, which distinguishes the difference between the noise and reduction of SD due to maintenance interventions.[2] After detecting previous maintenance, linear regression models were fitted to the inspections between two maintenance interventions, as shown by dotted lines in Figure 1 (b). The slope of the fitted line that covers the collection date of GPR represents long period degradation for that specific date. Section 2.2.1 presents a comparison between the effectiveness of short and long period degradation and their correlation with BFI.

To ensure the reliability of fitted lines and, in turn, calculated track degradation rates, the team calculated performance metrics such as root mean of squared error (RMSE), mean of absolute percentage error (MAPE), and mean of absolute deviation (MAD). If the MAPE of a fitted line was higher than 15 percent, the calculated degradation rate was considered unreliable, and the observation was removed from the dataset. Only 12 percent of datapoints were recognized as unreliable observations.

$$RMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(A_i - P_i)^2} \tag{1}$$

$$MAPE = \frac{\sum_{i=1}^{N}|A_i - P_i|}{\sum_{i=1}^{N} A_i} \times 100 \tag{2}$$

$$MAD = \frac{\sum_{i=1}^{N}|A_i - P_i|}{N} \tag{3}$$

A_i and P_i are the actual and predicted values of SD for ith observation, respectively. N is the number of inspections that are selected for fitting a linear regression model.

2.2 Categorizing the Dataset

In stochastic analysis, dividing the entire dataset into subgroups based on their properties results in more accurate estimations [11]. This is attributed to the fact that smaller groups with similar properties are more likely to have similar behavior compared to the entire dataset. The types of properties chosen for categorizing the dataset directly affects the accuracy of estimated output. Categorization should be typical, so that it can be used on other datasets, and it should be

[2] These conditions and details are provided in a previous paper by the authors of this study [1].

effective so it increases the accuracy of estimations. In this study, the team chose subsurface and drainage conditions of tracks for categorizing data as they have a significant impact on track degradation rate. These categories are typical and can be used for other passenger tracks with similar conditions. The project team also considered using class of track as another category. However, as revealed in the previous study [1], TDR is relatively similar for Classes 6, 7, and 8. Thus, class of track was not an effective category and consequently was not considered in this section.

2.2.1 Categorizing Based on Subsurface Condition

Subsurface conditions of tracks can be determined from GPR inspections. Important properties such as BFI and ballast thickness can be extracted from interpolated GPR data [12], [13]. In this study, the research team chose BFI to represent subsurface condition as it is the most important feature that can be extracted from GPR data. Because three antennas were used to collect GPR, BFI data are provided for three channels (i.e., center of track and left and right shoulders). The interval of BFI data along the track is 16 ft (4.88 m); therefore, there are 15 BFI values in each 80-ft (24.38-m) block, as shown in Figure 2. These values should be summarized into a single number (using averaging or maximizing) so that a representative number can be used for categorizing the dataset.

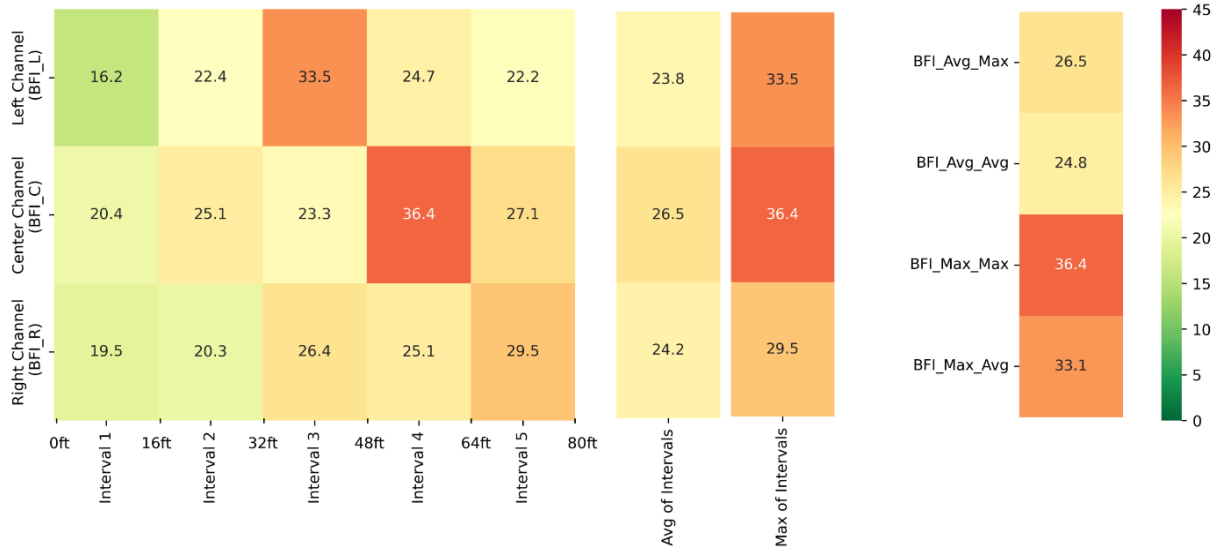


Figure 2. Example BFI data block. The final value of BFI in a block strongly depends on the summarizing approach. Blue: BFI=0, red BFI=45.

The correlation between TDR and different approaches for defining the representative BFI values in a block should be investigated to find the best representative for the block. These correlations provide valuable information for creating the right solution for cleaning the track. For example, if TDR has a high correlation with the maximum of BFI values, the fouled ballast should be replaced with clean ballast even if BFI is high only in one interval (i.e., 16 ft). On the other hand, if TDR has a high correlation with the average of BFI values, the cleaning should be scheduled only for blocks that have high BFI in all intervals. Therefore, the summarizing approach to calculate the appropriate BFI representative has a significant impact on maintenance

planning. The following approaches are employed for calculating the BFI representative value in a block (Figure 2):

- BFI_Avg_Max: Calculates the average for each channel then considers the maximum value of three calculated averages.
- BFI_Avg_Avg: Calculates the average for each channel then considers the average value of three calculated averages (i.e., overall average).
- BFI_Max_Max: Calculates the maximum value for each channel then considers the maximum value of three calculated maximums (i.e., overall maximum).
- BFI_Max_Avg: Calculates the maximum value for each channel then considers the average value of three calculated maximums.

Figure 3 shows the Pearson correlation between different approaches for calculating BFI representative value and TDR.

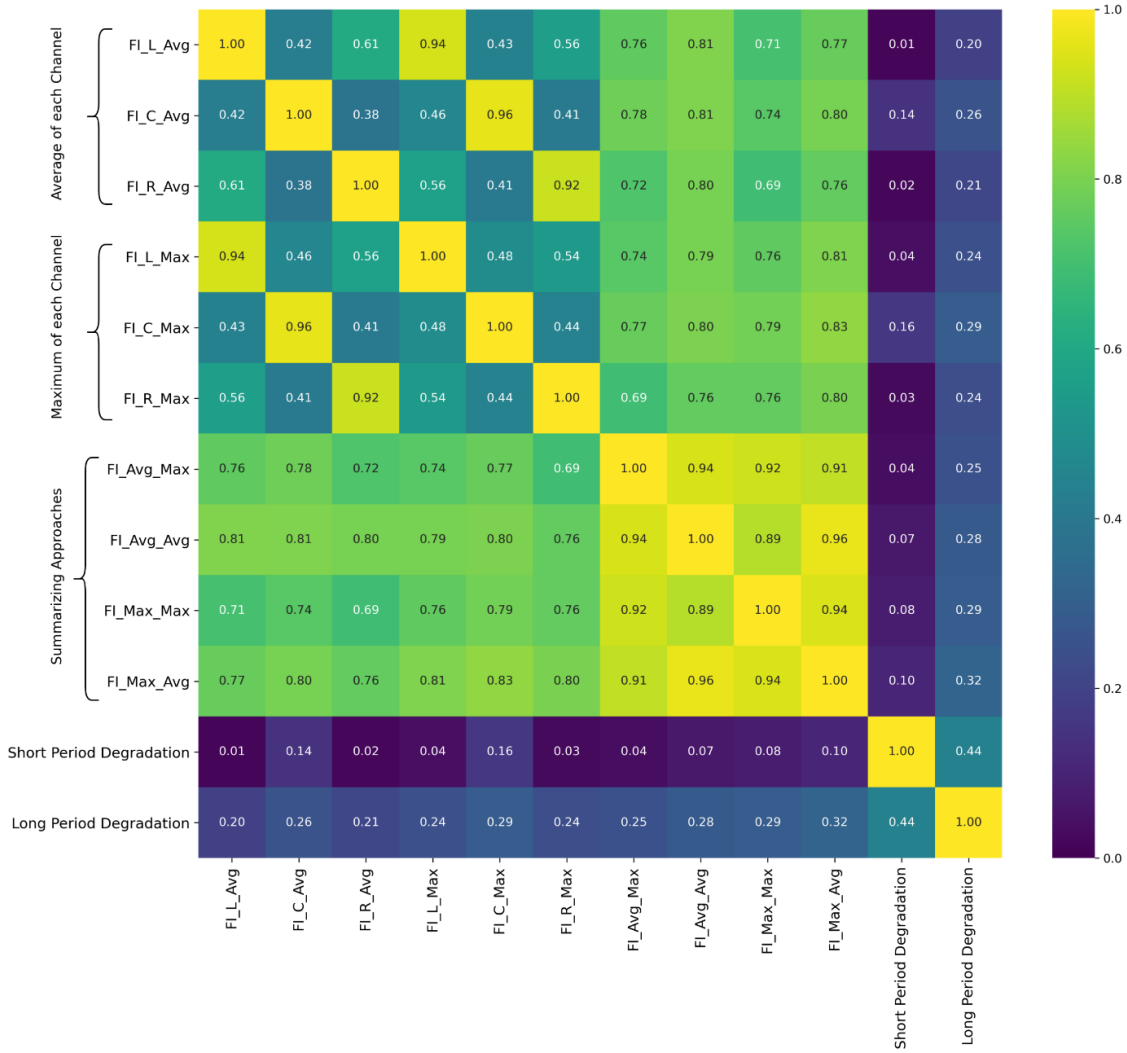


Figure 3. The correlation plot using Pearson equation; violet: Pearson Correlation=0, yellow: Pearson correlation=45.

BFI has a higher correlation with long period degradation than with short period degradation regardless of BFI channel or summarizing approach. Thus, long period degradation can better represent track behavior. Among the summarizing approaches, BFI_Max_Avg has the best correlation with TDR. Thus, this approach was used to calculate the final representative value of BFI for each block and categorize the dataset based on subsurface conditions. The entire dataset is divided into four categories from clean ballast to highly fouled ballast, as presented in Table 1.

Table 1. Categorizing dataset based on subsurface condition of track

Name of Category	BFI Limits	Description
G1	5£ BFI <15	Clean Ballast
G2	15£ BFI <25	Moderately Fouled Ballast
G3	25£ BFI <35	Fouled Ballast
G4	35£ BFI <45	Highly Fouled Ballast

Figure 4 shows the box plot of TDR with respect to subsurface categories.

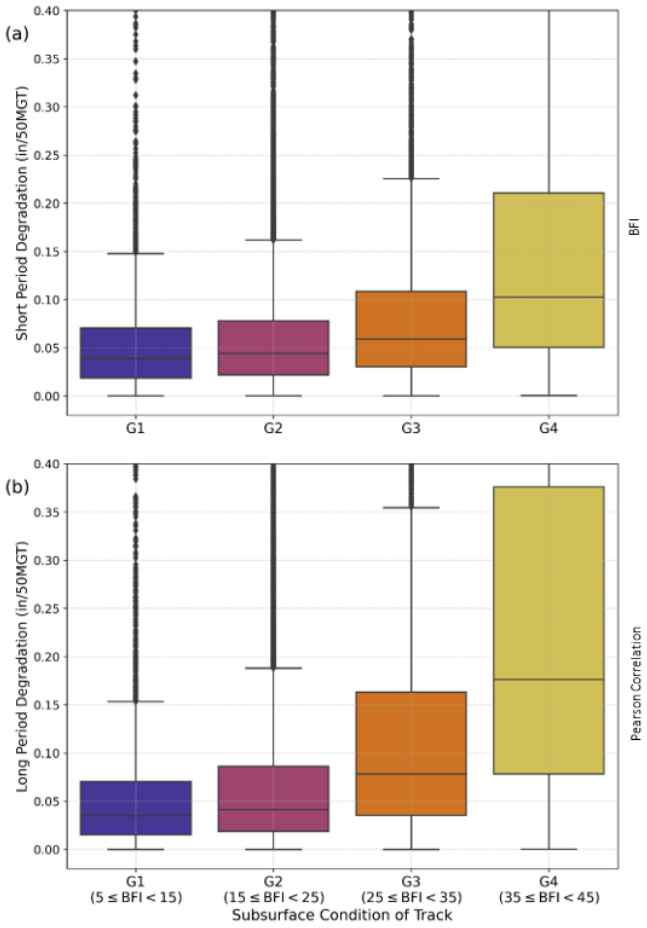


Figure 4. Box plot of track degradation rate with respect to subsurface conditions; (a) short period degradation; (b) long period degradation.

TDR increases with increasing fouling in ballast both for short period and long period methods of calculating degradation rate. However, the difference between categories of long period

degradation is much higher than short period degradation, which indicates the superiority of this method to represent track behavior. Therefore, only long period degradation is employed for the rest of the analysis in this report to avoid presenting redundant results. The direct correlation between BFI and track settlement is observed in many laboratory studies [4], [5], [14]. There are many outliers indicated on the box plots, but those outliers represent a very small percentage of the samples. The outliers are more dispersed on the higher end of the distribution than the lower end because the values represented are all positive (>0).

2.2.2 Categorizing Based on Cross Section and Drainage Conditions

In addition to BFI, track drainage conditions affect TDR. If the external drainage of track is poor, the water from precipitation will remain in the track and result in high degradation rates. The negative effects of additional water in track are observed by box tests [5], [14]. To investigate the drainage condition of track, the combination of ditch geometry and track cross-section derived from LiDAR data were analyzed. For example, if track is on fill, precipitation will be drained from track, and a ditch is not required. Conversely, ditch geometry plays a significant role in the drainage condition of cut or at-grade cross sections. To analyze large datasets, the research team proposed an algorithm to automatically evaluate drainage quality. This algorithm addresses the following challenges:

- The distance between adjacent tracks and the distance between ditch and center of track varies along the track. The algorithm should be able to find the exact location of a ditch if any ditch exists.
- The algorithm should be able to discriminate ditch from fill cross sections when the relative elevation of adjacent terrain decreases.
- If there are multiple tracks, the ditch and terrain further from a track should not be considered in drainage quality if the distances are too high.

The proposed algorithm in this study solves these challenges. The flowchart for this algorithm is shown in Figure 5. The algorithm divides blocks into three categories: cut without good ditch (herein “cut”), at-grade or others (AGO), and fill cross sections. The cut category represents a poor drainage condition as precipitations will accumulate close to the track. Fill cross section represents a good drainage condition as it allows for proper drainage of precipitation away from the track, while AGO represents a moderate drainage condition. To determine the drainage condition of a cross section of track, researchers find the starting locations of the adjacent terrain, and compare the average elevation of terrain with the track elevation. In cut cross sections, if the depth of the adjacent ditch is deeper than 2.5 ft (0.76 m) from top of tie (TOT), the cross-section is changed to AGO, which represents moderate drainage condition. In other words, the existence of a good ditch in cut cross sections changes drainage conditions from poor to moderate as it keeps the surface water away from the track. The 2.5 ft threshold was chosen based on the recommended standard depth of ditch [3].

Figure 6 shows some applications of the proposed algorithm for evaluating the drainage conditions and cross sections of tracks. The relative elevation has been zeroed on the TOT of T2. The black “x” marks show the starting locations of ditches and adjacent terrains, which vary along the track. To find this starting location, the relative elevations are checked from the center of T2 to 35 ft (10.67 m) lateral distance in both side directions. The first location in which the absolute elevation is more than 1 ft (shown by red dotted lines) from the TOT of T2 is marked as

the start of ditch and adjacent terrain. The depth of ditch is assessed from the starting point (x mark) up to 15 ft (4.57 m) on each side (double hatched in Figure 6), while adjacent terrain considers up to 50 ft (15.24 m) on each side (double hatched plus single hatched in Figure 6). The drainage conditions and cross sections of track are evaluated based on the depth of the ditch and average elevation of adjacent terrain with thresholds of 2.5 (0.76 m) ft for depth of ditch and 3 ft (0.91 m) for average elevation of adjacent terrain. More details are presented in the flowchart of the algorithm (Figure 5).

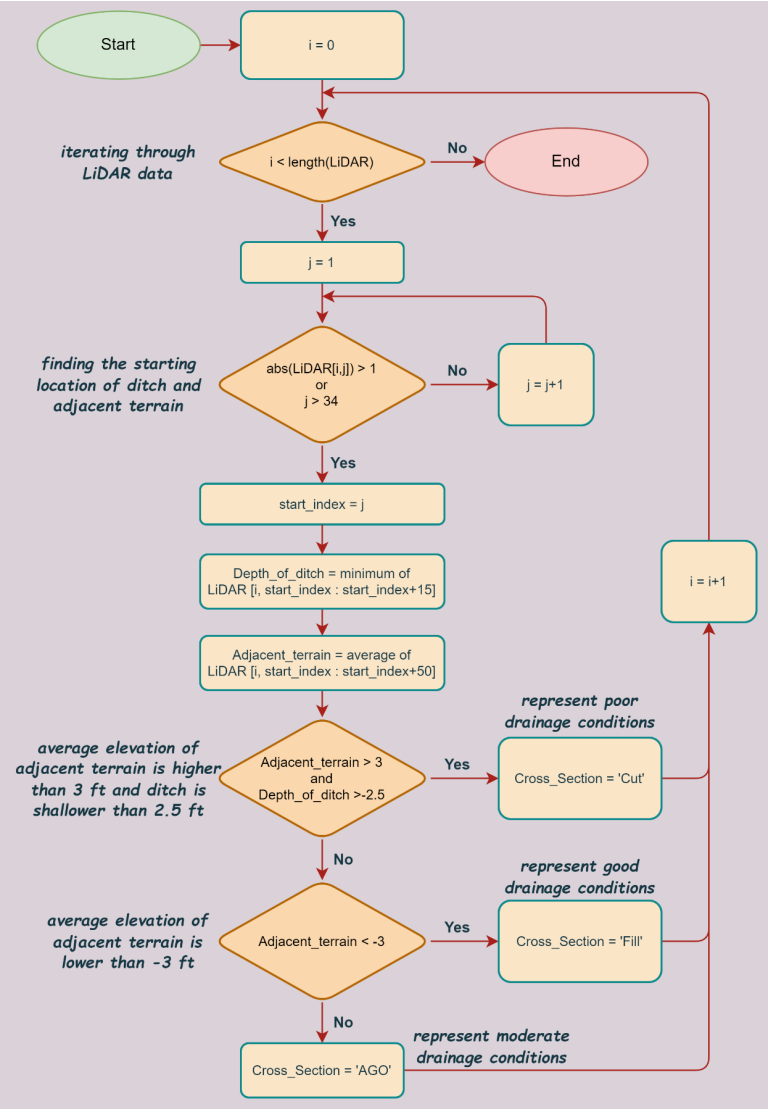


Figure 5. Flowchart for evaluating drainage and cross section condition from LiDAR data.

Figure 6 (a) shows a cross-section of track in a cut. Although there are ditches on both sides of the tracks, they are not deep enough to keep the water away from the track, so the drainage condition is poor. Figure 6 (b) shows a typical at-grade cross-section where the drainage condition is moderate since water that falls on adjacent terrain would not accumulate on track. Figure 6 (c) shows a typical fill cross-section, representing good drainage conditions. Figure 6 (d) shows AGO for the left cross-section and cut for the right cross-section. Although the

average elevation of the left terrain is 3.2 ft and is higher than the threshold, it was categorized as AGO because the drainage condition is moderate due to the existence of a good ditch. The depth of ditch is 3.5 ft, which is deep enough to keep water adequately away from the track.

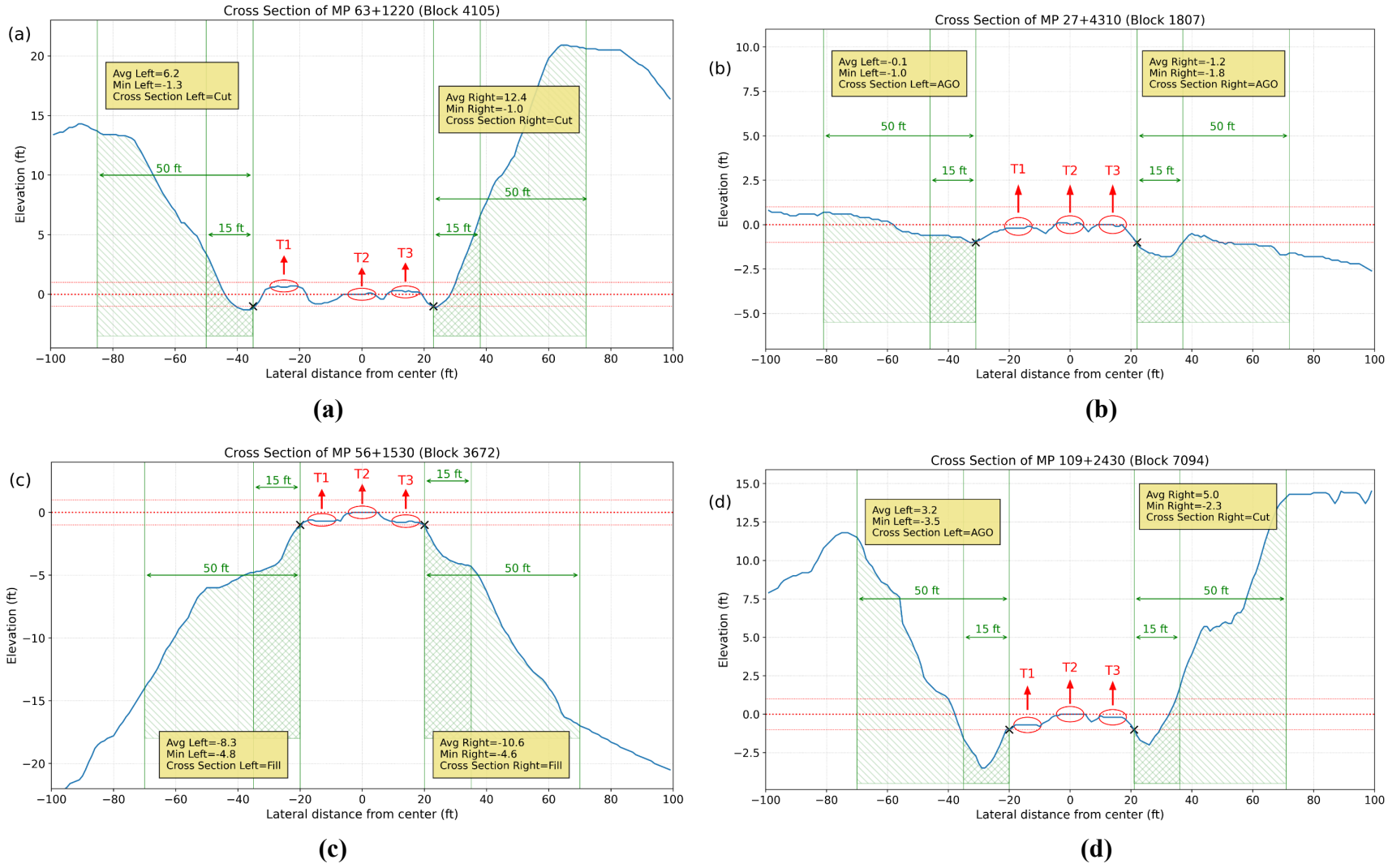


Figure 6. Applications of the proposed algorithm for determining cross-section and drainage condition of tracks; (a) typical cut cross-section (poor drainage condition) for all three tracks; (b) typical AGO cross section (moderate drainage condition); (c) typical fill cross section (good drainage condition); (d) cut cross-section for right terrain and AGO cross-section for left terrain due to existence of a good ditch.

Figure 8 illustrates TDR box plots with respect to the cross-section and drainage conditions of track. TDR is calculated using the long period method as it better represents track behavior. The median and 75th percentile of TDR in the cut category is much higher than the other two categories, which indicates the detrimental effects of poor drainage conditions on tracks. Although track degradation rate in the fill category is lower than AGO, the difference between these categories is not remarkable, showing that moderate drainage condition is enough for reducing the track degradation rate.

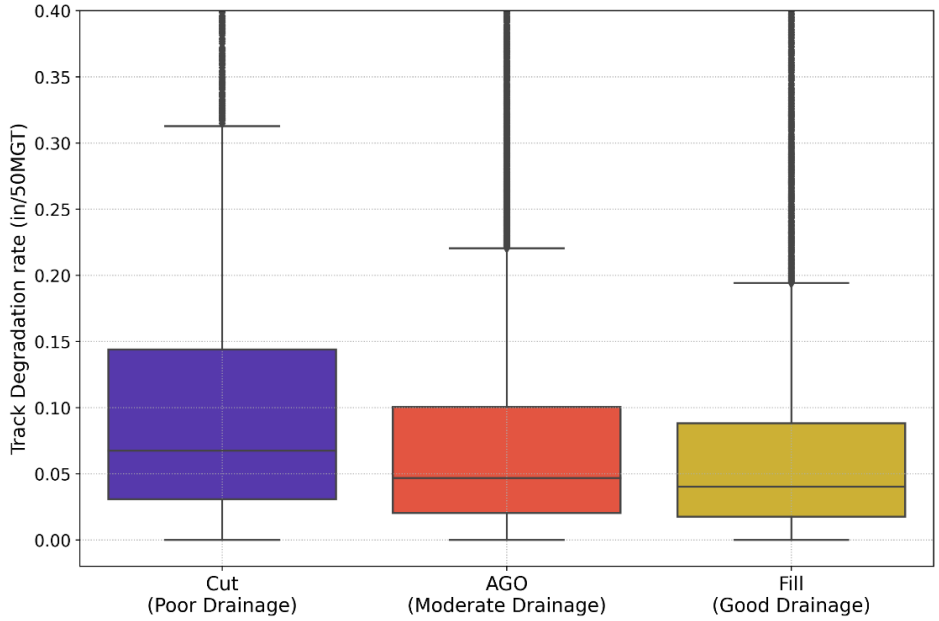


Figure 7. Box plot of track degradation rate with respect to drainage conditions.

2.3 Stochastic Models

TDR is a function of many variables with an uncertain nature, but measuring all variables that affect TDR is tedious. Even if all variables are measured, there are uncertainties incorporated with them that would be reflected in the track degradation rate. Thus, stochastic and probabilistic models are good approaches for estimating the expected track degradation rate [2]. To propose stochastic models, two variables that highly affect the track degradation rate (i.e., subsurface and drainage conditions) are considered. The entire dataset is divided into 12 categories based on these variables. Table 2 presents the statistical values of track degradation rate in each category. The skewness of all categories is positive, which indicates more concentration of datapoints on the left side of the histograms. The counts of two categories (i.e., “G1 and Cut” and “G4 and Fill”) are much lower than the other categories, which might be an indicator of the rarity of occurrence of these two categories in the field. This is presumably due to a better interaction between internal and external drainage in the fill cross sections than cut cross sections. Better interaction between internal and external drainage in fill cross sections will help the water to flow through the ballast matrix resulting in migration of fine particles (i.e., fouling) out of the ballast layer. However, this interaction in cut cross sections is poor and lead to accumulation of fouling particles in the track. Figure 8 shows a flowchart diagram that summarizes all steps for stochastic analysis of track degradation rate.

Table 2. Statistical values of track degradation rate in each category. The distribution is represented with the minimum value, median value, maximum value, standard deviation, mean value skewness and kurtosis.

Name of Category	Count	Min	Median	Max	Standard Deviation	Mean	Skewness	Kurtosis
G1 and Cut	180	0.0025	0.050	0.480	0.0890	0.080	2.39	6.4
G1 and AGO	1,586	0.0007	0.034	0.405	0.0510	0.050	1.92	4.3
G1 and Fill	1,503	0.0008	0.034	0.388	0.0630	0.054	2.52	7.5
G2 and Cut	2,635	0.0015	0.048	0.610	0.0860	0.079	2.22	5.8
G2 and AGO	15,285	0.0009	0.041	0.511	0.0770	0.066	2.52	7.7
G2 and Fill	7,534	0.0009	0.038	0.573	0.0800	0.065	2.61	8.2
G3 and Cut	1,643	0.0059	0.114	1.642	0.2490	0.195	3.11	11.4
G3 and AGO	4,022	0.0021	0.081	1.646	0.2170	0.154	3.43	14.7
G3 and Fill	734	0.0013	0.071	1.110	0.1460	0.120	2.71	9.8
G4 and Cut	530	0.0109	0.179	2.172	0.3790	0.312	2.57	7.0
G4 and AGO	1,067	0.0048	0.162	2.525	0.3940	0.304	2.69	8.5
G4 and Fill	163	0.0141	0.237	1.613	0.2960	0.299	2.16	5.5

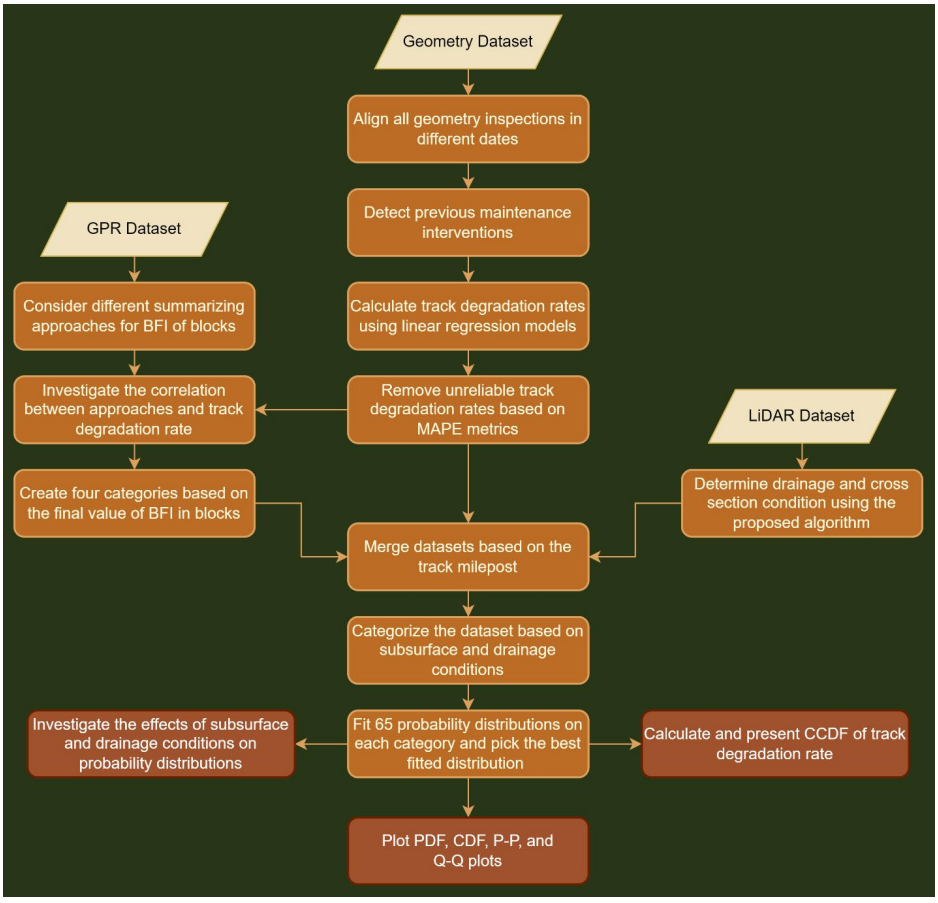


Figure 8. Flowchart diagram of approach for stochastic analysis of track degradation rate.

2.3.1 Fitting Probability Distributions

To propose stochastic models for TDR, 65 probability distributions were fitted to each category. Probability distributions that fit well with right-skewed datasets (e.g., Lognormal, F-distribution, Dagum, Power Lognormal, Beta prime, Burr, and Johnson SB) are ideal for TDR. Goodness-of-fit tests such as Kolmogorov-Smirnov (K-S), Cramér Von-Mises (CvM), and Chi-Square were conducted for each distribution. The best fitted distribution was chosen based on the Kolmogorov-Smirnov test as it is the most common test for continuous variables [16].

Table 3 presents the goodness-of-fit values and parameters of the best-fitted distribution for each category. The null hypothesis in goodness-of-fit is that the dataset follows a specific distribution. Thus, higher p-values in tests represent better fits. The p-value depends on test statistics and critical value, which is a function of the number of datapoints and significance level. For a constant critical value, a lower test statistic results in a higher p-value. The CvM test is a refined version of the K-S test that considers the area between fitted CDF and actual CDF rather than the maximum distance between the curves. Although Chi-Square is usually employed for categorical variables, it can be used for continuous data by dividing the dataset into nonoverlapping bins [17].

Probability distributions are defined by location, scale, and shape parameters. Making a change in location moves the distribution along the x axis, while changing the scale either stretches or squeezes the distribution. In addition to the location and scale, some distributions have shape parameters, which makes them more flexible. These parameters are presented in Table 3 for the best fitted distribution of each category.

Table 3. Goodness-of-fit values and parameters of the best fitted distribution on each category.

Name of Category	Name of Best Fitted Distribution	K-S* Statistic	K-S p-value	CvM** Statistic	CvM p-value	c2 Statistic	Scale	Location	Shape Parameter1	Shape Parameter2
G1 and Cut	Inverse Weibull	0.0335	0.983	0.0309	0.973	3.74	0.061	-0.02400	1.813	
G1 and AGO	Beta prime	0.0131	0.942	0.0437	0.913	7.35	0.200	0.00013	1.146	5.227
G1 and Fill	F-distribution	0.0157	0.844	0.0674	0.768	4.74	0.041	-0.00001	2.547	6.550
G2 and Cut	Lognormal	0.0133	0.729	0.0895	0.639	16.82	0.052	-0.00319	1.017	
G2 and AGO	Dagum	0.0084	0.226	0.2649	0.170	18.74	0.066	0.00001	1.064	1.857
G2 and Fill	Lognormal	0.0089	0.577	0.1373	0.431	19.88	0.039	-0.00170	1.117	
G3 and Cut	Inverse Gaussian	0.0205	0.511	0.1125	0.527	11.89	0.213	-0.01934	0.829	
G3 and AGO	Power Lognormal	0.0171	0.197	0.2432	0.197	10.79	21.509	-0.00037	47.594	2.558
G3 and Fill	Beta prime	0.0211	0.895	0.0522	0.863	7.13	0.757	0.00002	0.978	7.481
G4 and Cut	Johnson SB	0.0353	0.571	0.1135	0.523	9.18	0.959	-0.00244	1.366	0.889
G4 and AGO	Johnson SB	0.0206	0.793	0.0824	0.678	5.63	0.877	-0.00118	1.094	0.679
G4 and Fill	Johnson SB	0.0416	0.945	0.0296	0.978	2.24	0.658	-0.00236	0.538	0.673

* Kolmogorov–Smirnov
** Cramér Von-Mises

The railway community has two options for estimating expected track degradation rates proposed in this report: (i) employing the presented parameters in Table 3 to recreate probability

distributions for a specific category and (ii) referring to Appendix A, which presents the CCDF (Complementary Cumulative Distribution Function, which is $1 - CDF$) for each category. The reason for using CCDF rather than CDF is that maintenance scheduling is based on the exceedance probability of a specific track degradation rate.

Figure 9 displays histograms and the best fitted distributions of categories from clean ballast (G1) to highly fouled ballast (G4) and AGO cross-section. The datapoints are more concentrated on the left side of the histograms (i.e., right skewed) regardless of category. Moving from clean ballast to highly fouled ballast stretches the distribution. The probability of having high track degradation rates in G4 is much higher than G1, which indicates the negative effects of fouling on degradation rates.

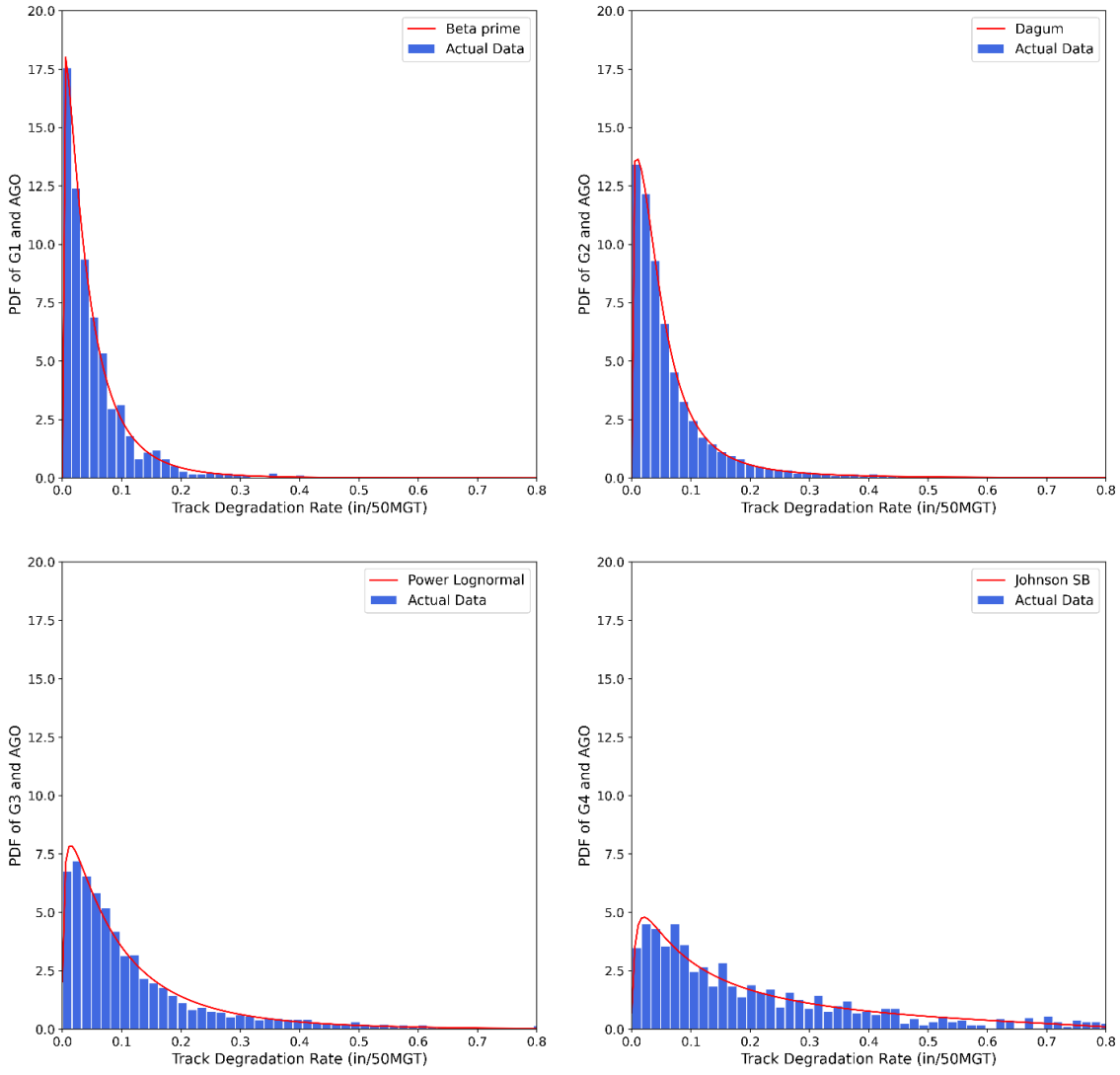


Figure 9. Histograms and the best fitted distributions of categories from clean ballast (G1) to highly fouled ballast (G4) and AGO cross-section.

Figure 10 shows the CDF of actual data and the best fitted distribution of categories from clean ballast (G1) to highly fouled ballast (G4) and AGO cross-section. Making a comparison between

actual data and fitted CDF indicates that probability distributions fitted very well with the data in all four categories. The K-S p-value of G2 and G3 are lower than the other two categories, which is attributed to the very low critical value in these groups. More clearly, the number of observations in G2 and G3 are extremely high, which results in very low critical values for hypothesis testing. Thus, even a small test statistic does not yield a high p-value.

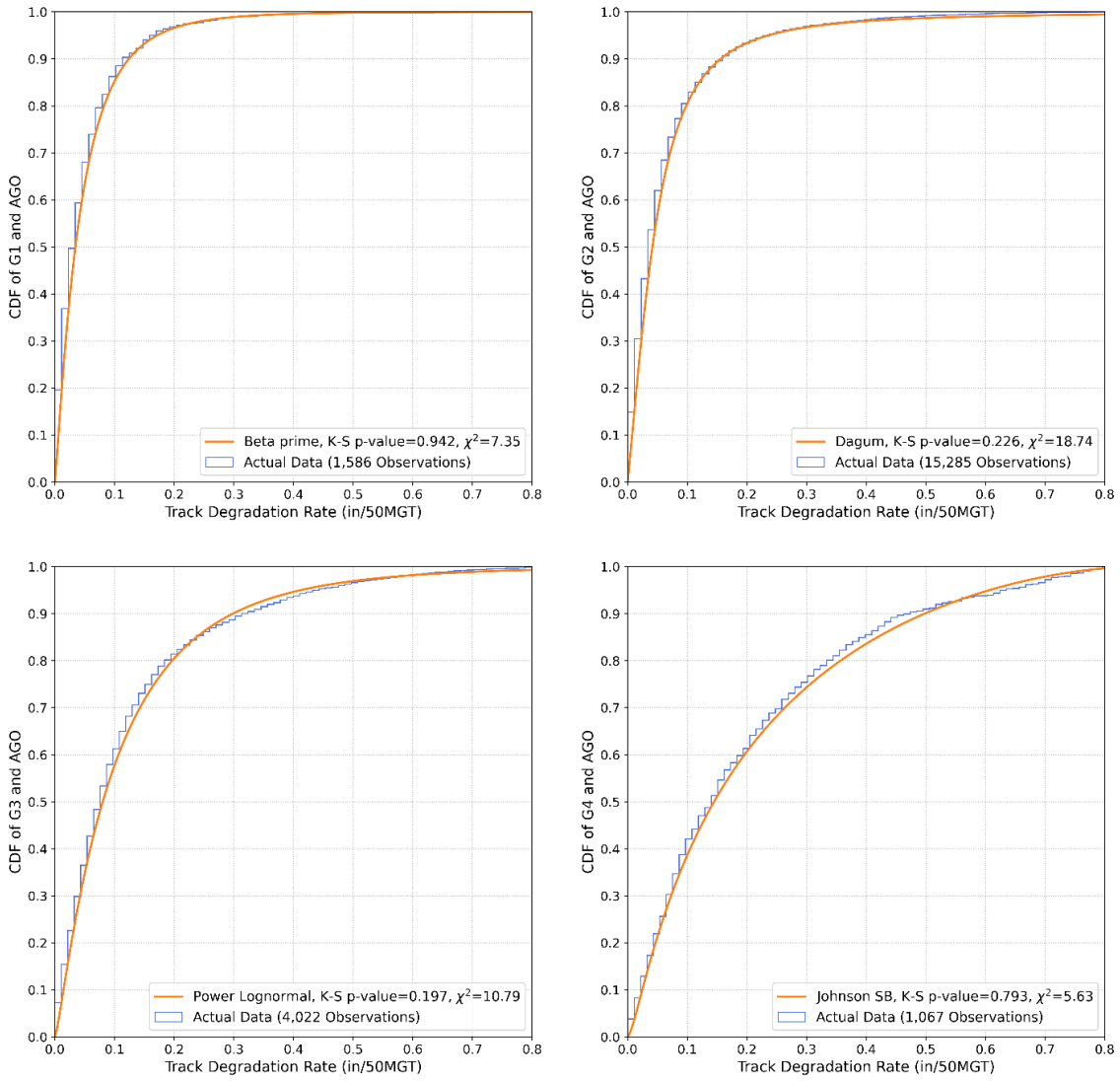


Figure 10. CDF of actual data and the best fitted distribution of categories from clean ballast (G1) to highly fouled ballast (G4) and AGO cross-section.

To further indicate that the distributions fit the data, Q-Q and P-P plots for G3 (the worst p-value among categories) are presented in Figure 11. Q-Q and P-P plots are used for visually assessing if the dataset comes from the fitted distribution. As shown in the Q-Q plot, there are only a few datapoints that have small deviations from the red dotted line; however, the majority of datapoints fall exactly on the diameter line, which indicates that the distribution fits the data. The deviation of datapoints in the P-P plot, which compares the empirical CDF with the fitted CDF,

is also small. All other categories have higher K-S p-values than G3 and AGO. Thus, their theoretical distributions fit even better with the corresponding datasets.

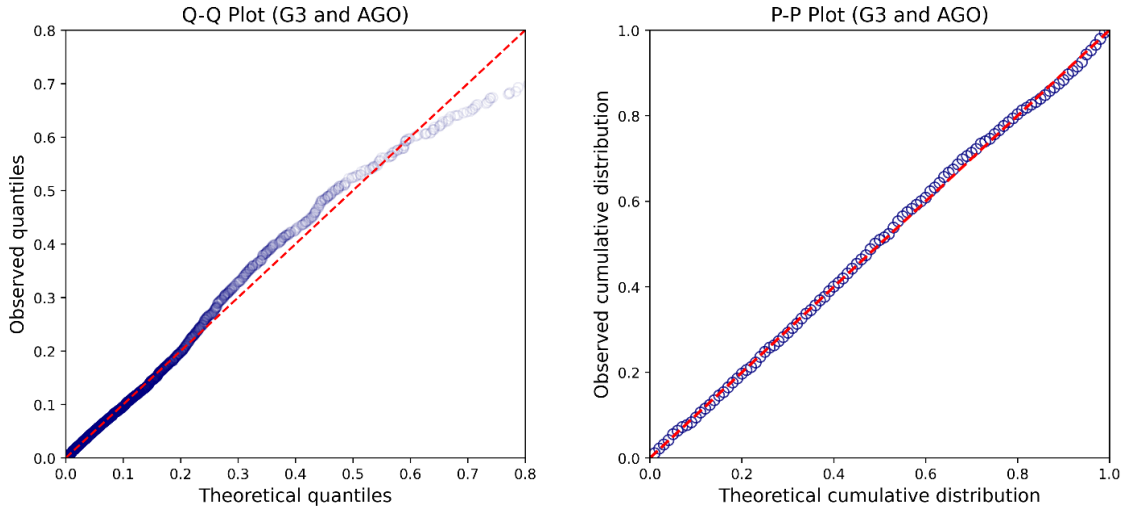


Figure 11. Q-Q and P-P plots of category "G3 and AGO," which has the lowest K-S p-value among categories.

2.3.2 Effect of Subsurface Conditions

The proposed stochastic models can be employed for estimating expected track degradation rates, which are required for scheduling geometry maintenance such as tamping. Investigating the effect of subsurface conditions on track degradation rate provides valuable information for making decisions about other types of maintenance, such as undercutting. Maintenance operations like undercutting that clean the ballast layer are costly, but they result in a reduction in the track degradation rate, which in turn decreases the need for tamping and increases track safety. By comparing the cost of undercutting and the amount saved from the reduction of track degradation rates, an informed maintenance decision can be made. Figure 12 shows the effect of subsurface conditions on track degradation rates using PDF and CCDF plots.

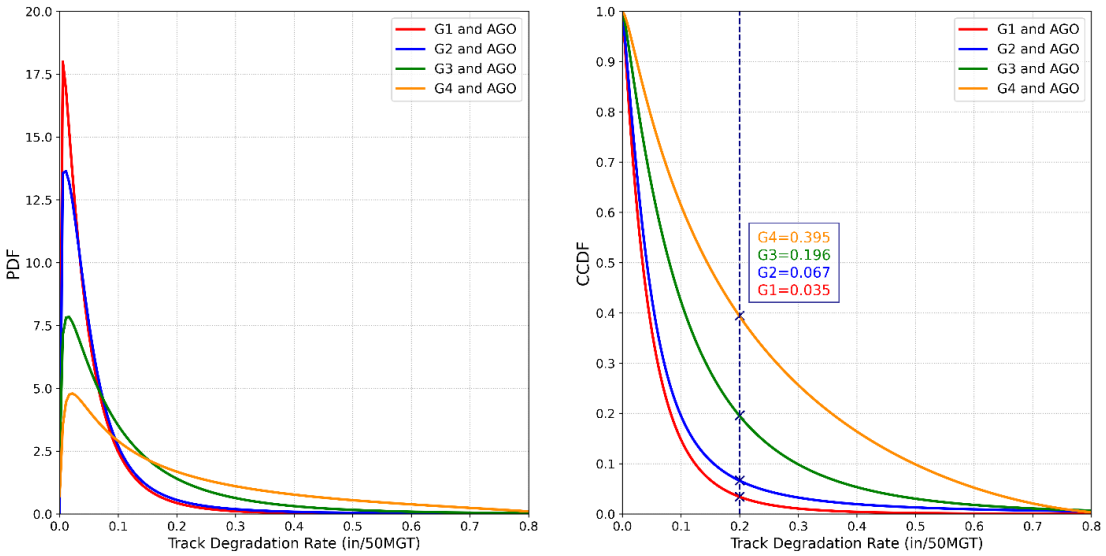


Figure 12. Effect of subsurface conditions on track degradation rates.

The probability of having high track degradation rates significantly increases by increasing ballast fouling (i.e., moving from G1 to G4). By considering 0.2 in/50MGT as a threshold, the probability of having high track degradation rates in G4 (i.e., highly fouled ballast) is 10 times than in G1 (i.e., clean ballast). The difference between G2 (BFI between 15 and 25) and G3 (BFI between 25 and 35) is significant, which indicates a BFI higher than 25 considerably increases track degradation rates. This finding is consistent with previous laboratory and data analysis studies [1], [5].

2.3.3 Effect of Drainage and Cross-section Conditions

Although negative effects of adding water to the track have been studied in laboratory testing [15] [6], the effects of drainage conditions on track degradation rates have not been investigated. Analyzing large datasets from the field provides valuable and reliable information for maintaining current or building new ditches to improve drainage conditions. Figure 13 shows the effect of cut, AGO, and fill cross sections, which represent poor, moderate, and good drainage conditions, respectively. By considering 0.2 in/50MGT as a threshold, the blocks with poor drainage conditions (i.e., cut) are 30 percent more likely to have high track degradation rates compared to blocks with moderate drainage conditions (i.e., AGO). As explained in Section 2.2.2, building a ditch with a depth of 2.5 ft (0.76 m) in cut cross sections changes drainage conditions from poor to moderate. Consequently, the probability of having high track degradation rates is reduced by building ditches. As illustrated in the figure, the difference between moderate and good drainage condition is not significant, which indicates moderate drainage condition is enough for keeping precipitation away from the track.

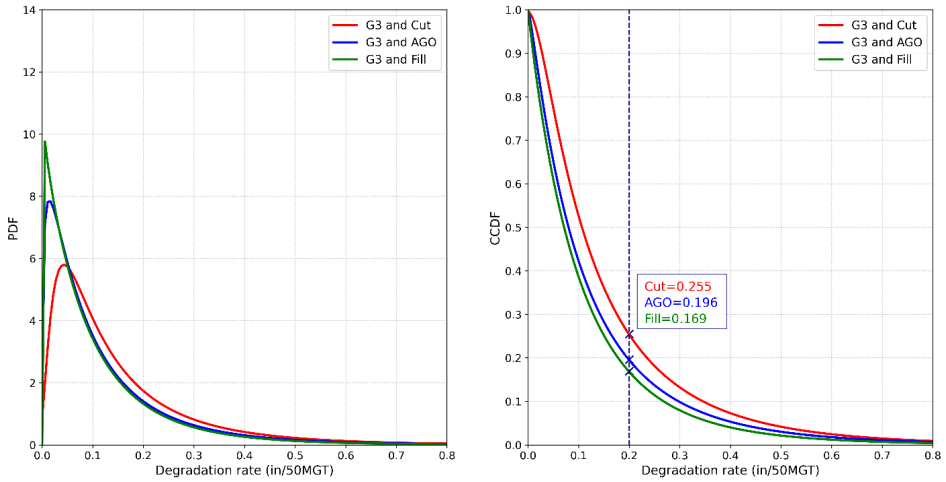


Figure 13. Effect of drainage and cross-section conditions on track degradation rates.

3. Predicting Value of TDR Using ML Models

Section 2 explored the estimation of TDR exceedance probability using stochastic methods. Building upon this foundation, this section advances to predicting the actual value of TDR, rather than estimating its likelihood of surpassing a certain threshold. This method requires a significantly larger amount of historical data and is applicable when such data is readily available. The first part of this section is devoted to data preprocessing and examining the impact of ditch properties on TDR. The second part introduces four ML models: Random Forest, SVR, XGBoost, and ANN, each designed for TDR prediction. This approach not only deepens the understanding of the factors influencing TDR but also uses advanced predictive models to provide more precise and useful insights for the maintenance and safety of railway systems.

3.1 Extracted Features in a Block

TDR is calculated by fitting linear regression models on inspections between two maintenance interventions. Figure 14 presents an example of maintenance cycles and TDR for one block. TDR is predicted at the time GPR data collection, which in this block is May 2018. Inspections prior to this date are considered historical, while those after are future inspections, represented by faded pentagons in the figure. Predicting TDR involves estimating the slope of the line fitted to the current maintenance cycle, shown as a dotted purple line.

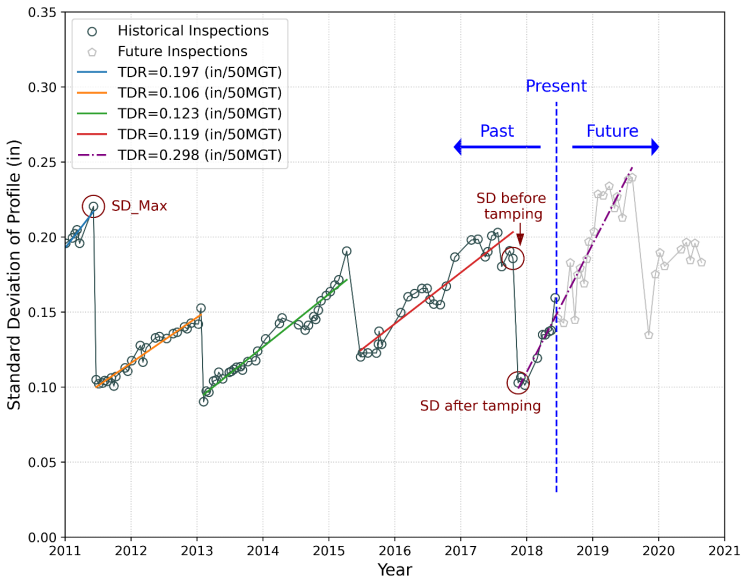


Figure 14. An example of SD vs time for a block. The goal of prediction models is to predict the TDR in current cycle, which is the slope of dotted purple line.

To accomplish this, 12 features are considered forming a broad spectrum of factors influencing TDR, including track subsurface conditions, drainage conditions, and track geometry history. The list of these features and their descriptions is presented in Table 4. Given that TDR is influenced by a multitude of factors, incorporating various features for predicting TDR is beneficial as it helps capture the complex interdependencies among these features, ultimately leading to more accurate and robust predictions.

Table 4. List of Extracted Features for Predicting TDR.

Number	Short name of feature	Description
1	Fouling Index	The value of Ballast Fouling Index (BFI) at the time of prediction
2	Depth of ditch	Depth of ditch from Top Of the Tie (TOT)
3	Distance of ditch	Distance of ditch from the center of track
4	Longitudinal cond of ditch	Ditch condition regarding water flow in the longitudinal direction
5	Current SD	SD of profile at the time of prediction
6	SD_Max	The maximum value of SD in the block since 2011
7	SD before prev tamping	The SD of profile in block before the last tamping
8	SD after prev tamping	The SD of profile in block after the last tamping
9	Number of tampings	The number of detected tamping since availability of geometry data (2011)
10	TDR in previous cycle	The value of TDR in the previous cycle
11	Kurtosis	The kurtosis of inspection at the time of prediction
12	Skewness	The Skewness of inspection at the time of prediction

3.2 Effect of Ditch Properties on TDR

To conduct a robust analysis on the impact of ditch properties on TDR, a large dataset consisting of 280 miles of passenger revenue track was examined. This extensive dataset mitigates the influence of localized or exceptional conditions, enabling the consideration of diverse track conditions and producing results that can be generalized to other tracks. For this large dataset analysis, ditch properties (e.g., depth and distance of ditch) are automatically assessed from LiDAR data. Figure 15 displays the cross section of a block as an example.

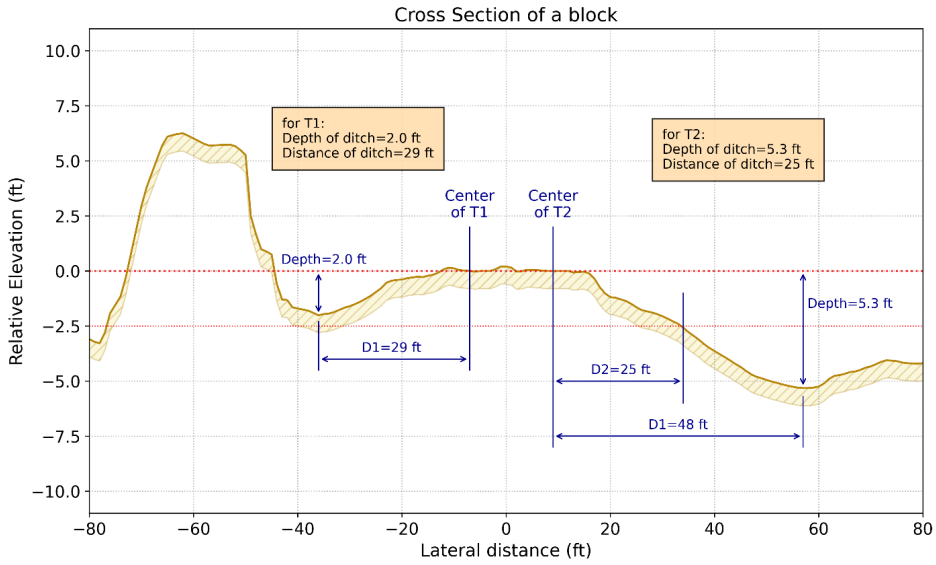


Figure 15. Cross section of a block as an example to show depth and distance of ditch; Distance of ditch is minimum value between D1 and D2.

Ditch depth refers to the elevation change from the TOT to the ditch bottom, while ditch distance is the minimum value between D1 and D2, as illustrated in the figure. D1 is the distance from the track center to the ditch bottom, and D2 is the distance from the track center to the point where the ditch depth exceeds 2.5 ft. When the ditch is shallower than 2.5 ft, D2 is null, and the ditch distance equals D1. Considering the minimum value of D1 and D2 ensures that the ditch distance is not incorrectly considered long when the ditch is deep and its bottom is far from the track (as shown in Figure 15 for Track 2.)

Ditches are not necessary for fill cross sections, as water from precipitation can move away from the track. Therefore, blocks with fill cross sections should be filtered out for an accurate investigation of the impact of ditch properties on TDR. To automatically identify fill cross sections from LiDAR data, the framework presented in [18] is used. This framework detects the starting location of adjacent terrain and compares the average elevation of the adjacent terrain to the track elevation to determine whether the track is in a fill, cut, or at-grade cross section. More details about this framework can be found in [18]. The influence of ditch properties on TDR is only examined for blocks that are either in cut or at-grade cross sections.

3.2.1 Effect of Depth of Ditch on TDR

Figure 16 presents the boxplot of TDR for blocks with ditches either shallower or deeper than 2.5 ft (76 cm). Two reasons for selecting 2.5 ft (76 cm) as the threshold for classifying the boxplot are: (i) it is a recommended standard for ditch depth [4], and (ii) it divides the dataset into two balanced groups with a relatively similar number of blocks. The TDR in blocks with ditches shallower than 2.5 ft (76 cm) is higher than in blocks with ditches deeper than 2.5 ft, highlighting the importance of ditch depth. The number of blocks in each group is displayed in maroon text within the figure.

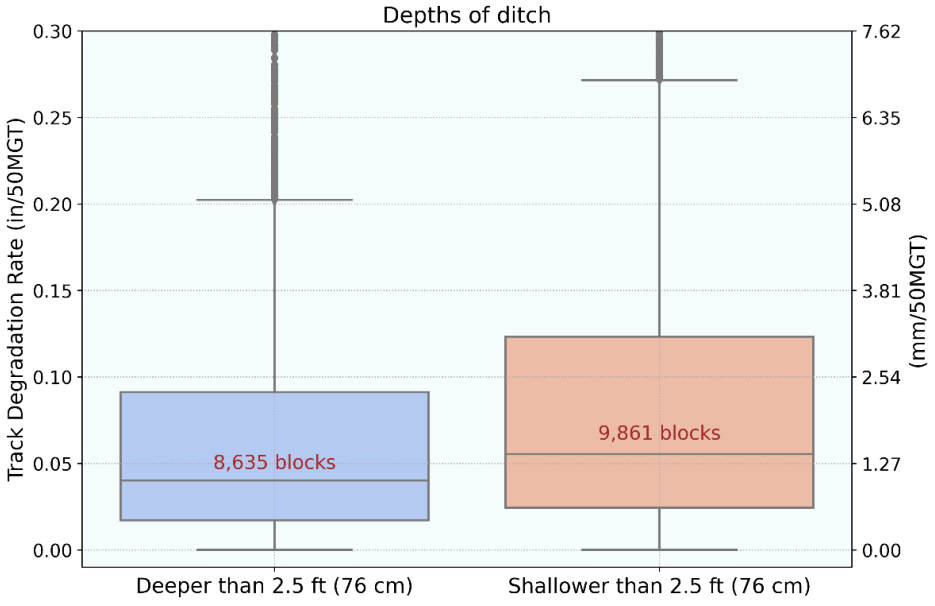


Figure 16. Effect of ditch depth on TDR.

3.2.2 Effect of Distance of Ditch on TDR

Another critical aspect of ditches is their distance from the track. Ditches positioned too far away from the track are less effective, as water from precipitation needs to travel a longer distance to reach them. Figure 17 displays the boxplot of TDR for numerous blocks with different distances from the track. Splitting the blocks into two groups at a 15 ft (4.57 m) distance results in balanced groups, which minimizes the risk of biased or misleading results, as each group has an equal representation and influence on the comparison, making the conclusions more reliable and robust. Blocks with ditches farther from the track exhibit higher TDR than those with closer ditches.

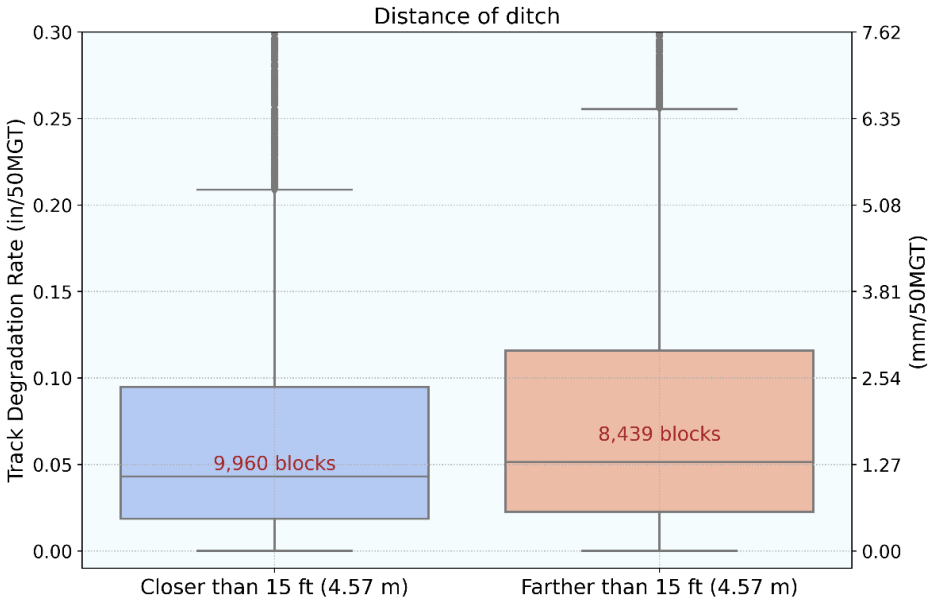


Figure 17. Effect of ditch distance from center of track on TDR.

3.2.3 Effect of Lateral Ditch Condition on TDR

Considering both lateral offset and depth of the ditch simultaneously allows for a more comprehensive investigation of the track's behavior and analysis of interactions between these properties. Investigating these interactions is crucial for understanding the combined effects of ditch depth and distance on TDR, which can help optimize track maintenance planning. Figure 18 displays the boxplot of TDR considering both depth and distance of blocks. As depicted in the figure, the impact of ditch distance on reducing TDR is more pronounced when the ditch is shallower than 2.5 ft. In other words, the results suggest that when a ditch is shallow, it needs to be close to the track; otherwise, TDR will increase substantially.

The considerable impact of external drainage quality on TDR can be assessed by differentiating between good and poor lateral drainage conditions. Good lateral drainage is defined for blocks with a ditch deeper than 2.5 ft and closer than 15 ft to the track's center, while poor lateral drainage refers to blocks with a ditch shallower than 2.5 ft and farther than 15 ft. Figure 19 compares the TDR for good and poor lateral drainage conditions across the 25th, 50th, and 75th percentiles. The TDR ratio for poor to good lateral drainage, represented by the maroon "x" mark, remains consistent at approximately 1.55 for different percentiles. This finding suggests

that TDR for poor lateral drainage quality is roughly 55 percent higher than for good lateral drainage quality, which is a significant difference. Given the negative effects of high TDR (e.g., the need for more frequent correction of track geometry through surfacing maintenance and reduced ride quality), maintaining high-quality ditches is important. Since this analysis was carried out on a large dataset of around 18,500 blocks, various track conditions were considered and examined. Consequently, the results are robust, reliable, and provide valuable insights for decision-making regarding ditch maintenance.

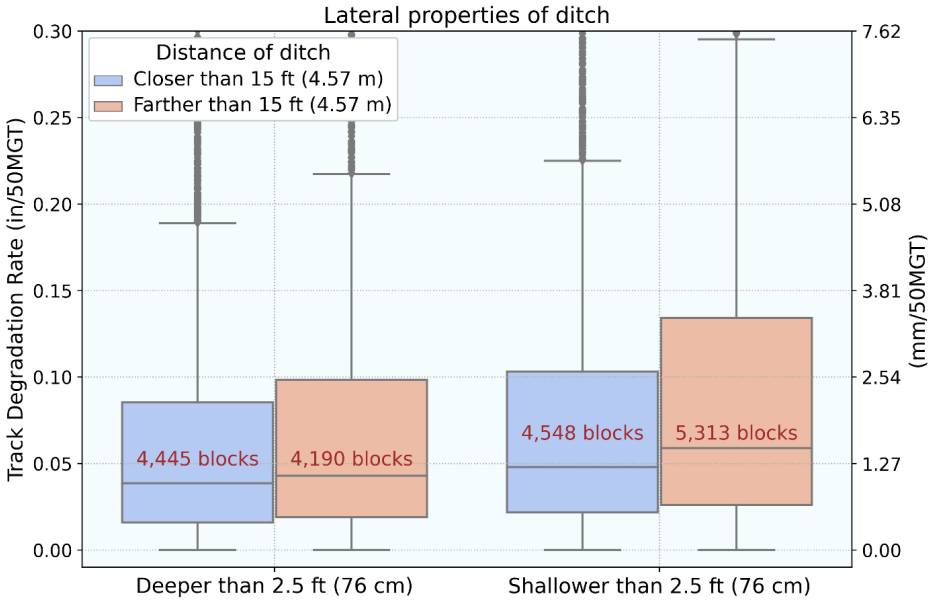


Figure 18. The boxplot of TDR considering both distance and depth of ditch; the effect of ditch distance is more significant when the ditch is shallow.

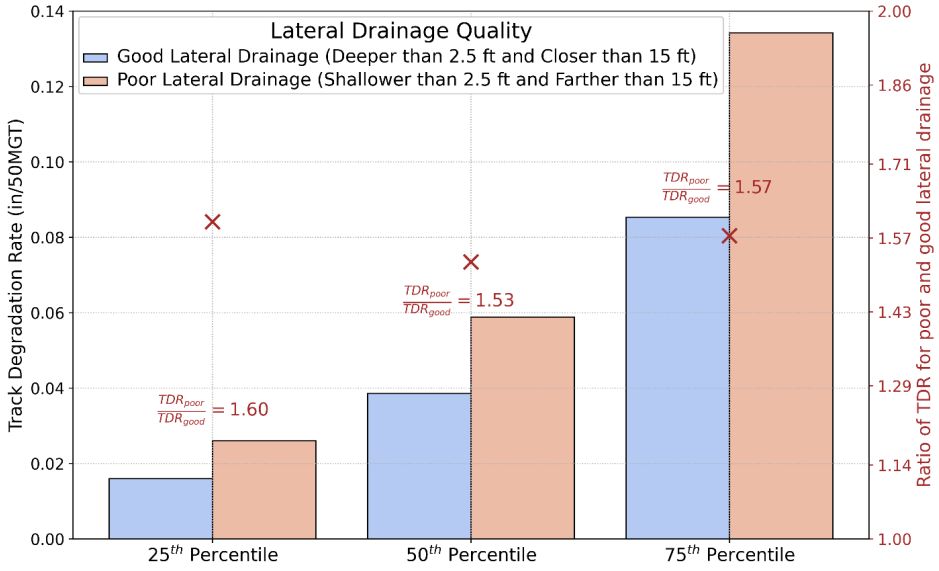


Figure 19. Comparing TDR for good and poor lateral drainage quality across different percentiles; the ratio of TDR for poor and good lateral drainage is shown by maroon "x" mark and is around 1.55 for different percentiles.

3.2.4 *Effect of Longitudinal Ditch Condition on TDR*

In addition to the lateral condition of a ditch, its longitudinal condition also impacts the overall drainage quality of the track. Lateral drainage is responsible for the flow of water from the track to the adjacent ditch, while longitudinal drainage ensures the water flows within the ditch, keeping it dry. A ditch with poor longitudinal condition accumulates water and hinders its flow, while a ditch with good longitudinal condition allows water to flow freely. Figure 20 illustrates schematic longitudinal conditions of ditches for both poor and good conditions.

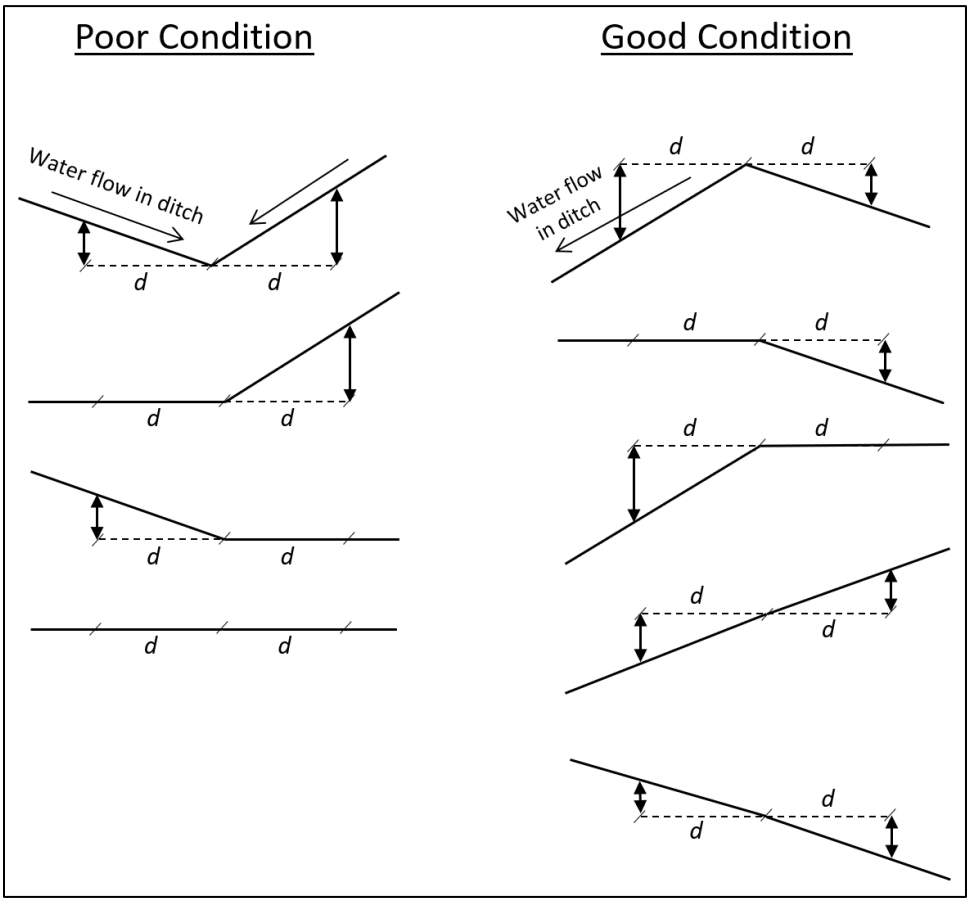


Figure 20. Schematic representation of longitudinal condition of ditch; in poor longitudinal drainage, water accumulates in the ditch, while in good longitudinal drainage, water flows freely.

To investigate the longitudinal condition of the ditch, topography contour lines are plotted using LiDAR data, as shown in Figure 21. The location of the ditch is automatically determined using LiDAR data and is represented by the green line in the figure. At each ditch location, the longitudinal slope is determined by measuring the elevation changes at the bottom of the ditch 50 ft in front of and behind that specific point. After calculating the longitudinal slope of the ditch, drainage conditions are evaluated based on the group they fall into as shown in Figure 20. This process is repeated for all blocks to evaluate their longitudinal drainage conditions. Figure 22 presents the boxplot of TDR for good and poor longitudinal conditions. Although TDR for poor longitudinal drainage is higher than for good longitudinal drainage, the difference is not as

substantial as it was for lateral drainage conditions. This indicates that lateral drainage conditions have a more significant impact on TDR than longitudinal drainage conditions.

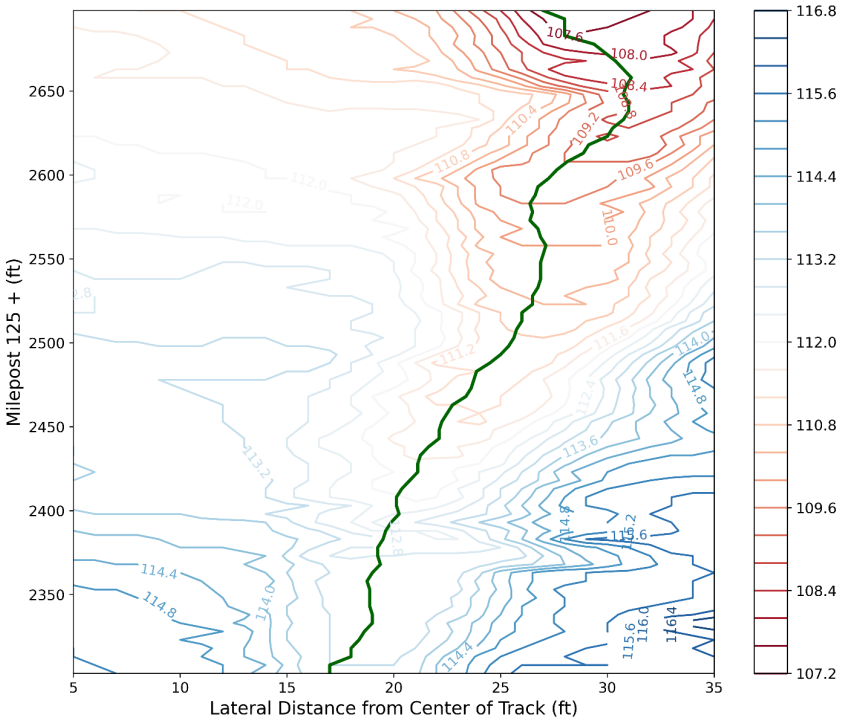


Figure 21. Topography contour lines at a milepost; the ditch location is automatically determined using LiDAR data and depicted by a green line. Note that this figure is not a geospatial map, but rather an indication of elevations at a distance from the centerline of track. It is also important that this figure is independent of the variations in alignment of the track.

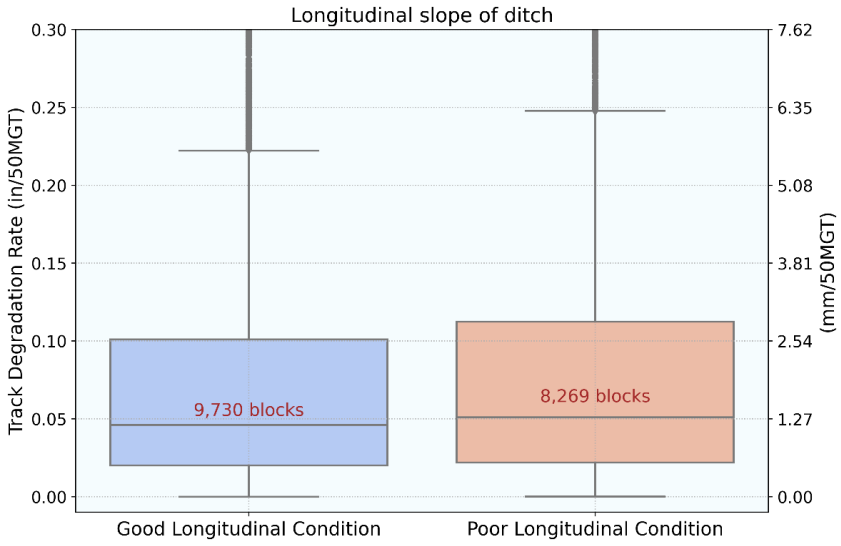


Figure 22. The effect of longitudinal drainage condition on TDR.

3.3 Machine Learning Models

TDR is predicted using multiple ML models with features such as track subsurface, drainage condition, and geometry history. Using multiple ML models offers several advantages. Each model has its own strengths and weaknesses, and by employing a diverse set of models, it becomes possible to capture a more comprehensive understanding of the underlying relationships in the data. This approach can lead to improved overall performance and robustness, as well as providing a more reliable interpretation of the results [19].

Additionally, there is a trade-off between model complexity and interpretability. Complex models, such as deep neural networks, might achieve high accuracy but can be challenging to interpret, while simpler models can be more interpretable but may not capture complex relationships in the data. By using multiple models, including Random Forest, SVR, XGBoost, and ANN, the analysis can strike a balance between complexity and interpretability. This enables the investigation of both simpler and more complex relationships in the data, while still providing insights that can be more easily understood and communicated. In the following subsections, a brief overview of each ML model's methodology is provided.

3.3.1 Random Forest

Random Forest is an ensemble learning method that constructs a multitude of decision trees during training and outputs the mean prediction of the individual trees for regression tasks [20]. It is built upon the concept of bagging, or bootstrap aggregating, which involves training several base models independently on random subsets of the training data and then combining their predictions through averaging for regression [21]. In Random Forest, the base models are decision trees. The combination of multiple weak learners (i.e., decision trees) produces a strong learner with improved performance and resistance to overfitting. The Random Forest algorithm can be represented as follows:

$$RF = \{T_1(x), T_2(x), \dots, T_k(x)\} \tag{4}$$

where RF is the Random Forest model, $T(x)$ is a decision tree, and k is the number of trees in the forest.

During the construction of each decision tree, a random subset of the training data is drawn with replacement, typically equal in size to the original dataset. This process is known as bootstrapping. Additionally, when selecting the optimal split at each node, only a random subset of features is considered, which further contributes to the diversity of the trees. The decision tree is grown to its maximum depth without pruning, resulting in a low-biased but high-variance tree. By averaging the predictions of multiple trees, the variance is reduced, leading to a more accurate and robust model [22]. The final prediction of the Random Forest model for regression is the average of the predictions made by individual trees.

3.3.2 Support Vector Regression

SVR is a powerful machine learning technique used for prediction tasks in the railway industry [23]. It is an extension of the Support Vector Machine (SVM) algorithm, originally designed for classification problems. The main objective of SVR is to find a function that best fits the data points while maintaining a specified margin around the function [24]. This is achieved by

minimizing the error between the predicted values and the actual values, subject to a certain tolerance ε.

The SVR model is formulated by introducing a set of slack variables ξ_i and ξ_i^* to measure the deviation of the training samples from the ε-insensitive zone. The optimization problem can be written as:

$$\text{minimize: } (1/2)\|w\|^2 + C\sum(\xi_i + \xi_i^*) \tag{5}$$

$$\text{subject to: } |y_i - (w \cdot x_i + b)| \leq \varepsilon + \xi_i,\ \xi_i \geq 0,\ \xi_i^* \geq 0 \tag{6}$$

where w is the weight vector, b is the bias term, x_i is the input feature vector, y_i is the value of TDR, C is a regularization parameter that controls the trade-off between minimizing the error and maximizing the margin, and ε is the insensitivity parameter. The objective function seeks to minimize the sum of the squared weights while penalizing the deviations outside the ε-insensitive zone.

To solve this optimization problem, the dual problem is formulated using Lagrange multipliers, and the resulting dual problem is a convex quadratic programming problem [24]. The kernel trick can be applied to SVR, allowing for nonlinear regression by mapping the input features into a higher-dimensional space. Some common kernel functions used in SVR include linear, polynomial, radial basis function (RBF), and sigmoid kernels. The optimal solution is then found using specialized optimization algorithms, such as the Sequential Minimal Optimization (SMO) algorithm or the gradient descent method.

3.3.3 XGBoost

XGBoost (Extreme Gradient Boosting) is a powerful and efficient machine learning algorithm, which is employed both for regression [25] and classification [26] in the railway industry. It is an implementation of gradient boosted decision trees, which combines the strengths of both tree-based models and boosting techniques [27]. Unlike Random Forest, which constructs multiple independent decision trees and combines their predictions through averaging, XGBoost builds an ensemble of weak learners (typically shallow decision trees) sequentially, where each new tree is trained to correct the errors made by the previous trees. The final prediction is obtained by summing the predictions of all the individual trees in the ensemble [27]. The objective function for XGBoost can be expressed as:

$$Obj(\theta) = L(\theta) + \Omega(\theta) \tag{7}$$

where θ represents the model parameters, $L(\theta)$ is the loss function that measures the difference between the predicted TDR and actual TDR values, and $\Omega(\theta)$ is the regularization term that penalizes the model complexity, which is defined as:

$$\Omega(\theta) = \gamma T + (1/2)\lambda\|w\|^2 \tag{8}$$

where T is the number of leaves in the tree, w is the vector of leaf weights, γ is the complexity control parameter, and λ is the L2 regularization parameter. The regularization term helps prevent overfitting by penalizing complex models.

The key innovation in XGBoost is its ability to find the optimal tree structure by using a second-order Taylor expansion to approximate the loss function [27]. This approximation enables XGBoost to learn the optimal splits and leaf weights simultaneously, making it more efficient

and accurate than traditional gradient boosting methods and differentiating it from Random Forest. Additionally, XGBoost incorporates several advanced features, such as column block for parallel learning, regularization, and early stopping, which further enhance its performance and scalability [28].

3.3.4 Artificial Neural Network

ANN is a machine learning model that draws inspiration from the biological neural networks found in the human brain [29]. ANN is capable of learning complex patterns and relationships within data, making it suitable for predicting TDR, which has extremely complicated behavior. The fundamental building block of an ANN is the artificial neuron, also known as a node or perceptron, which receives input signals from other neurons, processes them, and generates an output signal. These artificial neurons are organized into layers, with the input layer receiving the raw data features, the output layer producing the final predictions, and the hidden layers performing intermediate computations [30].

A typical feedforward ANN for regression consists of an input layer, one or more hidden layers, and an output layer. The connections between neurons in adjacent layers are associated with weights, which determine the strength of the signal transmitted between neurons [31]. In a fully connected ANN, every neuron in a layer is connected to all neurons in the subsequent layer. The output of a neuron in layer l, denoted as a_i^l, can be computed as follows:

$$a_i^l = f\left(\sum_j w_{\{ij\}}^l a_j^{l-1} + b_i^l\right) \tag{9}$$

where $w_{\{ij\}}^l$ is the weight connecting neuron j in layer $(l-1)$ to neuron i in layer l, a_j^{l-1} is the output of neuron j in layer $(l-1)$, b_i^l is the bias term for neuron i in layer l, and f is the activation function, which is chosen to be rectified linear unit (ReLU) in this study.

The learning process in an ANN involves adjusting the weights and biases to minimize the difference between the predicted TDR and actual TDR values, which is typically measured by a loss function. The backpropagation algorithm is commonly used to compute the gradients of the loss function with respect to the weights and biases [32]. These gradients are then used to update the model parameters through an optimization algorithm, which is called Adam. By iteratively updating the parameters, the ANN can learn to capture the underlying relationships between extracted features (e.g., subsurface condition) and the TDR values in the data and make robust predictions.

3.4 Prediction Models

3.4.1 Performance Metrics

Performance metrics are essential for evaluating the effectiveness of predictive models, as they quantify the differences between the predicted values and the actual data, helping researchers understand the model's accuracy and reliability. Using different performance metrics provides a more comprehensive evaluation of the predictive models for TDR. In this study, Mean Absolute Error (MAE), Mean Absolute Percentage Error (MAPE), and Root Mean Squared Error (RMSE) were employed as performance metrics. RMSE might not be the most appropriate metric for

TDR predictions, as it squares the errors, thereby placing excessive emphasis on large errors. For TDR values, which fall between 0 and 1, squaring these values will not produce the same effect, leading to a less informative evaluation of model performance. Nevertheless, RMSE has been included in the analysis for the sake of completeness and to facilitate comparisons with other studies that use RMSE. MAE and MAPE, on the other hand, are more appropriate for this task, as they provide insight into the average absolute deviation and percentage error, respectively. By combining these metrics, we can better assess the performance of the models and make more informed decisions about their suitability for predicting TDR in various applications. The equations for these metrics may be found in Section 2.1.3.

3.4.2 Splitting Dataset

To evaluate the performance of the ML models, the dataset consisting of 25,782 data points was divided into two parts: training and testing. The training set comprised 80 percent of the data (approximately 20,626 data points), while the remaining 20 percent (5,156 data points) was allocated to the test set. This division allowed the models to be trained on a larger portion of the dataset while still preserving a separate portion for evaluation purposes, ensuring that the models were tested on unseen data.

To obtain a more robust estimate of the model's performance and hyperparameters, a k-fold cross-validation technique with k=5 was employed on the training set. In k-fold cross-validation, the training set is randomly partitioned into k equal-sized subsets or "folds." The model is then trained on k-1 folds and validated on the remaining fold. This process is repeated k times, with each fold being used as the validation set once [1]. By employing k-fold cross-validation, the risk of overfitting is mitigated by exposing the model to various combinations of training and validation data.

3.4.3 Fine-tuning Hyperparameters

Hyperparameter tuning, also known as fine-tuning, is an essential step in the ML model building process. It involves adjusting the parameters of a model that are not learned during training, thereby optimizing the model's performance. Fine-tuning hyperparameters is important because the choice of hyperparameters can significantly impact a model's ability to generalize to unseen data and improve its overall performance [19]. For instance, the list of hyperparameters for XGBoost includes:

- Learning rate (η): controls the step size at each boosting iteration, with a smaller learning rate making the model more robust, but potentially requiring more boosting iterations to converge;
- Maximum depth of trees: deeper trees can capture more complex relationships in the data, but may also lead to overfitting;
- Number of trees: a higher number of trees may result in better performance, but also increases the risk of overfitting and computational cost; and
- Gamma: controls the minimum loss reduction required for a split to occur in a tree, with a larger value resulting in a more conservative model, reducing the likelihood of overfitting.

Figure 23 presents the fine-tuning of the number of trees and depth of trees for the XGBoost model. The fine-tuning of the other two hyperparameters (i.e., learning rate and Gamma) is not shown in the figure. To avoid a cluttered figure, results for only two max_depths are displayed. The variation of MAE for different folds is represented by the shaded area.

The same process of adjusting hyperparameters and evaluating their performance using k-fold cross-validation was also applied for fine-tuning the other ML models. However, the details regarding these models are not provided in this report, as the primary focus of this study is not on the process of fine-tuning the prediction models.

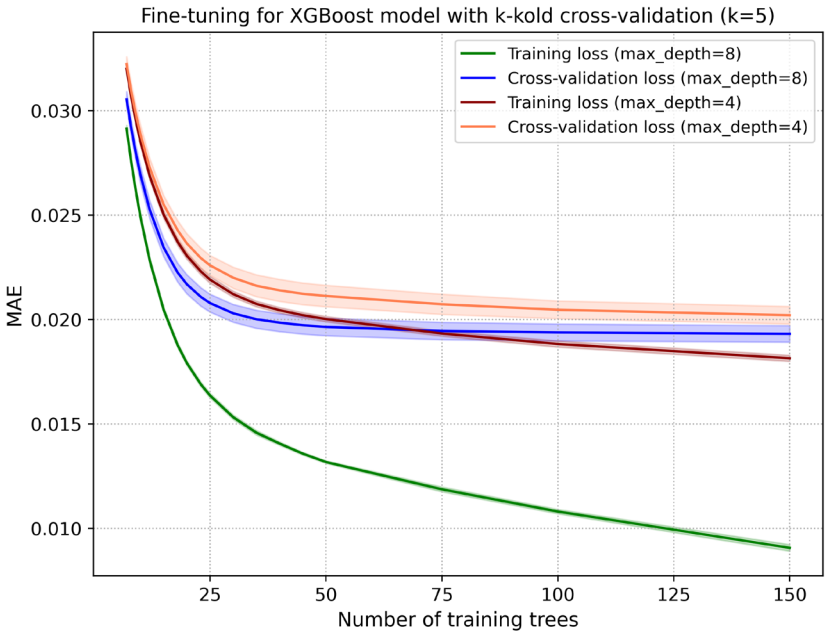


Figure 23. Fine-tuning the number of trees and maximum depth for the XGBoost model.

3.4.4 Results of Predicting TDR

After fine-tuning all prediction models, their performance was assessed on the test dataset, which the models have not previously encountered. Table 5 presents the performance of the models for both the testing and training datasets. While the performance on the training dataset does not directly evaluate the generalization of the model, it is beneficial to report it, as it can help identify potential overfitting or underfitting in the model's performance. The results indicate that the performance of Random Forest, XGBoost, and ANN models is quite impressive for predicting TDR, considering its complexity. However, XGBoost exhibits a slight edge over the other models. The SVR model was not complex enough to capture the relationship between the extracted features and TDR, resulting in less satisfactory outcomes.

Table 5. Statistical performance of developed models for predicting TDR.

Model	Testing			Training		
	MAE*	MAPE*	RMSE*	MAE*	MAPE*	RMSE*
Random Forest	0.0203	0.494	0.0301	0.0161	0.452	0.0228
SVM	0.0485	2.277	0.0535	0.0445	2.013	0.0496
XGBoost	**0.0192**	**0.460**	**0.0291**	0.0144	0.426	0.0205

ANN	0.0198	0.497	0.0294	0.0187	0.473	0.0279

*Mean Absolute Error (MAE); Mean Absolute Percentage Error (MAPE); Root Mean Squared Error (RMSE)

Figure 24 presents the scatter plot of the predicted TDR vs. the actual TDR values, specifically for the XGBoost model, as it outperformed the other models in this study. The plot illustrates the close correspondence between the predicted and actual TDR values, highlighting the effectiveness of the XGBoost model in capturing the complex relationships between the extracted features and TDR. A strong linear trend between the predicted and actual values can be observed, with most of the data points lying close to the diagonal dotted line, indicating that the model is able to provide accurate predictions for a wide range of TDR values. Data points that are farther away from the diagonal dotted line represent instances with larger prediction errors, but these cases are relatively few.

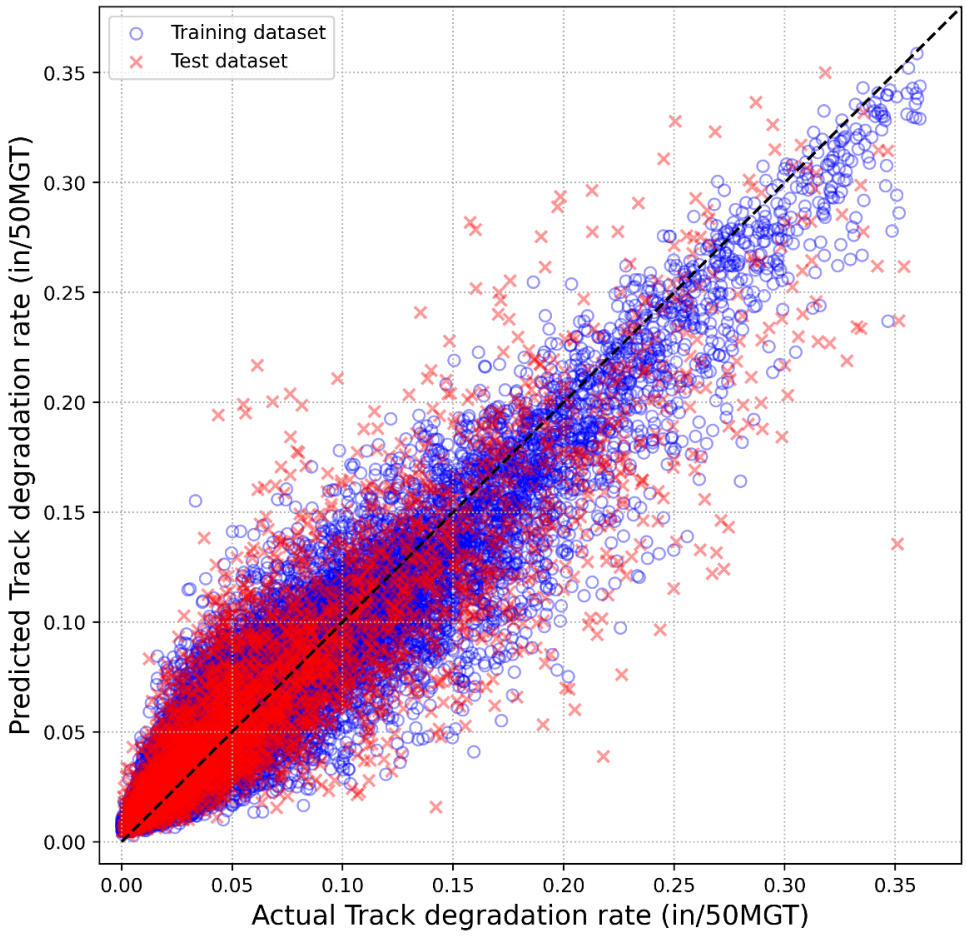


Figure 24. Scatter plot of predicted TDR vs. actual TDR values using XGBoost prediction model.

Due to the large number of data points in the test dataset (5,156), some details in Figure 25 may be difficult to discern. To better visualize the distribution and concentration of data points, a heatmap is presented in Figure 25. The heatmap clearly illustrates the proximity of data points to the diagonal dotted line, further confirming the accuracy of the XGBoost model in predicting

TDR values. Furthermore, the heatmap highlights the high concentration of data points at smaller TDR values, signifying that most blocks exhibit small TDR values.

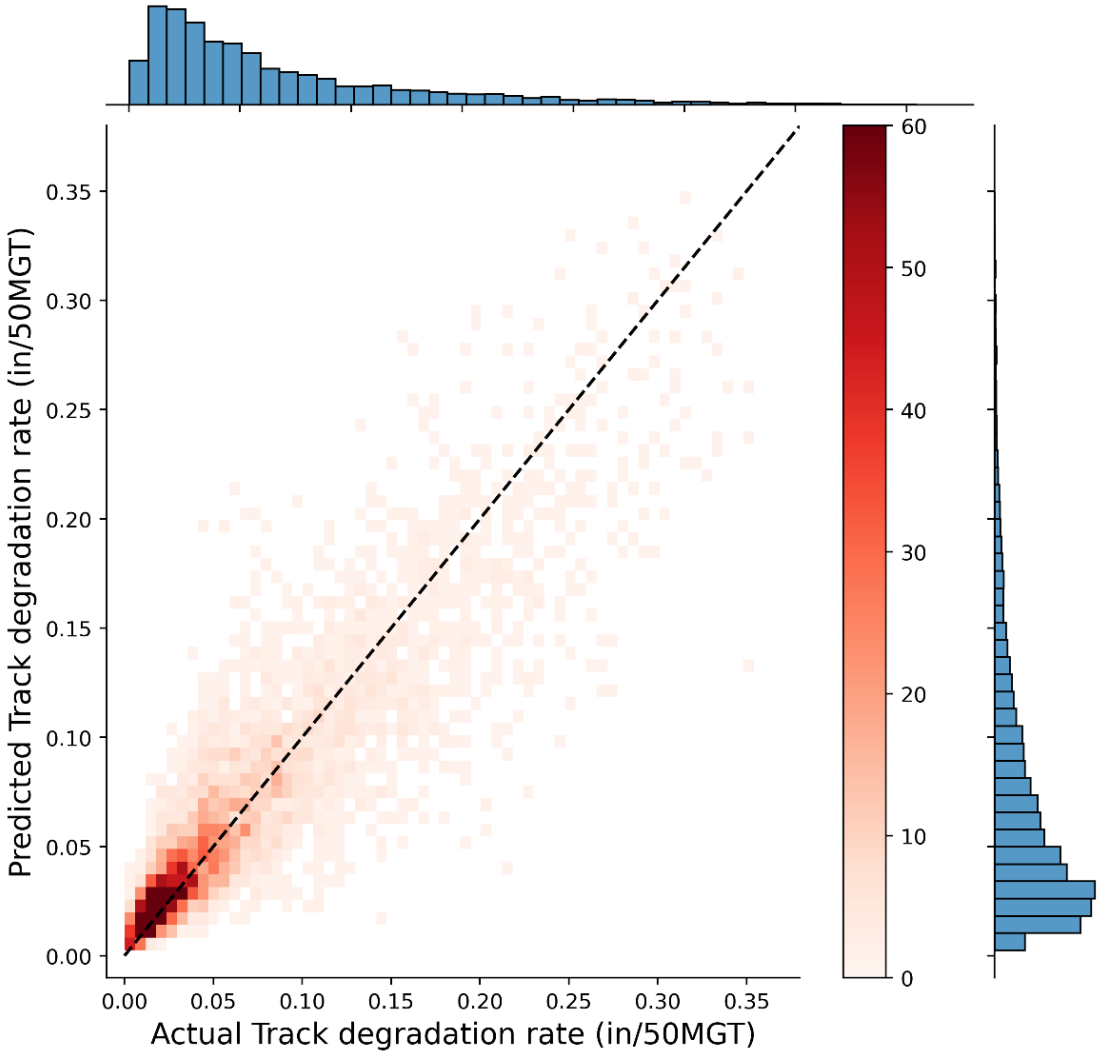


Figure 25. Concentration of data points for predicted TDR vs actual TDR in the test dataset, utilizing the XGBoost prediction model.

Figure 26 depicts the histograms of residuals for both the training and testing datasets. The residual distribution provides a visual tool to evaluate the quality of the predicted TDR values. The mean of both the training and testing residuals distributions is zero, which indicates that the model does not systematically overestimate or underestimate the TDR. Moreover, it is found that 95 percent of the residuals for the testing dataset fall within the range of -0.063 to +0.063 in/50 MGT. This indicates that most of the model's predictions deviate from the actual values by less than 0.063 in/50 MGT. This narrow range of residuals demonstrates the accuracy and reliability of the model's predictions. Figure 27 illustrates the application of the proposed model on a block as an example. The calculated 95 percent confidence interval is represented by the shaded area. In this example, not only does the predicted TDR fall within the confidence interval, but also most actual inspections are within this region, which further indicates the validity of the model.

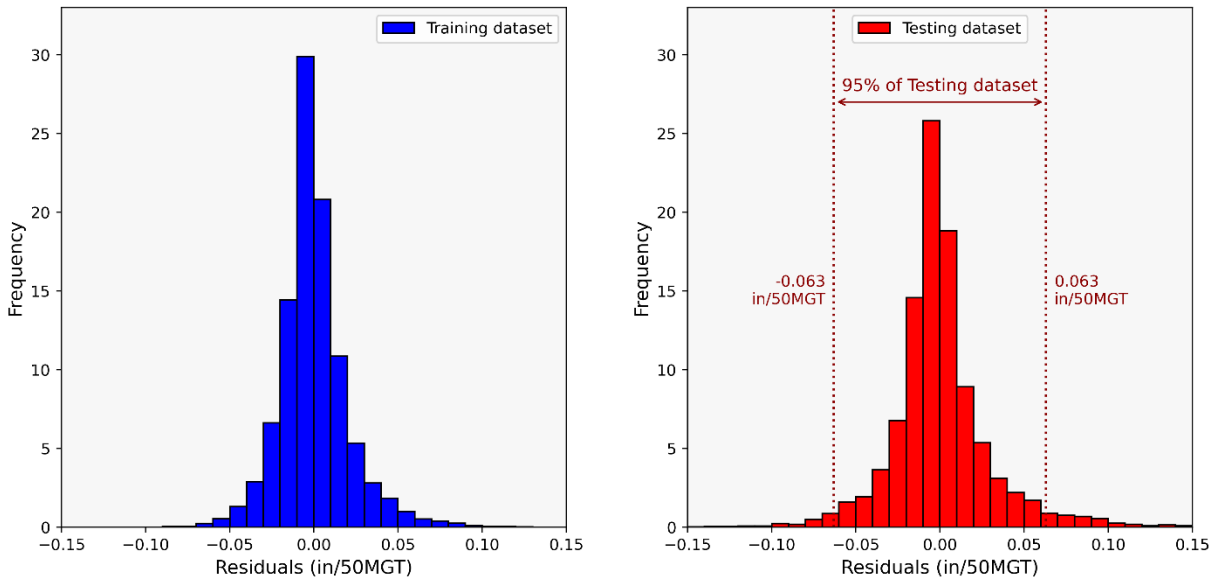


Figure 26. Histograms of residuals for both the training and testing datasets.

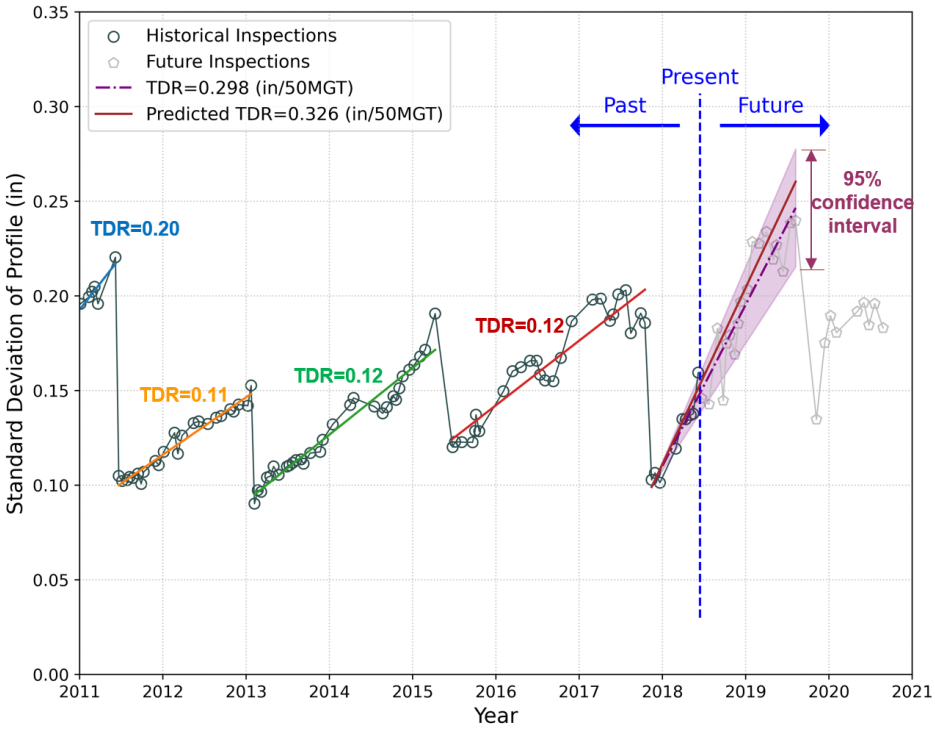


Figure 27. The application of the proposed ML model on a block as an example.

3.4.5 Feature Importance

The XGBoost model, which outperformed the other models in this study, provides a robust means of determining feature importance. This capability stems from the model's use of decision trees, where the importance of a feature is often derived from the number of times a feature is used to split the data across all trees. In other words, feature importance is quantifying a score

that indicates how valuable each feature was in the construction of the boosted decision trees within the model. More specifically, the importance of a feature is calculated based on the number of times a feature appears in a tree across all the trees in the model. Mathematically, the importance I_j of feature j is computed as:

$$I_j = \sum_{t=1}^{T} \sum_{i=1}^{I} 1 \; if \; (f_{ti} = j) \; and \; 0 \; otherwise \tag{10}$$

Where f_{ti} denotes the feature to split at the i^{th} node and t^{th} tree; T and I represent the total number of trees and total number of nodes in each tree, respectively. After calculating I_j the score is normalized to create relative importance values that sum to one.

Figure 28 displays the importance of features for predicting TDR using the XGBoost model. The results indicate that current_SD, TDR in the previous cycle, and fouling index are the most influential features for predicting TDR. The importance of the number of tampings is lower than expected. This could be attributed to the fact that the number of tampings has only been measured since 2011, when geometry data became available. Tampings conducted before 2011 were not considered due to data limitations. The feature importance for both depth of ditch and distance of ditch is relatively high, further supporting the high impact of lateral drainage quality on TDR and suggesting that the properties of ditch should be measured for more accurate TDR prediction. The influence of kurtosis on the occurrence of geometry defects was also observed by Soleimanmeigouni et al. [33].

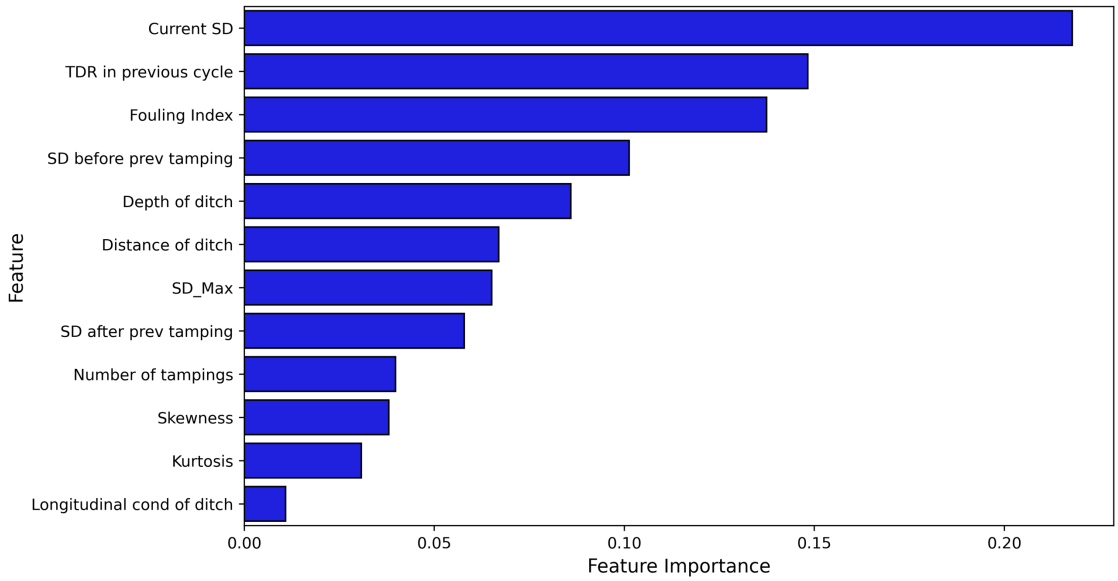


Figure 28. The importance of features for predicting TDR using the XGBoost model.

4. Conclusion and Publications

4.1 Major Conclusions

This report presents an analysis of predicting the need for maintenance operations on railway tracks. For this purpose, extensive datasets from U.S. passenger railway tracks, spanning a cumulative length of 287 miles, were examined. The research involved a detailed analysis of geometry, GPR, and LiDAR data collected from 2011 to 2021. Throughout this project, two methods were employed to estimate TDR, each tailored to suit the availability of historical data and the desired level of precision. These methods, while distinct in their approach, complement each other and collectively offer the railway industry versatile and effective tools for TDR estimation and maintenance scheduling, highlighting their reliability and adaptability. The first method uses stochastic analysis to estimate the probability of TDR exceedance based on subsurface and external drainage conditions. This approach is especially valuable in situations where data availability is limited, providing a probabilistic assessment of TDR and its impact on maintenance needs. The second method, which demands a more extensive dataset, applies a range of ML techniques to accurately predict the actual values of TDR. The major conclusions of this project are below:

- Stochastic models are proposed for estimating the exceedance probability of TDR with respect to subsurface and drainage conditions of track. Results can be used either by rebuilding the probability distributions from the presented parameters or referring to Appendix A, which presents CCDF for intervals of TDR.
- Investigating the effect of subsurface conditions indicated that the probability of having high track degradation rates in highly fouled ballast is 10 times more than in clean ballast, which indicates the detrimental effect of fouling on track behavior.
- The difference between track degradation rates of moderately fouled ballast (15<BFI<25) and fouled ballast (25<BFI<35) is significant. This indicates that increasing BFI values over 25 results in high track degradation rates.
- A framework is proposed that can automatically evaluate drainage and cross-section conditions of tracks from LiDAR data.
- TDR in conditions of poor lateral drainage, where the ditch is shallower than 2.5 ft (76 cm) and farther than 15 ft (4.57 m), is roughly 55 percent higher than tracks with good lateral drainage, where the ditch is deeper and closer than the dimensions.

- MAE of 0.019 and RMSE of 0.029 inches per 50 MGT.

- Among various machine learning models proposed for estimating the TDR, the XGBoost model was marginally superior, achieving successful predictions. The feature importance derived from the XGBoost model indicated that current_SD, TDR in the previous cycle, and the fouling index are the most influential features for predicting TDR.

4.2 Journal Reports and Conference Papers

The following journal reports and conference papers were generated from this research:

- Stochastic analysis for estimating track geometry degradation rates based on GPR and LiDAR data: *Construction and Building Materials* (also presented at *TRB 2023*) [34]
- Investigating Track Degradation Rates: A Machine Learning Approach for Maintenance Scheduling; presented at *Railway Engineering 2023* (also under revision in *Construction and Building Materials*)
- An Approach to Evaluate Track Geometry Degradation Rate (accepted for presentation in *TRB 2024*)

4.3 Recommendations for Future Research

While this research has successfully identified and used 12 significant features to predict the long-term condition of railway tracks, it is recognized that numerous other factors could potentially influence the TDR. Therefore, a promising direction for subsequent research would be the exploration and extraction of additional features that impact TDR. By expanding the feature set, future studies could refine the accuracy of predictive models, leading to more precise forecasting of TDR. This could involve the integration of new data sources, the application of more sophisticated data mining techniques, and the potential use of emerging machine learning algorithms that could uncover complex patterns and relationships within the data. Such advancements would not only enhance predictive performance but also contribute to the development of more robust and reliable maintenance strategies for railway infrastructure.

Another promising direction for future research involves leveraging geometry data to forecast the condition of various railway track components, such as the quality of ties or the subsurface condition. Theoretically, there are correlations between geometry data and other components of track. By identifying and exploiting these hidden correlations, it may be possible to predict the state of different track elements using primarily geometry data. This methodology holds considerable potential due to the more frequent collection of geometry data compared to other types of data. Consequently, this approach could enable quicker and more efficient diagnosis of issues in track components, leading to improved maintenance strategies and overall track safety.

5. References

[1] Goodarzi, S., Kashani, H. F., Oke, J., & Ho, C. L. (2022). Data-driven methods to predict track degradation: A case study. *Construction and Building Materials*, *344*, p. 128166.

[2] Attoh-Okine, N. O. (2017). *Big data and differential privacy: analysis strategies for railway track engineering*. John Wiley & Sons.

[3] Matthews, H. S., Hendrickson, C. T., & Weber, C. L. (2008). The importance of carbon footprint estimation boundaries. ACS Publications.

[4] Li, D., Hyslip, J., Sussmann, T., & Chrismer, S. (2015). *Railway geotechnics*. CRC Press.

[5] Kashani, H. F., Ho, C. L., & Hyslip, J. P. (2018). Fouling and water content influence on the ballast deformation properties. *Construction and Building Materials*, *190*, pp. 881–895.

[6] Rohrman, A. K., Kashani, H. F., & Ho, C. L. (2020). Effects of natural abrasion on railroad ballast strength and deformation properties. *Construction and Building Materials*, *247*, p. 118315.

[7] Mamou, A., Powrie, W., Priest, J. A., & Clayton, C. (2017). The effects of drainage on the behaviour of railway track foundation materials during cyclic loading. *Géotechnique*, *67*(10), pp. 845–854.

[8] Chrismer S. M., & Selig, E. T. (1991). Mechanics-based model to predict ballast-related maintenance timing and costs. *International Heavy Haul Railway Conference*, Vancouver, CANADA.

[9] Federal Railroad Administration (2018). Track and Rail and Infrastructure Integrity Compliance Manual: Volume II - Chapter 1 &2, Track Safety Standards.

[10] Goodarzi, S., Kashani, H. F., Chrismer, S., & Ho, C. L. (2021). Using large datasets for finding the correlation between the rate of track settlement and changes in geometry indices. *Transportation Geotechnics*, *31*, p. 100665.

[11] Mosayebi, S.-A., Esmaeili, M., & Zakeri, J.-A. (2020). Dynamic train–track interactions and stress distribution patterns in ballasted track layers. *Journal of Transportation Engineering, Part B: Pavements*, *146*(1), p. 4019042.

[12] Vale, C., & Lurdes, S. M. (2013). Stochastic model for the geometrical rail track degradation process in the Portuguese railway Northern Line. *Reliability Engineering & System Safety*, *116*, pp. 91–98.

[13] Kashani, H. F., Ho, C. L., Clement, W. P., & Oden, C. P. (2016). Evaluating the correlation between the geotechnical index and the electromagnetic properties of fouled ballasted track by a full-scale laboratory model. *Transportation Research Record: Journal of the Transportation Research Board*, *2545*(1), pp. 66–78.

[14] Silvast, M., Nurmikolu, A., Wiljanen, B., & Levomaki, M. (2010). An inspection of railway ballast quality using ground penetrating radar in Finland. *Proceedings of the Institution of Mechanical Engineers, Part F: Journal of Rail and Rapid Transit*, *224*(5), pp. 345–351.

[15] Kashani, H. F., Hyslip, J. P., & Ho, C. L. (2017). Laboratory evaluation of railroad ballast behavior under heavy axle load and high traffic conditions. *Transportation Geotechnics*, *11*, pp. 69–81.

[16] Pham, H. (2006). *Springer handbook of engineering statistics*, *49*. Springer.

[17] Ivošević, Š., Meštrović, R., & Kovač, N. (2019). Probabilistic estimates of corrosion rate of fuel tank structures of aging bulk carriers. *International Journal of Naval Architecture and Ocean Engineering*, *11*(1), pp. 165–177.

[18] Goodarzi, S., Kashani, H. F., Saeedi, A., Oke, J., & Ho, C. L. (2023). Stochastic analysis for estimating track geometry degradation rates based on GPR and LiDAR data. *Construction and Building Materials*, *369*, p. 130591.

[19] James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). *An introduction to statistical learning*, *112*. Springer.

[20] Breiman, L. (2001). Random forests. *Machine Learning*, *45*(1), pp. 5–32.

[21] Biau, G. (2012). Analysis of a random forests model. *Journal of Machine Learning Research*, *13*(1), pp. 1063–1095.

[22] Sharma, S., Cui, Y., He, Q., Mohammadi, R., & Li, Z. (2018). Data-driven optimization of railway maintenance for track geometry. *Transportation Research Part C: Emerging Technologies*, *90*(September), pp. 34–58.

[23] Huang, P., Wen, C., Fu, L., Peng, Q., & Li, Z. (2020). A hybrid model to improve the train running time prediction ability during high-speed railway disruptions. *Safety Science*, *122*, p. 104510.

[24] Drucker, H., Burges, C. J., Kaufman, L., Smola, A., & Vapnik, V. (1996). Support vector regression machines. *Advances in Neural Information Processing Systems*, *9*.

[25] Zhang, C., Zhao, Q., Shen, T., & Sun, B. (2022). Rail Defect Detection Method Based on Improved XGBoost. *Proceedings of the 12th International Conference on Computer Engineering and Networks*, pp. 911–920.

[26] Mohammadi, R., He, Q., Ghofrani, F., Pathak, A., & Aref, A. (2019). Exploring the impact of foot-by-foot track geometry on the occurrence of rail defects. *Transportation Research Part C: Emerging Technologies*, *102*, pp. 153–172.

[27] Chen T., & Guestrin, C. (2016). Xgboost: A scalable tree boosting system. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 785–794.

[28] Hak Lee, E., Kim, K.,. Kho, S.-Y., Kim, D.-K., & Cho, S.-H. (2021). Estimating express train preference of urban railway passengers based on extreme gradient boosting (XGBoost) using smart card data. *Transportation Research Record: Journal of the Transportation Research Board*, *2675*(11), pp. 64–76.

[29] Braspenning, P. J., Thuijsman, F., & Weijters, A. J. M. M. (1995). *Artificial neural networks: an introduction to ANN theory and practice*, *931*. Springer Science & Business Media.

[30] Popov, K., *et al.* (2022). Rail Track Monitoring Using AI and Machine Learning. *Trends on Construction in the Digital Era: Proceedings of ISIC 2022*, pp. 513–523.

[31] Alidoust, P., Goodarzi, S., Tavana Amlashi, A., & Sadowski, Ł. (2023). Comparative analysis of soft computing techniques in predicting the compressive and tensile strength of seashell containing concrete. *European Journal of Environmental and Civil Engineering*, 27(5), pp. 1853–1875.

[32] Khajehei, H., Ahmadi, A., Soleimanmeigouni, I., Haddadzade, M., Nissen, A., & Latifi Jebelli, M. J. (2021). Prediction of track geometry degradation using artificial neural network: a case study. *International Journal of Rail Transportation*, pp. 1–20.

[33] Design, L., Soleimanmeigouni, I., Ahmadi, A., Nissen, A., & Xiao, X. (2019). Prediction of railway track geometry defects : a case study. *Structure and Infrastructure Engineering*, *0*(0), pp. 1–15.

[34] Goodarzi, S., Kashani, H. F., Saeedi, A., Oke J., and Ho, C.L. (2023). Stochastic analysis for estimating track geometry degradation rates based on GPR and LiDAR data. *Construction and Building Materials, 369.*

Appendix A.
The Exceedance Probability (i.e., CCDF) for Intervals of TDR in Each Category

TDR (in/50MGT)	G1 and Cut	G1 and AGO	G1 and Fill	G2 and Cut	G2 and AGO	G2 and Fill	G3 and Cut	G3 and AGO	G3 and Fill	G4 and Cut	G4 and AGO	G4 and Fill
0.00	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
0.02	0.833	0.673	0.673	0.786	0.735	0.700	0.940	0.857	0.815	0.974	0.922	0.957
0.04	0.596	0.445	0.443	0.571	0.514	0.476	0.832	0.714	0.671	0.914	0.829	0.897
0.06	0.425	0.301	0.303	0.423	0.361	0.341	0.718	0.596	0.555	0.842	0.747	0.837
0.08	0.313	0.208	0.214	0.321	0.261	0.254	0.615	0.500	0.462	0.769	0.676	0.779
0.10	0.239	0.147	0.156	0.249	0.195	0.195	0.526	0.422	0.386	0.699	0.614	0.726
0.12	0.188	0.106	0.117	0.198	0.150	0.154	0.451	0.359	0.324	0.634	0.559	0.676
0.14	0.152	0.078	0.089	0.159	0.118	0.124	0.389	0.306	0.274	0.574	0.511	0.629
0.16	0.125	0.059	0.070	0.130	0.096	0.101	0.336	0.262	0.232	0.519	0.467	0.584
0.18	0.105	0.044	0.055	0.107	0.079	0.084	0.291	0.226	0.197	0.469	0.428	0.543
0.20	0.089	0.035	0.045	0.090	0.067	0.071	0.255	0.196	0.169	0.423	0.395	0.503
0.22	0.077	0.027	0.037	0.076	0.056	0.060	0.221	0.169	0.144	0.382	0.361	0.466
0.24	0.067	0.021	0.030	0.064	0.048	0.051	0.194	0.147	0.123	0.344	0.331	0.430
0.26	0.059	0.017	0.025	0.055	0.042	0.044	0.170	0.128	0.106	0.309	0.304	0.397
0.28	0.052	0.013	0.021	0.048	0.037	0.038	0.150	0.112	0.092	0.278	0.280	0.364
0.30	0.047	0.011	0.018	0.041	0.033	0.033	0.132	0.099	0.080	0.249	0.257	0.334
0.32	0.042	0.009	0.015	0.036	0.029	0.029	0.117	0.087	0.069	0.223	0.235	0.304
0.34	0.038	0.007	0.013	0.032	0.026	0.026	0.104	0.077	0.060	0.200	0.215	0.276
0.36	0.035	0.006	0.012	0.028	0.024	0.023	0.092	0.068	0.053	0.178	0.197	0.250
0.38	0.032	0.005	0.010	0.025	0.021	0.021	0.082	0.060	0.046	0.158	0.180	0.224
0.40	0.029	0.004	0.009	0.022	0.020	0.018	0.073	0.053	0.040	0.141	0.164	0.199
0.42	0.027	0.004	0.008	0.020	0.018	0.017	0.065	0.048	0.035	0.124	0.149	0.176
0.44	0.025	0.003	0.007	0.017	0.016	0.015	0.058	0.042	0.031	0.110	0.135	0.153
0.46	0.023	0.003	0.006	0.016	0.015	0.013	0.052	0.038	0.028	0.096	0.122	0.132

TDR (in/50MGT)	G1 and Cut	G1 and AGO	G1 and Fill	G2 and Cut	G2 and AGO	G2 and Fill	G3 and Cut	G3 and AGO	G3 and Fill	G4 and Cut	G4 and AGO	G4 and Fill
0.48	0.021	0.002	0.006	0.014	0.014	0.012	0.047	0.034	0.024	0.084	0.110	0.111
0.50	0.020	0.002	0.005	0.013	0.013	0.011	0.042	0.030	0.022	0.073	0.099	0.092
0.52	0.019	0.002	0.004	0.012	0.012	0.010	0.038	0.027	0.019	0.064	0.088	0.074
0.54	0.017	0.001	0.004	0.010	0.011	0.009	0.034	0.025	0.017	0.055	0.078	0.071
0.56	0.016	0.001	0.004	0.010	0.011	0.008	0.030	0.022	0.015	0.047	0.069	0.068
0.58	0.015	0.001	0.003	0.009	0.010	0.008	0.027	0.020	0.014	0.040	0.060	0.065
0.60	0.015	0.001	0.003	0.008	0.009	0.007	0.025	0.018	0.012	0.033	0.052	0.058
0.62	0.014	0.001	0.003	0.007	0.009	0.007	0.022	0.016	0.011	0.028	0.045	0.048
0.64	0.013	0.001	0.003	0.007	0.008	0.006	0.020	0.015	0.010	0.023	0.038	0.041
0.66	0.012	0.001	0.002	0.006	0.008	0.006	0.018	0.013	0.009	0.019	0.032	0.035
0.68	0.012	0.001	0.002	0.006	0.007	0.005	0.016	0.012	0.008	0.015	0.026	0.029
0.70	0.011	0.001	0.002	0.005	0.007	0.005	0.015	0.011	0.007	0.012	0.021	0.026
0.72	0.011	0.000	0.002	0.005	0.007	0.004	0.014	0.010	0.006	0.009	0.016	0.018
0.74	0.010	0.000	0.002	0.004	0.006	0.004	0.012	0.009	0.006	0.007	0.012	0.015
0.76	0.010	0.000	0.002	0.004	0.006	0.004	0.011	0.008	0.005	0.005	0.009	0.012
0.78	0.009	0.000	0.001	0.004	0.006	0.004	0.010	0.008	0.005	0.004	0.006	0.009
0.80	0.009	0.000	0.001	0.004	0.005	0.003	0.009	0.007	0.004	0.002	0.004	0.005

Abbreviations and Acronyms

ACRONYM	DEFINITION
A_i	Actual Value (ith)
AGO	At-Grade or Other (profile category)
ANN	Artificial Neural Network
BFI	Ballast Fouling Index
CvM	Cramér Von-Mises (goodness of fit test)
FRA	Federal Railroad Administration
BFI_Avg_Max	Subroutine to calculate the average for each channel then considers the maximum value of three calculated averages
BFI_Avg_Avg	Subroutine to calculate the average for each channel then considers the average value of three calculated averages (i.e., overall average)
BFI_Max_Max	Subroutine to calculate the maximum value for each channel then considers the maximum value of three calculated maximums (i.e., overall maximum)
BFI_Max_Avg	Subroutine to calculate the maximum value for each channel then considers the average value of three calculated maximums
CCDF	Complementary Cumulative Distribution Function
CDF	Cumulative Distribution Function
GPR	Ground Penetrating Radar
GPS	Global Positioning System
K-S	Kolmogorov–Smirnov (goodness of fit test)
LiDAR	Light Detection and Ranging
MAD	Mean of Absolute Deviation
MAE	Mean Absolute Error
MAPE	Mean of Absolute Percentage Error
MGT	Million Gross Tons
ML	Machine Learning
N	Number of Inspections
P_i	Predicted Value (ith)
P-P	Probability comparison of two sets
Q-Q	Quantile comparison of two sets
RBF	Radial Basis Function
RF	Random Forest (ML method)

ACRONYM	DEFINITION
RMSE	Root Mean Squared Error
SD	Standard Deviation
SMO	Sequential Minimal Optimization
SVR	Support Vector Regressor (ML method)
T1, T2, T3	Tracks 1, 2, and 3
TDR	Track Degradation Rate
TOT	Top of Tie
XGBoost	Extreme Gradient Boosting (ML method)
χ^2	Chi squared