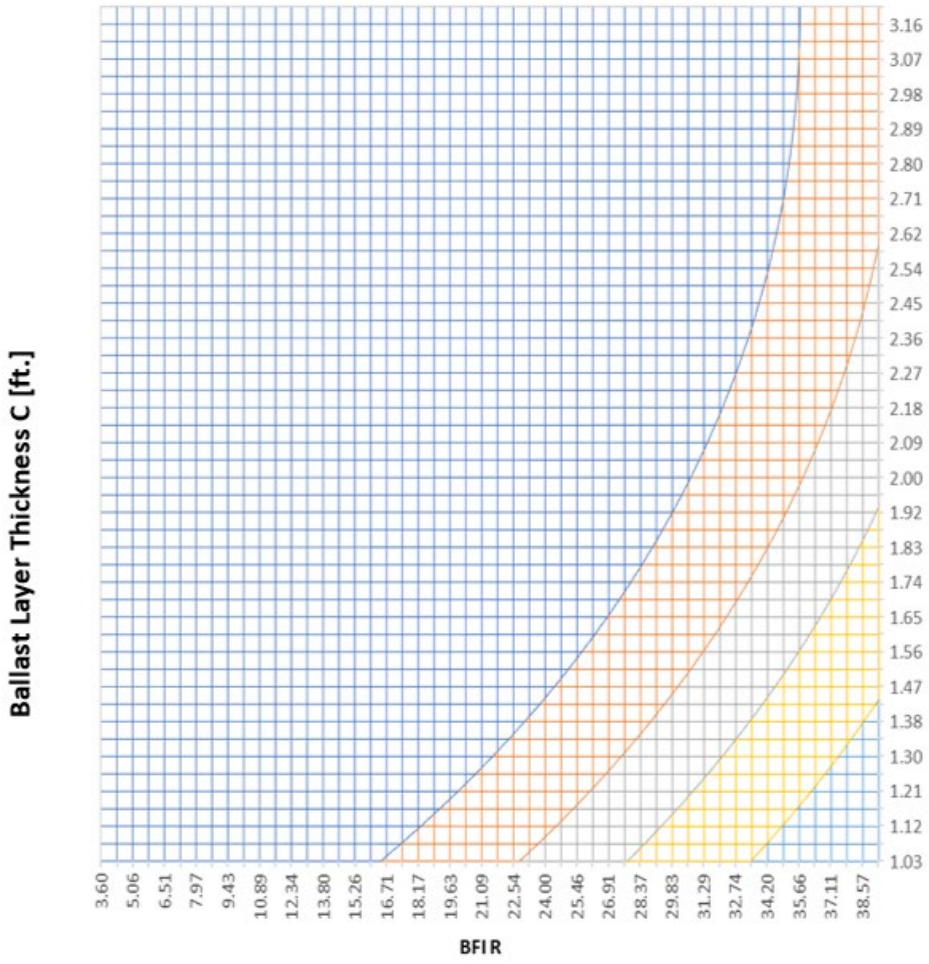
U.S. Department of Transportation

Federal Railroad Administration

Relationship Between Track Geometry Defects and Measured Track Subsurface Condition

Office of Research,
Development
and Transportation
Washington, DC 20590



DOT/FRA/ORD-20/07

Final Report
February 2020

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 0704-0188

1. AGENCY USE ONLY (Leave blank)	2. REPORT DATE	3. REPORT TYPE AND DATES COVERED
	February 2020	Technical Report, Sept. 1, 2016–March 31, 2019

4. TITLE AND SUBTITLE
Relationship Between Track Geometry Defects and Measured Track Subsurface Condition

5. FUNDING NUMBERS
DTFR53-16-C-00021

6. AUTHOR(S)
Dr. Allan Zarembski, Dennis Yurlov, Joseph Palese, and Dr. Nii Attoh-Okine

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)
University of Delaware
Newark, DE 19716

8. PERFORMING ORGANIZATION REPORT NUMBER
CIEG372133

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)
U.S. Department of Transportation
Federal Railroad Administration
Office of Railroad Policy and Development
Office of Research, Development and Technology
Washington, DC 20590

10. SPONSORING/MONITORING AGENCY REPORT NUMBER
DOT/FRA/ORD-20/07

11. SUPPLEMENTARY NOTES
COR: Hugh Thompson

12a. DISTRIBUTION/AVAILABILITY STATEMENT
This document is available to the public through the FRA website.

12b. DISTRIBUTION CODE

13. ABSTRACT (Maximum 200 words)

This report presented the results of a comprehensive study regarding the development of a probability model for occurrence of track geometry defects as a function of key subgrade parameters, measured by Ground Penetrating Radar (GPR). The analysis used multiple track geometry runs and the associated track geometry degradation behavior combined with GPR data. Statistical analyses examined the relationship between the probability of significant geometry degradation and measured GPR parameters (e.g., ballast fouling index [BFI], ballast layer thickness [BLT]). A data analytics approach used hybrid analysis to include hierarchical clustering analysis with histogram data, Logistic Regression (LR) analysis, and performed an application of higher degree polynomial model. The result was a higher order polynomial LR model for determination of the probability of a track geometry surface defect occurring at locations with measured ballast fouling and measured ballast thickness. The results showed that there was a statistically significant relationship between high rates of geometry degradation and poor subsurface conditions as defined by the GPR parameters: BFI and BLT. Furthermore, the development of a predictive model determined the probability of a high rate of geometry degradation as a function of these key GPR parameters.

14. SUBJECT TERMS
Railroad, track, track geometry, Ground Penetrating Radar, GPR, big data, ballast fouling index, BFI, ballast layer thickness, BLT, Logistic Regression, LR, research, study

15. NUMBER OF PAGES
162

16. PRICE CODE

17. SECURITY CLASSIFICATION OF REPORT	18. SECURITY CLASSIFICATION OF THIS PAGE	19. SECURITY CLASSIFICATION OF ABSTRACT	20. LIMITATION OF ABSTRACT
Unclassified	Unclassified	Unclassified	

METRIC/ENGLISH CONVERSION FACTORS

ENGLISH TO METRIC

LENGTH (APPROXIMATE)

1 inch (in)	=	2.5 centimeters (cm)
1 foot (ft)	=	30 centimeters (cm)
1 yard (yd)	=	0.9 meter (m)
1 mile (mi)	=	1.6 kilometers (km)

AREA (APPROXIMATE)

1 square inch (sq in, in^2)	=	6.5 square centimeters (cm^2)
1 square foot (sq ft, ft^2)	=	0.09 square meter (m^2)
1 square yard (sq yd, yd^2)	=	0.8 square meter (m^2)
1 square mile (sq mi, mi^2)	=	2.6 square kilometers (km^2)
1 acre = 0.4 hectare (he)	=	4,000 square meters (m^2)

MASS - WEIGHT (APPROXIMATE)

1 ounce (oz)	=	28 grams (gm)
1 pound (lb)	=	0.45 kilogram (kg)
1 short ton = 2,000 pounds (lb)	=	0.9 tonne (t)

VOLUME (APPROXIMATE)

1 teaspoon (tsp)	=	5 milliliters (ml)
1 tablespoon (tbsp)	=	15 milliliters (ml)
1 fluid ounce (fl oz)	=	30 milliliters (ml)
1 cup (c)	=	0.24 liter (l)
1 pint (pt)	=	0.47 liter (l)
1 quart (qt)	=	0.96 liter (l)
1 gallon (gal)	=	3.8 liters (l)
1 cubic foot (cu ft, ft^3)	=	0.03 cubic meter (m^3)
1 cubic yard (cu yd, yd^3)	=	0.76 cubic meter (m^3)

TEMPERATURE (EXACT)

[(x-32)(5/9)] °F = y °C

METRIC TO ENGLISH

LENGTH (APPROXIMATE)

1 millimeter (mm)	=	0.04 inch (in)
1 centimeter (cm)	=	0.4 inch (in)
1 meter (m)	=	3.3 feet (ft)
1 meter (m)	=	1.1 yards (yd)
1 kilometer (km)	=	0.6 mile (mi)

AREA (APPROXIMATE)

1 square centimeter (cm^2)	=	0.16 square inch (sq in, in^2)
1 square meter (m^2)	=	1.2 square yards (sq yd, yd^2)
1 square kilometer (km^2)	=	0.4 square mile (sq mi, mi^2)
10,000 square meters (m^2)	=	1 hectare (ha) = 2.5 acres

MASS - WEIGHT (APPROXIMATE)

1 gram (gm)	=	0.036 ounce (oz)
1 kilogram (kg)	=	2.2 pounds (lb)
1 tonne (t)	=	1,000 kilograms (kg)
	=	1.1 short tons

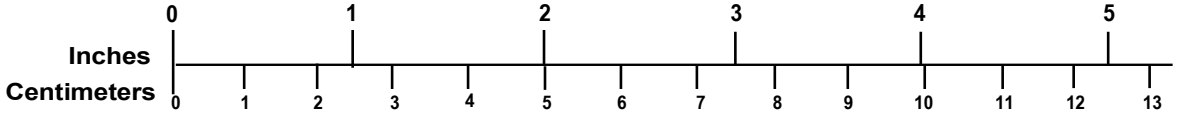
VOLUME (APPROXIMATE)

1 milliliter (ml)	=	0.03 fluid ounce (fl oz)
1 liter (l)	=	2.1 pints (pt)
1 liter (l)	=	1.06 quarts (qt)
1 liter (l)	=	0.26 gallon (gal)
1 cubic meter (m^3)	=	36 cubic feet (cu ft, ft^3)
1 cubic meter (m^3)	=	1.3 cubic yards (cu yd, yd^3)

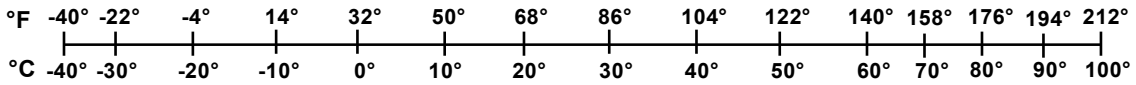
TEMPERATURE (EXACT)

[(9/5) y + 32] °C = x °F

QUICK INCH - CENTIMETER LENGTH CONVERSION



QUICK FAHRENHEIT - CELSIUS TEMPERATURE CONVERSIO



For more exact and or other conversion factors, see NIST Miscellaneous Publication 286, Units of Weights and Measures. Price $2.50 SD Catalog No. C13 10286

Updated 6/17/98

Acknowledgements

The authors would like to acknowledge the Federal Railroad Administration's (FRA) Hugh Thompson, Program Manager of the Track Research division and FRA's Gary Carr, former Chief Supervisory Engineer of the Track Research division, and Volpe National Transportation Systems Center's Ted Sussmann for their support, advice and assistance during this activity. The authors would also like to acknowledge and thank Mike Trosino and Amanda Kessler of Amtrak[1] for providing the track geometry data, and Jim Hyslip of HyGround Engineering for providing the Ground Penetrating Radar (GPR) data used in this analysis.

[1] Amtrak has approved the use of their data for the purposes of this research.

Contents

Illustrations

Tables

Executive Summary

Recent research shows a relationship between track geometry defects and track subsurface conditions as measured by Ground Penetrating Radar (GPR). This report presents the results of a comprehensive study funded by the Federal Railroad Administration (FRA) from 2016 and 2019. The study was conducted by the University of Delaware at their facility to review the development of a probability model for the growth of track geometry defects as a function of key subgrade parameters as measured by GPR.

The analysis made use of multiple track geometry runs and the associated track geometry degradation behavior combined with GPR data to include the ballast fouling index (BFI), moisture content, and ballast layer thickness (BLT). Correlation and statistical analyses were performed looking at the relationship between the probability of significant geometry degradation and measured GPR parameters (e.g., BFI and BLT).

A first order Logistic Regression (LR) model was developed showing a well-defined relationship between track geometry degradation and poor subsurface condition as defined by the ballast fouling (BFI and BLT).

A second order data analytics approach was performed using a hybrid analysis to include a hierarchical clustering analysis with histogram data, LR analysis, and an application of higher degree polynomial. The result was a high order polynomial LR model for determination of the probability of a track geometry surface defect occurring at locations with measured ballast fouling and measured ballast thickness. This model showed good correlation with the data and good predictive behavior.

The model results showed that there was a statistically significant relationship between high rates of geometry degradation and poor subsurface condition as defined by the GPR parameters: BFI and BLT. The predictive model allowed for the determination of the probability of a high rate of geometry degradation as a function of these key GPR parameters.

1. Introduction

Measurement of track geometry is one of the key track inspection approaches used for maintenance of the track structure for all types of railway operations to include the full range of passenger and freight operations. These track geometry measurements were generally used on a threshold exceedance basis, i.e., when a geometry parameter exceeds a predefined maintenance or safety limit. In addition, using statistically defined Track Quality Indices (TQI) based on these measurements, forecasting the rate of track geometry degradation was performed on a limited basis, usually in a detailed study environment [1] [2] [3]. Ground Penetrating Radar (GPR) is one of a new generation of inspection technologies that are being implemented in conjunction with and support of the fundamental track geometry measurements.

1.1 Background

The railroad industry in general and the Federal Railroad Administration (FRA) are actively involved in the development of improved track inspection technologies focusing on the track structure/substructure which includes crossties, ballast, sub-ballast, and subgrade elements of the track structure. This portion of the track structure is the primary area associated with track geometry degradation. The ability to predict the development of safety related track defects, such as geometry defects, using these improved inspection technologies would be of real value to both the railroads and to the FRA Office of Railroad Safety which is tasked with monitoring the safety of the railroads. One such inspection tool is GPR.

The relationship between these track substructure inspection tools and development of track geometry defects is key interest. This is because track geometry derailments represent one of largest category of track caused derailments. This relationship between track structure/substructure performance and geometry defects was discussed in the literature, but there is no well accepted and validated relationship between potential for developing a geometry defect and condition of track. Furthermore, there was little research into the forecasting of geometry defect development as a function of measured subsurface track parameters, particularly in the development of a quantitative relationship between these subsurface parameters and the probability of a geometry defect developing at the point of measurement.

1.2 Objectives

The objective of this research effort was to develop statistical relationships between inspection parameters from different subsurface and surface inspection technologies and the development of track geometry defects and other manifestations of track geometry degradation. These subsurface inspection technologies are to include technologies currently used by railroads as well as technologies currently undergoing developed or demonstration by FRA, such as technologies being implemented on the U.S. Department of Transportation (DOT) X218 car. This includes such measurement technologies as GPR and track deflection measurements (MRail). The objective was to allow for identification of track locations with the potential for development, growth, and propagation of track geometry defects.

1.3 Overall Approach

As part of this activity, analysis algorithms to correlate multiple inspection parameters, from the different inspection technologies were developed and evaluated. This inspection data was then correlated with track geometry data obtained from FRA and several U.S. railroads, as measured by their track geometry inspection vehicles. As such, the track geometry data included continuous (foot by foot) measurement of the individual track geometry parameters as well as exception data that included the type of geometry defect, its location, size, and date of detection.

The research focused on the use of multi-variate analysis tools combined with understanding of the interrelation of the different inspection technologies and their primary output parameters, as well as their correlation with the occurrence of geometry defects to allow for this combination of complementary data into useable inspection information. Such techniques as Logistic Regression (LR) and a more comprehensive data analytics-based approach using hybrid analysis were employed to develop relationships between subsurface inspection technologies (GPR, MRail) and development of track geometry defects. The result is a statistically significant relationship between track geometry defects and key track subsurface conditions as measured by GPR; specifically ballast fouling as measured by the ballast fouling index (BFI) and ballast layer thickness (BLT).

1.4 Scope

The scope of this activity encompassed two emerging inspection technologies, GPR and MRail. However, data issues with MRail resulting in a major focus on GPR, and the relationship between GPR measured ballast conditions and the development of track geometry defects. The resulting analysis addressed the relationship between the probability of significant geometry degradation and measured GPR parameters (e.g., BFI, BLT). Other inspection technologies were not considered in this work.

The overall relationship between GPR and track geometry deterioration can be of substantial value to railroad maintenance managers and maintenance planners because it allows for the prediction of the occurrence of geometry defects, particularly severe defects that can result in track slow orders or even derailments. By planning for this type of geometry maintenance, maintenance costs can be reduced through more efficient planning and scheduling, and failures can be averted.

1.5 Organization of the Report

This report is organized in order of the analysis approach used. Section 1 provides the introduction and background. Section 2 discusses the input data. Section 3 discusses data preparation and Exploratory Data Analysis (EDA). Section 4 discusses the initial LR analyses. Sections 5 and 6 discusses expanded analyses to include expanded LR analysis (Section 5) and a new hybrid analysis combining hierarchical clustering with a high polynomial based LR analysis. Section 7 summarizes the results and potential applications of the research work.

2. Input Data

The focus of this study is on two subsurface inspections technologies; specifically, ground penetrating radar and vertical track deflection as collected on several U.S. railroads. These two subsurface inspection technologies can be described as follows:

- GPR uses a reflection of radar waves in the 300 to 400 MHz range to identify conditions in the ballast, sub-ballast, and subgrade. This technology has advanced to the point where it is considered useful, and railroads began to deploy this technology as a supplement to their traditional track geometry inspections. To date GRP is used by FRA, Amtrak, Burlington Northern Santa Fe Railway (BNSF), Norfolk Southern Corporation (NS), and Union Pacific Railway (UP).
- Vertical deflection measurements (MRail) make use of deflection of the track under load. This inspection technique is based on the Beam on Elastic Foundation theory, which develops a support parameter k (or u)[2] to define the track support condition. MRail uses a loaded and unloaded track measurement to determine a value for the track support condition. To date MRail was used, on some basis, by FRA, BNSF, Canadian Pacific Railway (CP), and UP.

As noted previously, these technologies are to be correlated with track geometry data to include continuous (foot by foot) measurement of the individual track geometry parameters as well as exception data.

2.1 Track Geometry Data

Track geometry data in the form of exception reports, to include red and yellow level exceptions, was obtained from:

- CSX Transportation: CSX's Peninsula Subdivision milepost (MP) 67–69, a 2 mile stretch of track on CSX, taken by the DOTX218 car on April 5, 2016
- Amtrak: Continuous track geometry data for 1.6 miles of track near Oakington Road, Havre de Grace, MD (see Figure 1); Track 2, MP 62.6–to 64.0; this represents monthly continuous geometry data from June 2013 to September 2016, a total of 31 runs

[2] The track support parameter (track modulus), k, sometimes referred to as u, represents the effect of the crossties, fasteners, tie pads, ballast, and sub-grades, which support the rail.

Figure 1: Amtrak Oakington Road site MP 62.6 to 61

As noted, the primary CSX data was a foot by foot measurement run by the FRA DOTX218 car on April 5, 2016, on CSX's Peninsula Subdivision between MP 67 and 69. In addition, CSX provided track geometry exception data for 5 years of inspection from 2008 to 2012.

In addition, as noted, track geometry data files were obtained from Amtrak on the Northeast Corridor near Oakington Road, Havre de Grace, MD, near MP 63.7 between Philadelphia and Washington, DC. This data is continuous track geometry measurement data (as opposed to exception data provided by BNSF and CSX) and represents approximately 1.6 miles of data representing multiple subgrade conditions, see Figure 1. Approximately 4 years' worth of geometry data was available representing monthly track geometry inspections. Approximately 31 data files were provided that contained key track geometry parameters to include surface (left, right) and profile (see Figure 2).

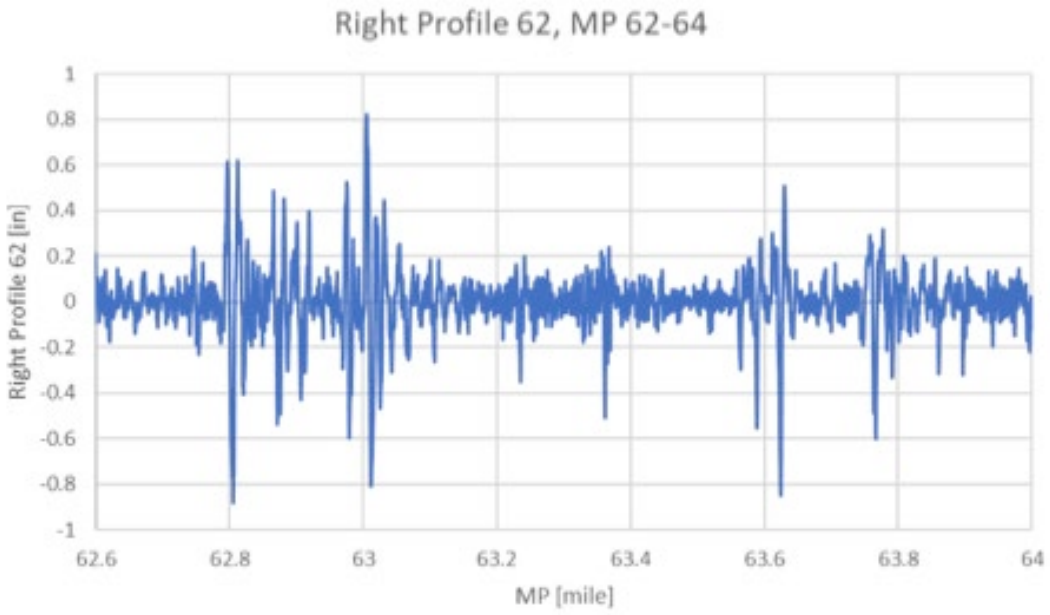


Figure 2: Amtrak continuous track geometry data from Oakington Road test site (Track Profile Right 62-foot chord)[3] MP 62.6 to 64.0 (test of December 2013)

2.2 MRail Data

The MRail data was obtained from an MRail unit mounted on the FRA DOTX218 track inspection vehicle as measured on CSX track. The CSX data taken from DOTX218 represents one continuous run made from MP 67 to 69 on the Peninsula Subdivision on April 2016. The MRail portion of the CSX DOTX218 data is presented in Table 1.

Table 1: MRail data from FRA's DOTX218 taken on CSX

MP	FEET	Speed	Track Class	Posted Speed	Track Number	YRel Left	YRel Right	Lat	Lon
Counts	Counts	Mph	Number	Mph	Number	Inches	Inches	Degrees	Degrees
67	0	49	4	79	5	0.0261	0.15276	37.46399	-77.1414
67	1	49	4	79	5	0.01618	0.14053	37.46399	-77.1414
67	2	49	4	79	5	0.01222	0.13481	37.46399	-77.1414
67	3	49	4	79	5	0.02503	0.10977	37.46399	-77.1414
67	4	49	4	79	5	0.02046	0.10093	37.46399	-77.1414
67	5	49	4	79	5	0.03794	0.12167	37.46399	-77.1414
67	6	49	4	79	5	0.05662	0.12906	37.46399	-77.1414
67	7	49	4	79	5	0.07493	0.13764	37.46399	-77.1414

[3] Right Profile 62 is the deviation from the vertical surface as measured by a 62 foot-chord on right rail

2.3 GPR Data

Consistent with the track geometry data, two sets of GPR data were obtained from FRA and Amtrak respectively.

The first represents a 2 mile stretch of track on CSX, taken by the DOTX218 car on April 5, 2016. The 2 miles are MP 67 to 69 on the CSX Peninsula Subdivision. Table 2 includes the data that was processed by ENSCO and includes specific parameters.

Table 2: GPR data taken by FRA's DOTX218 on CSX

Rail	Division	Sub-Division	Line Segment	Track ID	Collection Date	BMP	EMP	Latitude	Longitude	LRI_CAT_Left	LRI_CAT Center	LRI_CAT Right	BTI_CAT _Left
CSX	Huntington -East	Peninsula	Unknown	1	4/5/2016	67.000	67.003	37.464020	-77.14152	0	3	0	0
CSX	Huntington -East	Peninsula	Unknown	1	4/5/2016	67.003	67.006	37.464039	-77.14158	0	3	0	0
CSX	Huntington -East	Peninsula	Unknown	1	4/5/2016	67.006	67.009	37.464050	-77.14163	0	3	0	0
CSX	Huntington -East	Peninsula	Unknown	1	4/5/2016	67.009	67.012	37.464060	-77.14168	0	2	0	0
CSX	Huntington -East	Peninsula	Unknown	1	4/5/2016	67.012	67.015	37. 464073	-77.14174	0	2	0	0
CSX	Huntington -East	Peninsula	Unknown	1	4/5/2016	67.015	67.018	37. 464087	-77.14179	0	2	0	0
CSX	Huntington -East	Peninsula	Unknown	1	4/5/2016	67.018	67.021	37. 4640108	-77.14184	0	2	0	0
CSX	Huntington -East	Peninsula	Unknown	1	4/5/2016	67.021	67.024	37. 4640113	-77.14190	0	2	0	0
CSX	Huntington -East	Peninsula	Unknown	1	4/5/2016	67.024	67.027	37. 4640131	-77.14195	0	2	0	0

Table 2: GPR data taken by FRA's DOTX218 on CSX (continued)

BTI_CAT _Center	BTI_CAT _Center	FDL_Value _Left	FDL_Value _Right	FDL_CAT _Left	FDL_CAT _Center	FDL_CAT _Right	BFI_Value _Left	BFI_Value _Center	BFI_Value _Right	BFI_CAT _Left	BFI_CAT _Center	BFI_CAT _Right
5	0	10	9	-1	2	0	15	18	-1	3	3	0
5	0	11	9	-1	2	0	14	13	-1	3	3	0
5	0	11	9	-1	2	0	15	12	-1	3	3	0
5	0	10	9	-1	2	0	19	17	-1	3	3	0
4	0	10	9	-1	2	0	20	12	-1	3	3	0
3	0	10	10	-1	2	0	12	11	-1	3	3	0
4	0	10	10	-1	2	0	12	13	-1	3	3	0
4	0	10	10	-1	2	0	18	12	-1	3	3	0

2.3.1 Ballast Fouling Index (BFI)

The BFI results were calibrated to the Selig Fouling Index using data and ballast samples acquired in the U.S. in 2014. The range of values used are presented in Table 3.

Table 3: Ballast fouling indexes

BFI Category	Description	Modelled Fouling Index (Selig)
5	Clean	0 to <5
4	Moderately Clean	5 to <10
3	Moderately Fouled	10 to <25
2	Fouled	25 to <30
1	Highly Fouled	>30
0	Unavailable	n/a

Note: Values of 0 in the BFI category column and -999 in the BFI value column (see Table 2) indicate that the fouling index could not be calculated due to high EMI or the presence of a surface/sub-surface structure.

2.3.2 Layer Roughness Index (LRI)

The layer roughness index (LRI) provides a visual indication of the level of variation in the depth to the base of the primary track bed layer, designed to highlight areas where the interface with the underlying materials is highly irregular. Such areas can be indicative of sub-grade erosion, ballast pumping and wet-bed formation. The LRI is displayed as color-coded values for each channel of data, reported over three categories from good (green) to very poor (red) (see Table 4).

A “Good” rating indicates less than 2 inches of depth variance over 66 feet, 'Poor' indicates between 2 inches and 4 inches of variance, while “Very Poor” indicates greater than 4-inch variance over 66 feet. Both the wavelength and thresholds can be customized as required.

Table 4: Layer roughness indexes

Category	Description	Variance (inch)
3	Good	< 2
2	Poor	2–4
1	Very Poor	> 4
0	Unavailable	n/a

2.3.3 Ballast Thickness Index (BTI)

The ballast thickness index (BTI) provides an indication of sections of track where the thickness of the primary ballast layer falls outside of an optimum range as defined by the standard track bed design thickness. The category thresholds are relative to top of tie (Table 5).

Table 5: Ballast thickness indexes

Category	Description	Thickness (inch)
5	Positive Exceedance Level 2	> 23
4	Positive Exceedance Level 1	17–23
3	No Exceedance	11–17
2	Negative Exceedance Level 1	5–11
1	Negative Exceedance Level 2	< 5
0	Unavailable	n/a

2.3.4 Free Draining Layer (FDL) Depth Index

The free draining layer (FDL) interface represents the boundary between relatively clean ballast and highly fouled ballast and is determined by applying a BFI threshold to the 2D ballast fouling map that is used to determine the one-dimensional BFI.

The FDL depth index is designed to help highlight areas were the depth to the top of the ballast fouling is above or at the level of the base of the tie. In such instances, the track bed drainage may be significantly compromised. The FDL is reported relative to top of tie (see Table 6).

Table 6: Free draining layer depth indexes

Category	Description	Thickness (inch)
3	Good	>12
2	Poor	6-12
1	Very Poor	<6
0	Unavailable	n/a

The second set of GPR data was from Amtrak's Oakington Road site, MP 62.57 to MP 64.5 Track 2. The data provided is in a graphical format as shown in Figure 3. The data included a BFI, calculated fouling condition and relative moisture information, BLT, etc.

Figure 3 presents the GPR output for MP 62+3100 (62.6) to 63+3000 (63.58)[4] which is matched to the track geometry data section. Note, the GPR data encompasses three sections corresponding to the right, center and left portions of the track,[5] and includes multiple sets of information:

- Light Detection and Ranging (LIDAR) view (top level in Figure 3)
- Relative moisture for left, center and right of track 2 (2nd, 3rd, and 4th levels)
- Top and bottom of ballast layer depth for left, center and right of track 2 (2nd, 3rd, and 4th levels)

[4] A shorter GPR display is presented for clarity of viewing, the actual analysis used the full matching length MP 62.6 to 64.0.

[5] The GPR system used three measurement antennae with one between the rails (center antenna) and the other two on the field side of the left and right rail respectively.

- BFI (5th level) for
  - Left
  - Center
  - Right
- Running roughness of profile (62-foot chord)

This data was digitized manually, based on 16.7 foot intervals. Note, the BFI digitization used the color-BFI relationship shown in Table 7.

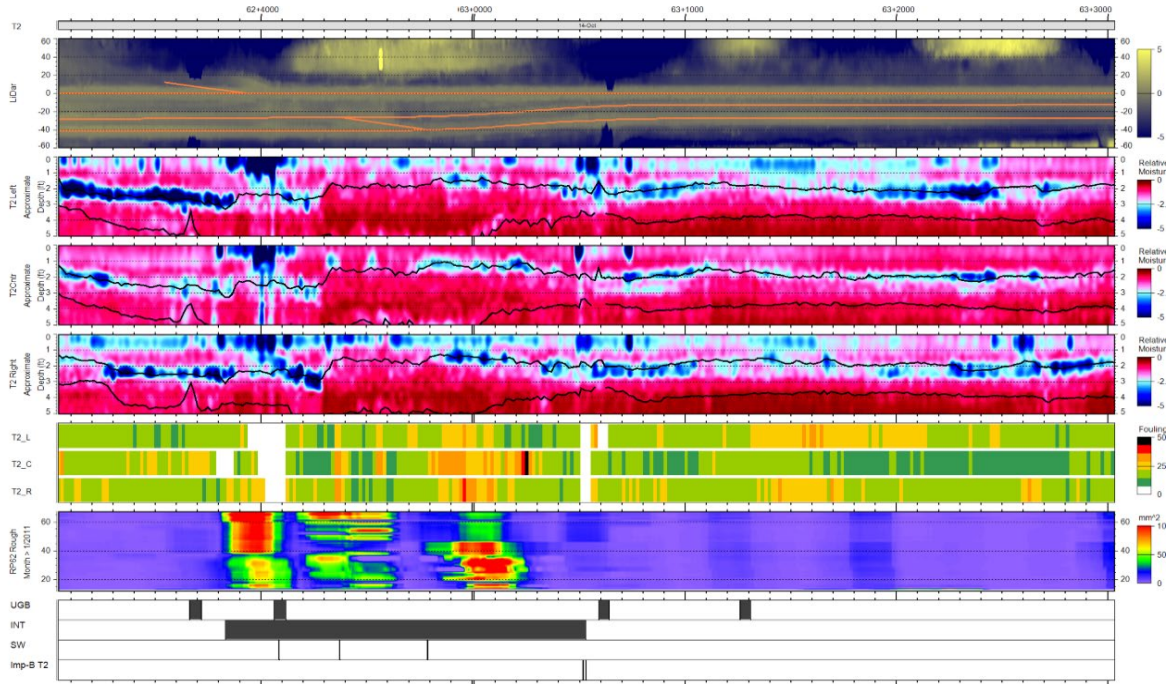


Figure 3: Amtrak Oakington Road GPR image data

Table 7: BFI conversion table

Color	Condition	Index	BFI Value
	Clean	1	3.6
Green		2	10.7
		3	17.9
Yellow	Moderately Fouled	4	25.0
		5	32.0
Red		6	39.3
Black	Highly Fouled	7	46.4

2.4 Consolidation of Data into the Common Database

The next step in the analysis process is the consolidation and pre-processing of the data into a common database. The consolidation and pre-processing of the data was done using a combination of Microsoft Excel and R software.[6]

2.4.1 Management of Data

The research team received more than 100 files in various file formats to include pdf, txt, Excel, and csv from FRA and Class I railroads. The data file distribution, by data type, is summarized in Table 8 below.

Table 8: Received data file distribution

Type of data	Number of files received
Geometry	71
Exceptions	9
TQI	13
DEF	18
MRail	9
MGT	7
GPR	2
total	129

The range and variability of the data required extensive data processing, correlation and follow up data analysis. The type of data issues addressed included data variability, data dispersion, data diversity, and data interdependence caused by such factors as:

- Variable data source (to include raw and processed data) from the different Class I railroads, FRA and its data analysis contractors.
- Type of the data (see Table 8)
- Differing inspection dates
- Differing locations to include division, sub-division, track and exact location (MP)
- Data reference and alignment errors
- Calibration and data drift
- Missing data
- Other data errors associated with large volumes of data.

[6] Programming language utilized for statistical computing and graphics. R Core Team (2018). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. The R Project for Statistical Computing

Based on the above, an essential task in the project was to organize and manage the data file. The approach taken here started with creating a traceability table (see Table 9) by file type, location, version, etc. This was extremely helpful in identifying any data organization problems.

Table 9: Data traceability example table

9	IM	POC IM	Type of meas./data	Measur. Comp.	POC Measur. Comp.	Division	Sub-Division	Line Name	MP from	MP to	File Name	Data Format	Date
1	BNSF		MRail				Cherokee		308.16	271.14	CSV__FT_BY_FT__MID8__UNK_CHEROKEE_0__MP_309_to_268.csv	CSV	20160915
2	BNSF		MRail				Cherokee		268	309	CSV__FT_BY_FT__MID7__UNK_CHEROKEE_0__MP_268_to_309.csv	CSV	20160915
3	BNSF		TQI			Southwest	Clovis		655000	897000	Clovis_TQI_2009to2015	CSV	2009-2015

2.4.2 Preprocessing and Development of Database

Consolidation of data is a key foundation step, and is critical in the tasks where analysis was performed on the consolidated database. Due to the disparate nature of the data, this step was time consuming. Incorrect data handling at this stage will lead to potential errors during analysis since all the data analysis are highly dependent on the quality of the data and the database.

In the data base construction, MP locations were used as the reference value for the combination and referencing of different data sets. However, the analysis goes beyond the simple reference locations e.g., such as for relationship development between variables. Since each database has its own "purpose," it was necessary to start with defining the research problem, so that the database can be purpose built for the analysis.

The steps used for the data base construction and evaluation included:

a. Creation, or converting of files to .csv[7] type files

b. Extraction, splitting, combination and grouping of the data frames[8] for a specific problem analysis of the data.

c. Aggregating data frames by variables and observation manipulation, e.g., consolidation, cleaning, filtering, aggregating and processing. For example: data base per division, defect level tags, date of collection, defect type, magnitude, track class, etc.

d. Organizing the data frame after simple plotting to find errors, missing data, or irregularities in the data

[7] Comma separated values

[8] Data frame – a two-dimensional array structure, in which each column contains measurements of single variable, and each row contains one observation. Observation may be a numeric, or character type data.

e. Alignment of the data/observations as applicable. For some of the larger data files, this step was done in the analysis stage of this project.

2.4.3 Development of Amtrak Oakington Road Data

The approach used to provide a consolidated database for the analysis of the Amtrak Oakington Road data required the consolidation of 31 geometry measurements in a single data folder. This further required conversion of data to a common .csv format and renaming the files to a common (and easily recognizable) programming identifying names. All 31 files were then loaded into the R software, and the key variables and observations defined to allow for effective data manipulation. Each geometry file contained 1 mile of foot by foot measurement data for the data variables, which are presented in Table 10.

Table 10: Amtrak Oakington Road geometry data

Mile	Feet	Track	Gage	Cross Level	Cross Level Rate	R Profile	R Prof 62	R Prof 124	L Profile	L Prof 62	L Prof 124	R Align	R Align 62	R Align 124
64	0	2	56.617	0.258	-0.048	-0.023	0.055	0.276	-0.051	0.025	0.226	0.023	-0.017	-0.004
64	1	2	56.618	0.255	-0.049	-0.026	0.052	0.282	-0.050	0.025	0.242	0.021	-0.020	-0.008
64	2	2	56.617	0.253	-0.049	-0.024	0.051	0.288	-0.046	0.026	0.254	0.016	-0.026	0.016

Table 11: Amtrak Oakington Road geometry data (continued)

L Align	L Align 62	L Align 124	Curvature	CTS	Speed	ALD	Class	Warp62	L Prof SC	R Prof SC	L Align SC	R Align SC	Sync Cnt	Sync Ft
0.027	-0.031	0.002	0.020	0.00	112	0	7	-0.159	0.328	0.385	0.005	0.004	189	1002
0.024	-0.028	-0.001	0.020	0.00	112	10	7	-0.161	0.333	0.391	0.003	0.002	189	1003
0.018	-0.028	-0.008	0.020	0.00	112	10	7	-0.166	0.337	0.397	0.00	-0.0008	189	1004

After identifying the variables, the research team defined an initial statement problem for the analysis of this data, which is that of finding a relationship between the GPR data and the track degradation as a function of geometry measurement type and chord length (for Amtrak 31, 64 and 124 foot chords). Table 12 illustrate a data frame constructed from the geometry data at Oakington Road, designed for this specific problem and analysis approach. The database was created using code developed in the R software. A data frame was created for each variable, e.g., Right Profile. Table 12 illustrates the data frame for right profile observations per each measurement date at MP 64.

Table 12: Amtrak Oakington Road geometry data frame per inspection date

Date	31 ft. chord	64 ft. chord	124 ft. chord
06_2013	-0.003	0.018	0.022

Date	31 ft. chord	64 ft. chord	124 ft. chord
07_2013	-0.125	-0.108	-0.077
08_2013	-0.111	-0.087	-0.078
09_2013	-0.117	-0.118	-0.072
10_2013	-0.117	-0.091	-0.078
12_2013	-0.116	-0.12	-0.09
01_2014	0.074	0.075	0.067
03_2014	0.064	0.049	0.034
04_2014	0.049	0.041	0.028
06_2014	0.066	0.041	0.028
07_2014	0.056	0.051	0.029
10_2014	0.001	0.025	0.017
11_2014	0.006	-0.009	0.008
12_2014	-0.015	-0.001	0.016
01_2015	-0.001	-0.004	0.024
02_2015	-0.002	-0.002	
03_2015	-0.008	-0.007	
04_2015	-0.025	0.016	
05_2015	-0.019	-0.008	
06_2015	0.001	-0.006	
07_2015	-0.03	-0.02	
08_2015	-0.029	-0.007	
11_2015	0.016	0.016	
12_2015	-0.081	-0.089	
01_2016	-0.018	-0.023	
02_2016	-0.022	-0.023	
03_2016	-0.021	-0.026	
04_2016	-0.018	-0.018	
05_2016	-0.005	-0.033	
06_2016	-0.028	-0.05	
09_2016	0.055	0.052	

The next stage for this data frame was data plotting and summary of the variables to find any error, or irregularity in the data. This was performed on all 16 variables[9] resulting in the creation of 16 data frames with one variable each. Additional data frames for combinations of variables were also be created for use in the analysis of the relationships.

[9] Gage, Cross Level, Right Profile 62 ft., Right Profile 124 ft., Left Profile 62 ft., Left Profile 124 ft., Right Alignment 124 ft., Left Alignment 62 ft., Left Alignment 124 ft., Curvature, Warp 62 ft.

Analysis of the continuous track geometry data frames showed what appeared to be a misalignment between different geometry runs, a not-uncommon occurrence. As a result, the data frames were carefully and individually examined and then re-alignment of the data was performed, as illustrated in Figure 4 and Figure 5. By examining and comparison of the identical inspection variables, e.g., Right Profile in different inspection dates, it was possible to identify with high probability which inspection data needed alignment and the required amount of alignment. Figure 4 and Figure 5 present an example of four sets of geometry signals before and after alignment (each color represents a different date of inspection).

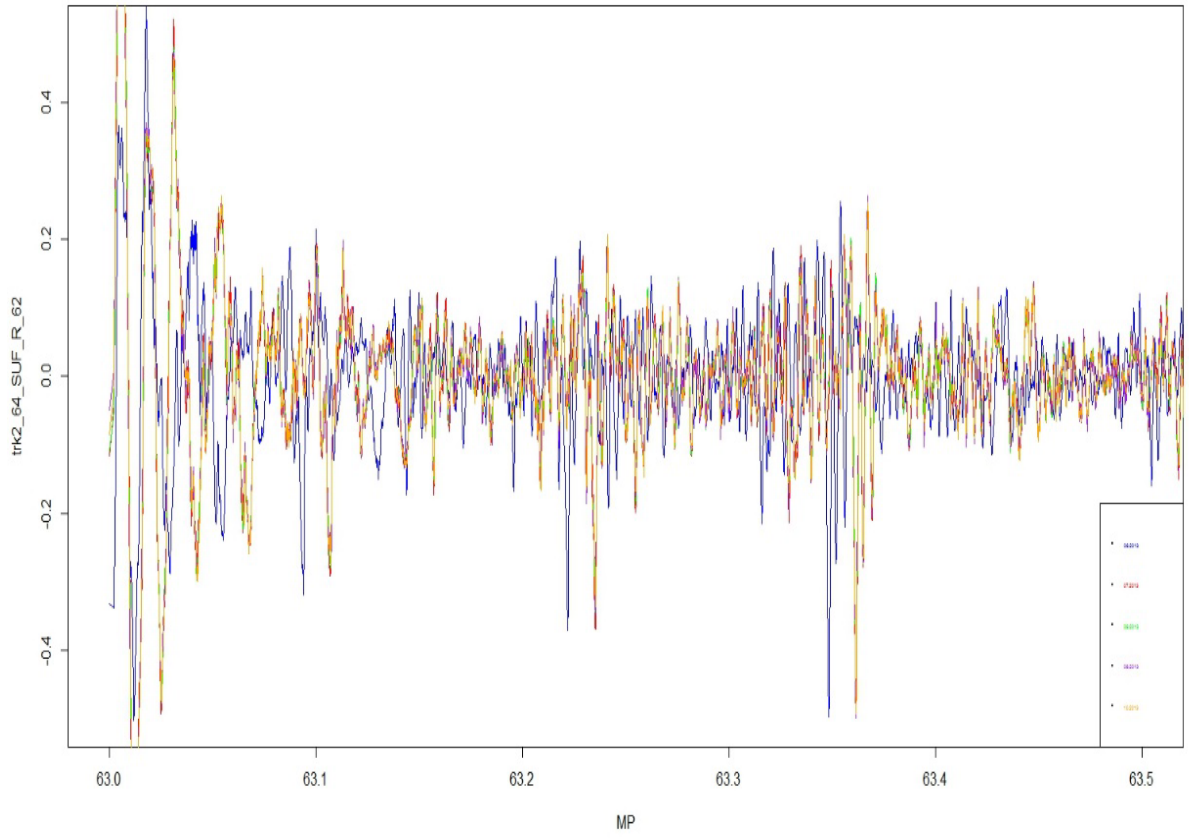


Figure 4: Data alignment example, before alignment[10]

[10] Note that variation is difficult to see between runs due to longitudinal dis-alignment

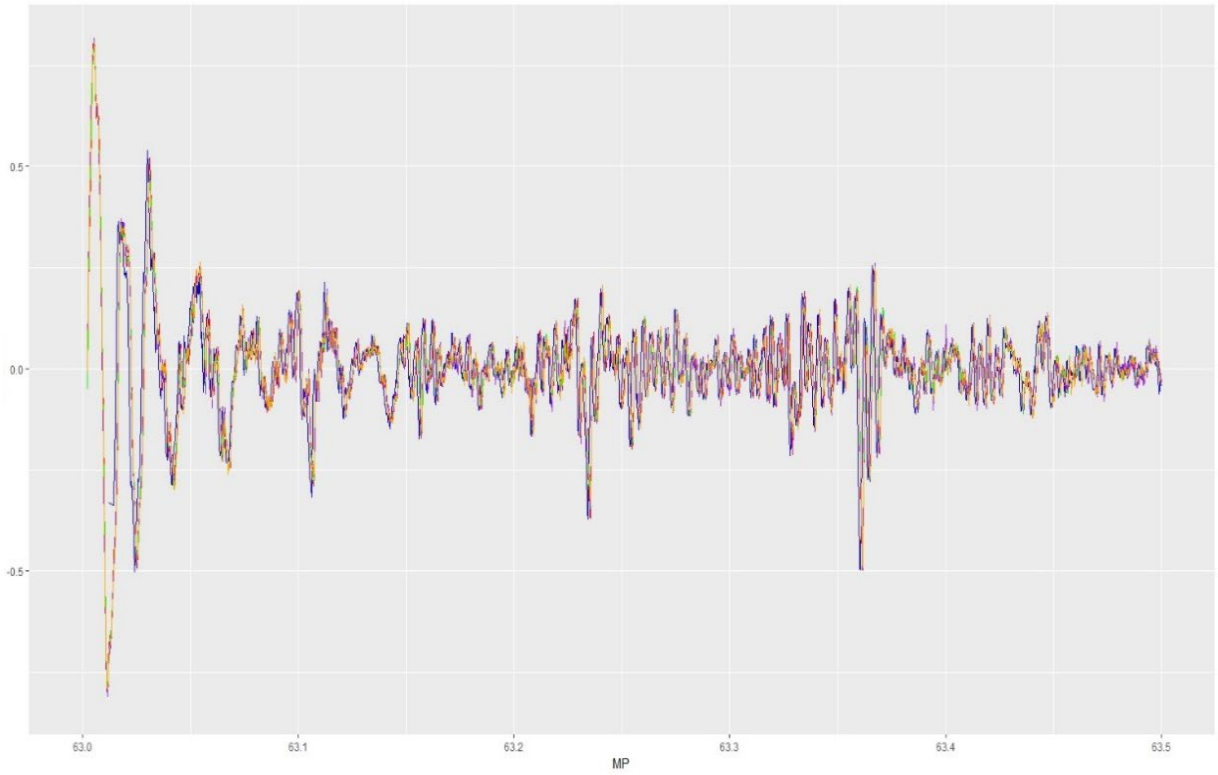


Figure 5: Data alignment example, after alignment

Using the location and inspection date as references, the data set was consolidated into a unified database. As noted, database preparation included matching GPR and geometry inspection measuring points, and creation of a mutual data frame of reference, considering different data sampling rates. After alignment of the inspection data by shifting the signal to match the peaks, the signals were consolidated into a common reference MP, noting that each inspection has different sampling steps and corresponding different number of measurements in the approximately 2 miles of data.

3. Data Preparation and Exploratory Data Analysis (EDA)

Once the data was collected into a common database, the following data preparation, data mining, and EDA steps were performed:

1) Data cleaning – removing noise and inconsistent data.
2) Data integration – grouping/combining multiple data source variables.
3) Data selection – creation of relevant datasets from the existing database for sake of specific analyses.
4) Data description and transformation – analysis of data "spread" and overall description of data set using descriptive statistics and other methods. This included transformation of data to a Track Quality Index (TQI) for preliminary analysis.
5) Data mining – process where intelligent methods are applied to find and extract patterns from data.
6) Pattern evaluation – process to identify patterns based on domain knowledge and origin of the data. Develop preliminary relationships between track subsurface data and track geometry. Identify potential correlation relationships between geometry defects and GPR data.
7) Data visualization – a technique to graphically represent valuable "knowledge" regarding the data.

EDA is an approach that allows a first insight into data by means of a variety of analysis techniques, many of them graphical [4]. EDA helps characterize the data whether there are anomalies in the variables (outliers), or if there are simple relationships within the variables, patterns etc. In this activity, EDA was used to explore each inspection/dataset separately and via multivariable analysis; using GPR data such as BFI and BLT. The objective was to identify potential simple correlation relationships between multiple datasets: geometry, GPR (BFI, BLT etc.). The approach included the following techniques:

- Measures of spread and overall description of a set of data
- Descriptive statistics
- Data class and structure
- Sample of the data head/tail, the beginning/end of an organic data frame
- Direct and principles of analytic using visualization technics
- Box and whisker plot
- Histograms
- Quantile-quantile plots
- Visualization and others

Appendix A provides a more comprehensive set of these EDA plots.

3.1 Data Analytics Using Visualization Techniques

Visualization and graphical analysis play a significant role in finding relationships and patterns in the data. In the EDA, several basic visualization approaches were used to examine if there is a direct relationship between the variables of the different inspection datasets. Visualization may point to relationships and patterns, as well as providing guidance towards which advanced analyses techniques are most promising.

In this section, several visualization techniques that were in this EDA activity are presented.

3.1.1 Base and Exploratory Graphics

The base plotting system in the R software provides many important tools for data visualization and representation. The following application of the base plotting system was performed on the CSX Peninsula Subdivision data MP 67 to 69.

Bivariable Visualization

Bivariable visualization is the simplest method of finding patterns between two variables.

Figure 6 below shows the relationship of two interesting variables of the CSX Peninsula Subdivision data MP 67 to 69; YRel Right (from MRail) and Right Profile 31 (track geometry car), and the possible correlation between them. A predominant decreasing linear relationship band appears to exist suggesting that there may be a relationship.

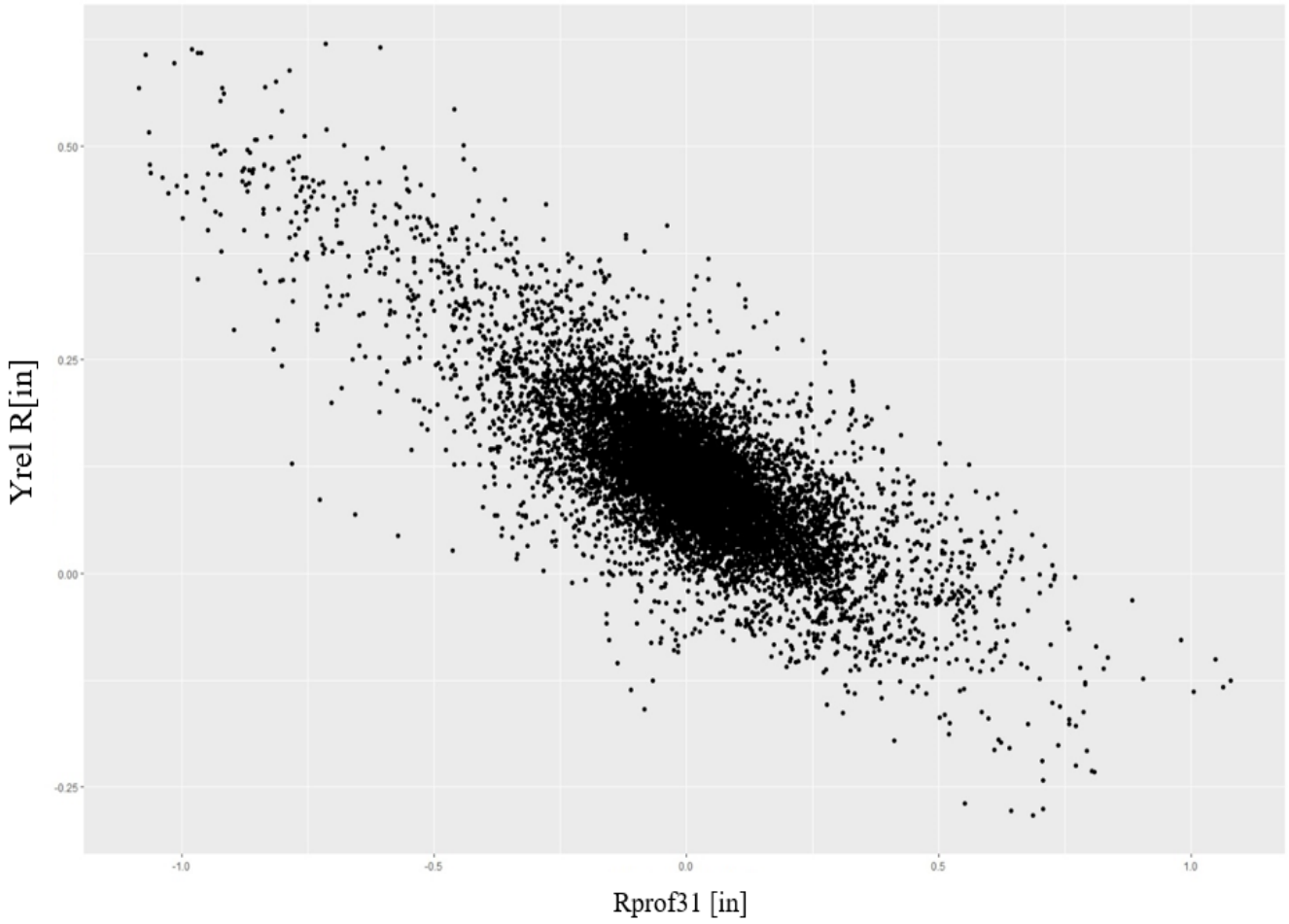


Figure 6: CSX Peninsula Subdivision MP 67–69; YRel Right and Right Profile 31 correlation plot multivariable visualization

Figure 7A and Figure 7B illustrate the graphical relationship between several of the variables of the data in this CSX dataset. These figures show a scatter plot of each variable against one another. The diagonal shows the variable of interest. The plots to the left and right of the variable have that variable as the ordinate, and the plots above and below the variable are abscissa. A distinguishable pattern in the plot shows a correlation between the two variables.

Each figure contains a combination of variables, for left and right rail respectively, that can be used to evaluate patterns, and identify changes and behaviors. Note that the figures are quite dense and difficult to define in detail; however, global relationships can be inferred.

The first variable in each figure is MP. The rest of the variables are unique, however in some cases, variables are intentionally repeated, e.g., to find patterns between the right, and left measurements. The second, third and fourth variables are Profile 31 (profile over a 31-foot chord), Profile 62 (profile over a 62-foot chord), and Profile 124 (profile over a 124-foot chord), for the left and right rails respectively. The last variable is YRail which is the MRail deflection values. The figures show a relationship between the three profile measurements, which is expected, as well as a possible relationship with YRail.

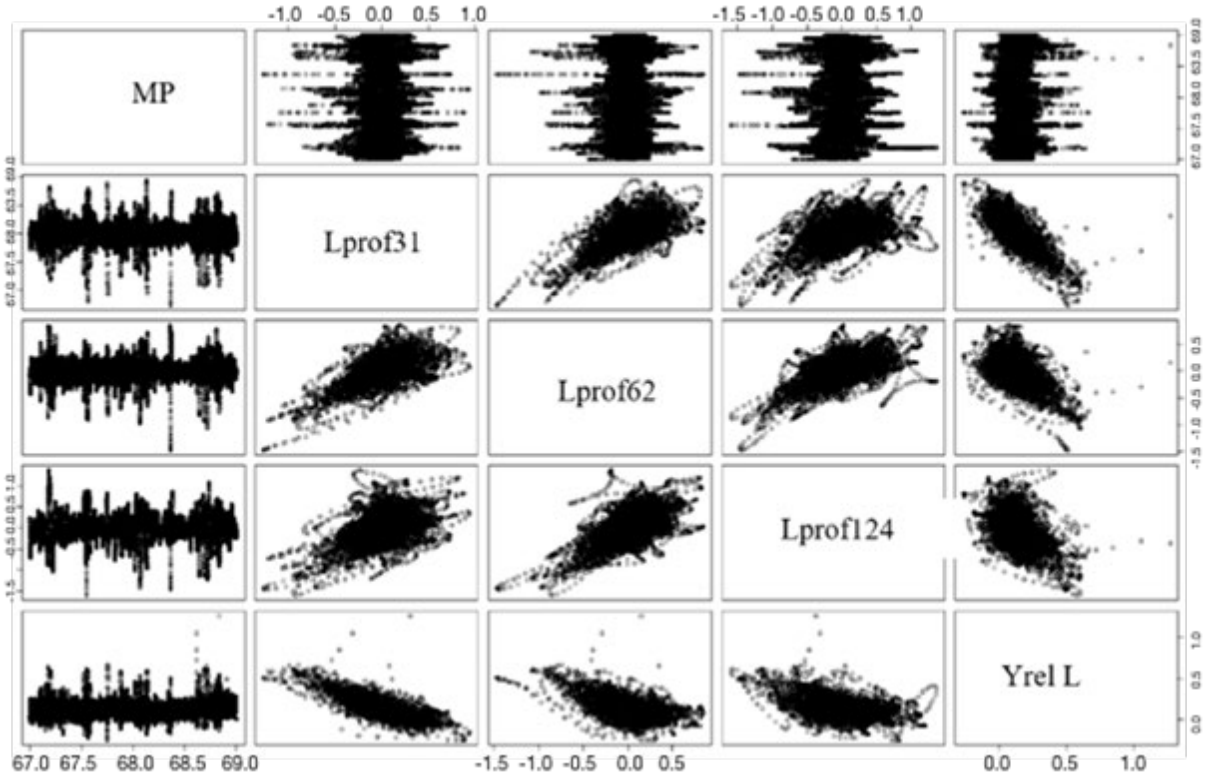


Figure 7A: CSX Peninsula Subdivision MP 67–69 multivariable plot-Left rail

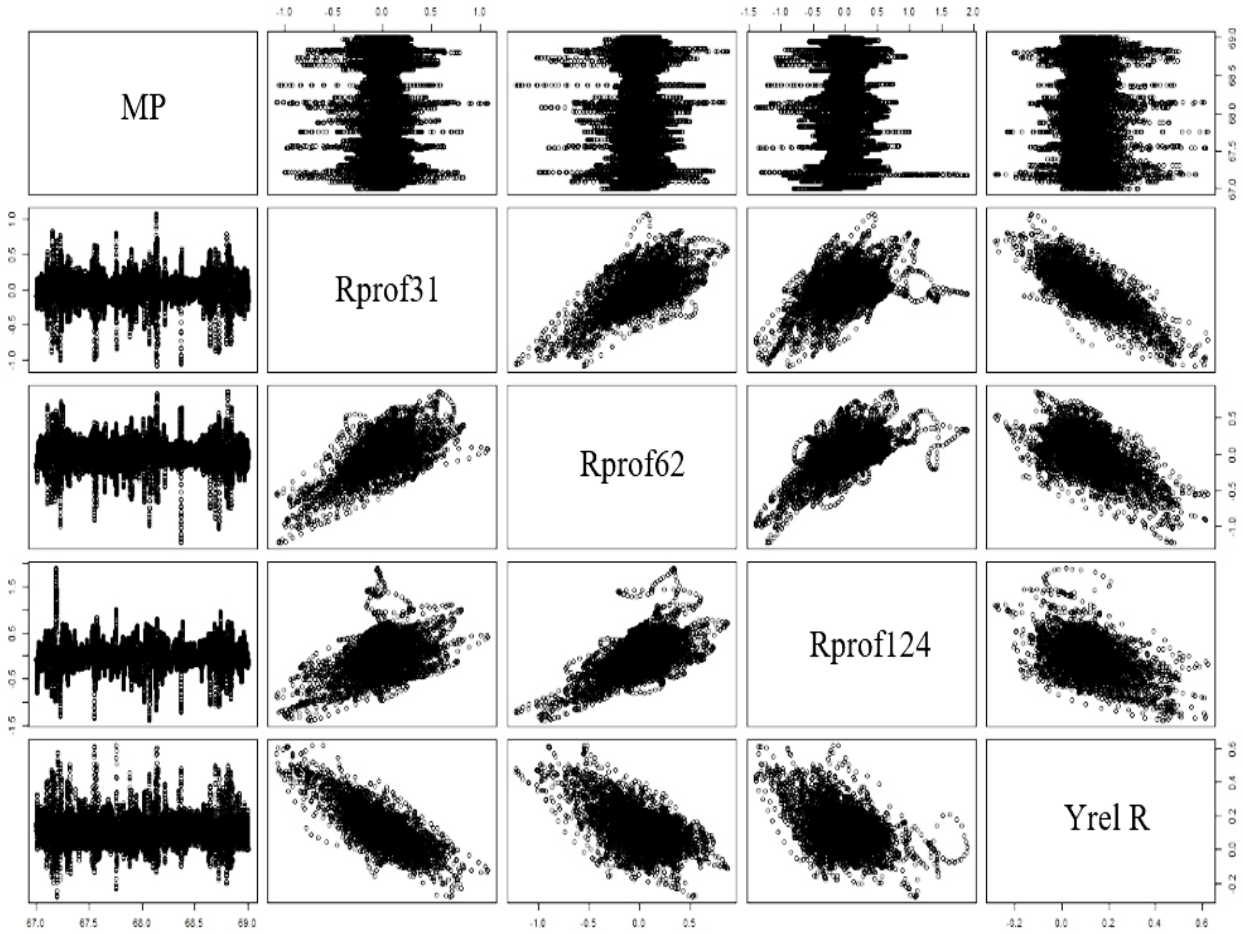


Figure 7B: CSX Peninsula Subdivision MP 67–69 multivariable plot-Right rail

To illustrate the visual ability of this approach, the geometry realignment of the different Amtrak track geometry measurements (corresponding to a time series data set) before and after alignment, as previously shown in Figure 4 and Figure 5, is presented in the same format in Figure 8 and Figure 9, which show the cross correlation for the several time series. Again, the first variable in each figure is MP. The rest of the variables represent different dates of measurement. If the data is properly aligned, the correlations should be very well behaved (a straight line with a slope of 1) to reflect the fact that this is the same section of track, just measured at different times.

However, as can be seen in Figure 8, there is a very poor correlation between the June inspection and the other dates of inspection, corresponding to the data misalignment problem shown previously in Figure 4. After alignment (Figure 9), there is a significant increase in correlation of the June and remaining inspection data, as reflected by the now well-defined linear relationship.

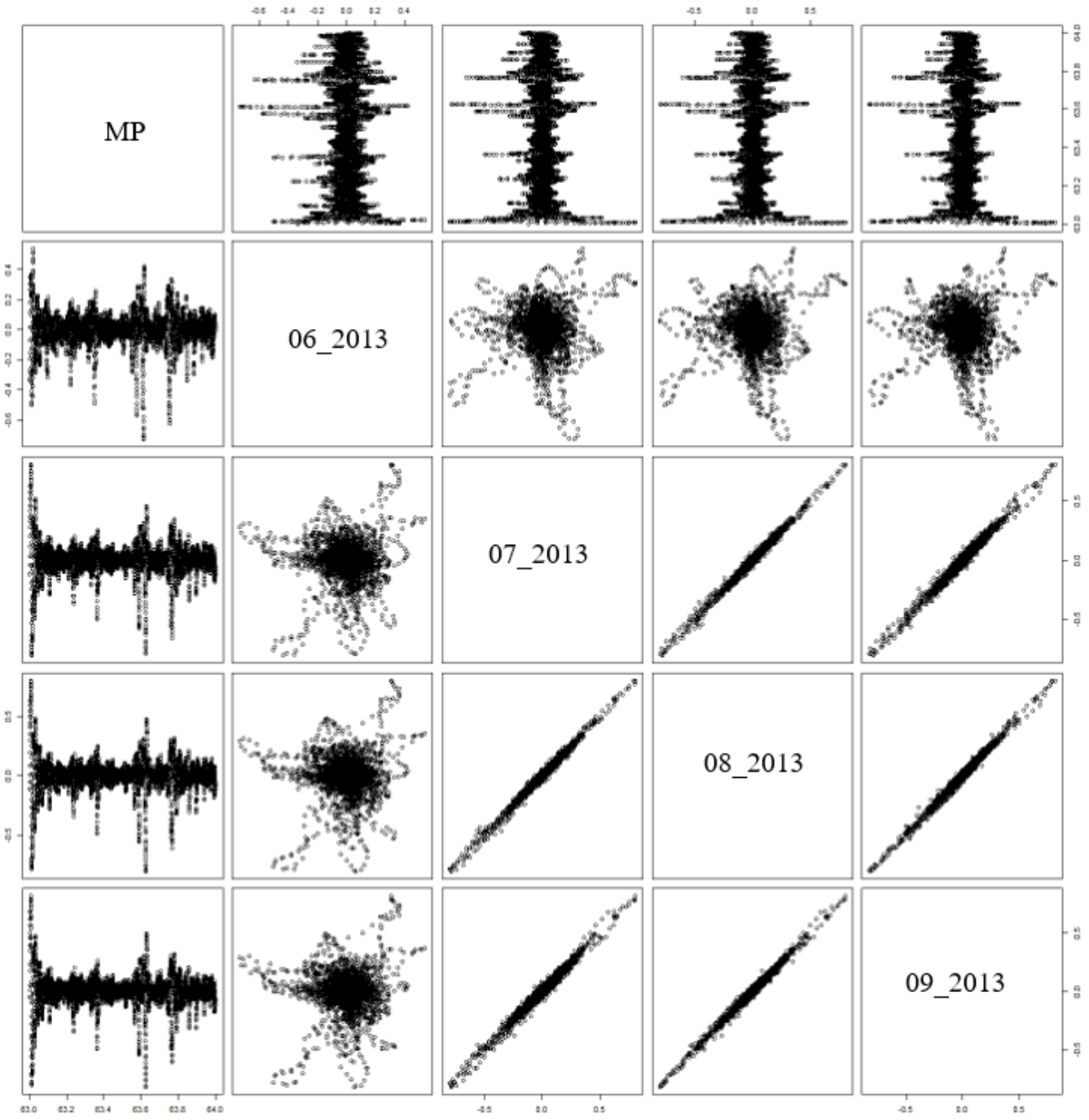


Figure 8: Graphical illustration of the time series variables correlation plot before alignment

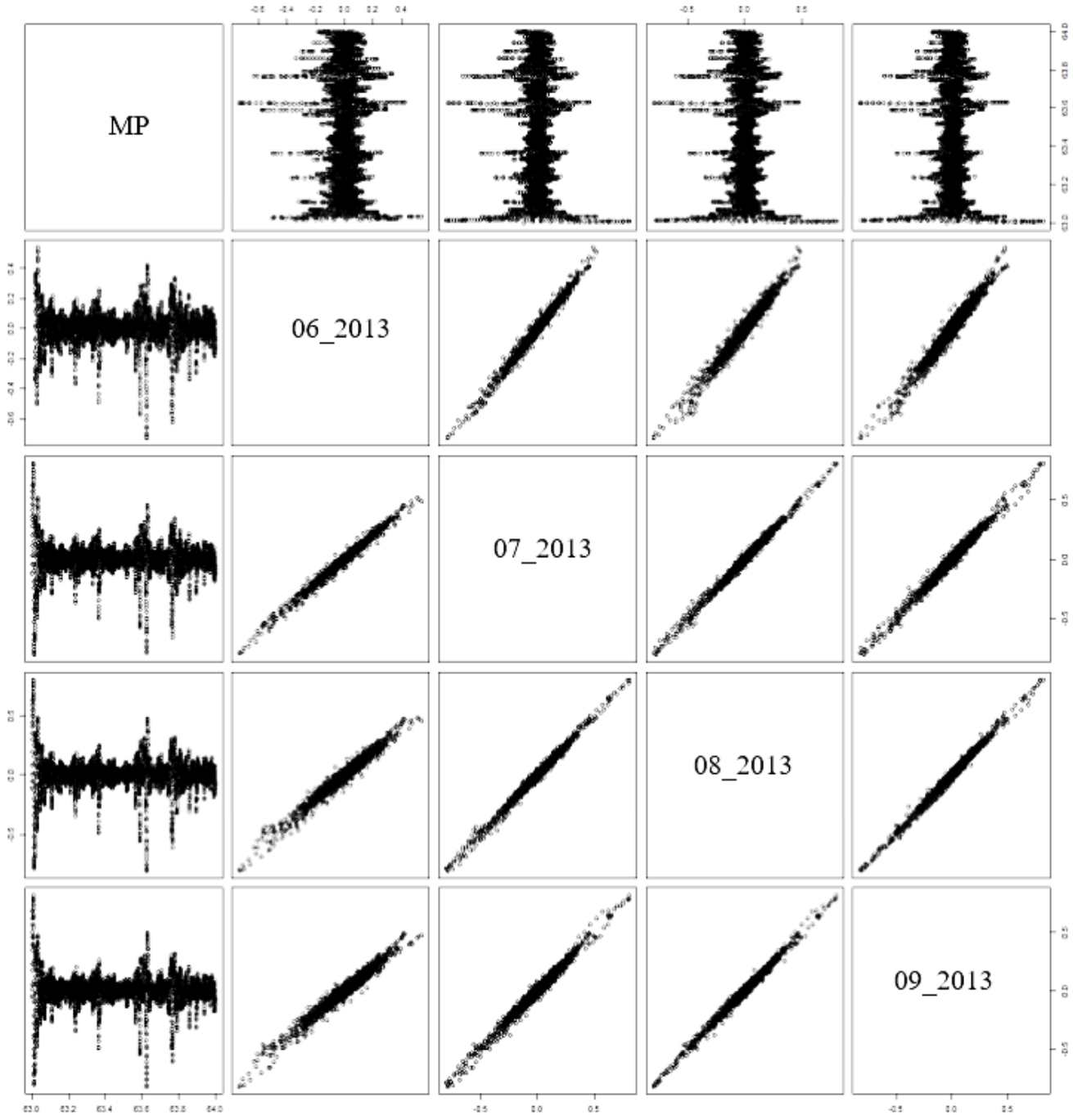


Figure 9: Graphical illustration of the time series variables correlation plot after alignment

3.1.2 3.1.2 Box and Whisker Plot

Box and whisker plots explicitly depict the shape of the data distribution, as well as its central value and variability.

As illustrated in Figure 10 below in a box and whisker plot, the median, 4 quartiles (0–25, 25–50, 50–75, and 75–100 percent), the interquartile range, and all the outliers can clearly be seen.

The outliers are observation which are from the mean and is defined as a sample that has more than 50 percent in the range of the data between it and the mean. In the case of box and whisker plots, outliers may be ignored, or deleted only in specific cases.[11]

In Figure 10 and Figure 11 below, the box plots of all Left and Right profiles of the CSX Peninsula Subdivision MP 67–69 data are presented together with the YRel data.

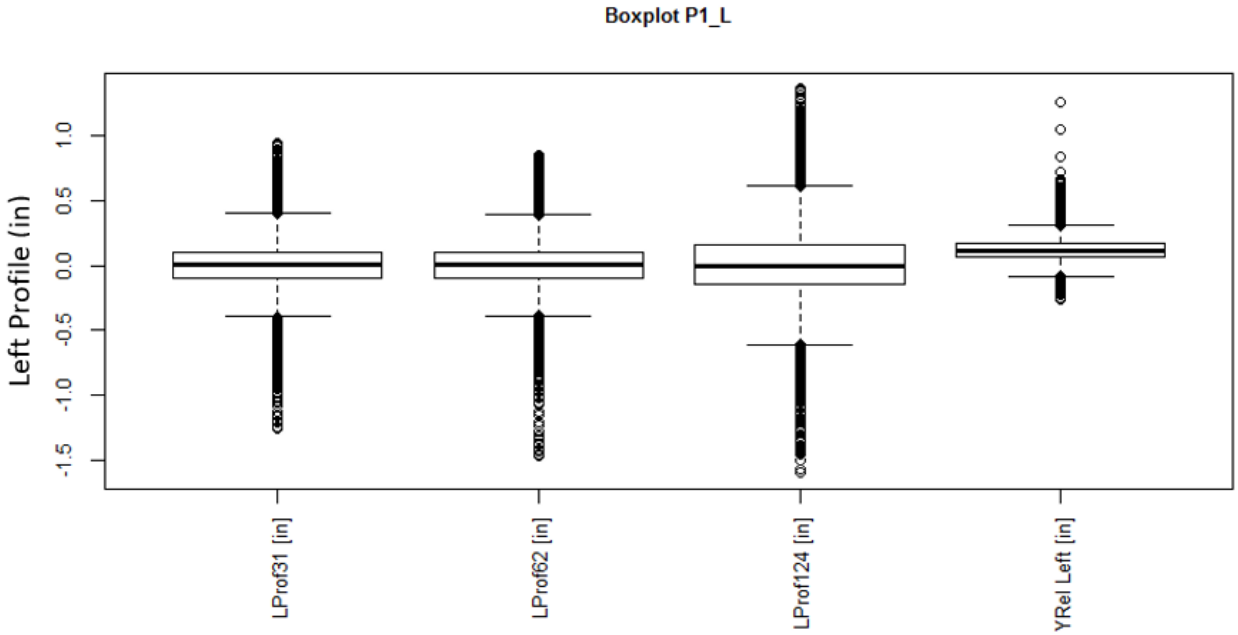


Figure 10: Box and whisker plot CSX Peninsula Subdivision MP 67–69 data plot-Left rail

[11] All outliers were investigated as to cause and eliminated if deemed appropriate.

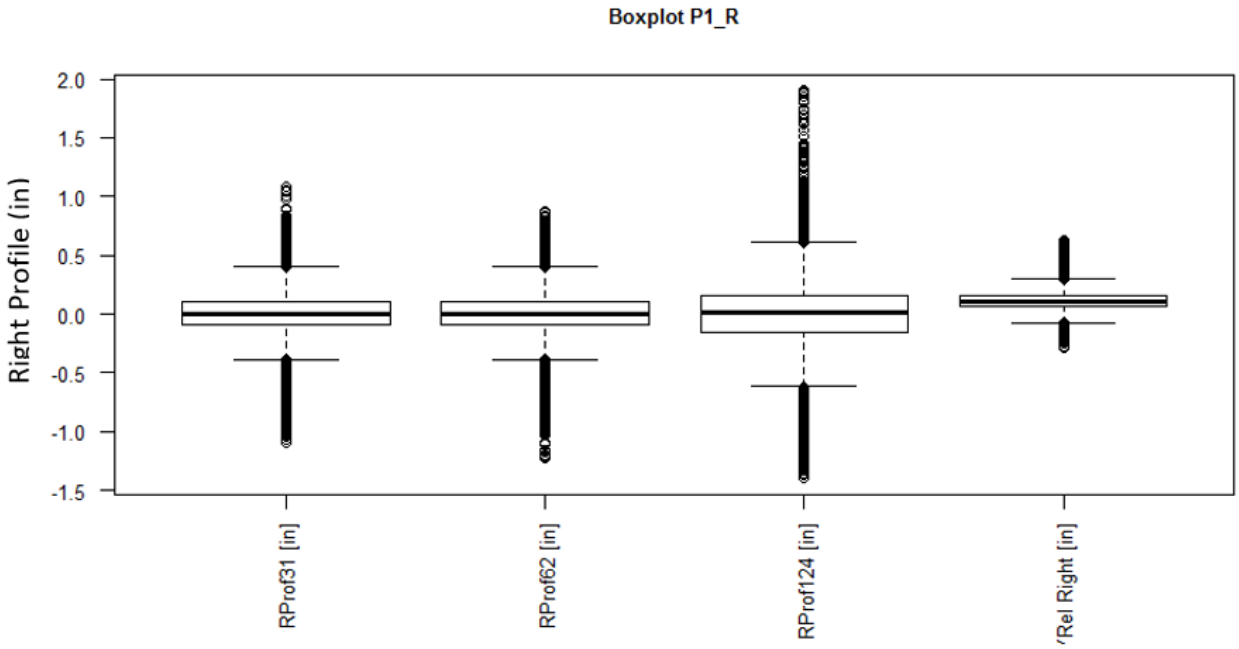


Figure 11: Box and whisker plot CSX Peninsula Subdivision MP 67–69 data plot-Right rail

3.1.3 Combination of Box and Whisker Plot and Scatter Graphical Representation

Figure 12 below presents a combination of a box and whisker plot and a bi-variable scatter plot. It shows the variability in the data along the track. This figure is a good example of the power of the box plot illustration, for the same CSX data presented previously; YRel Left and MP. This shows that the track support has a relative mean with variation of stiffness along the track, including localized soft/stiff sections.

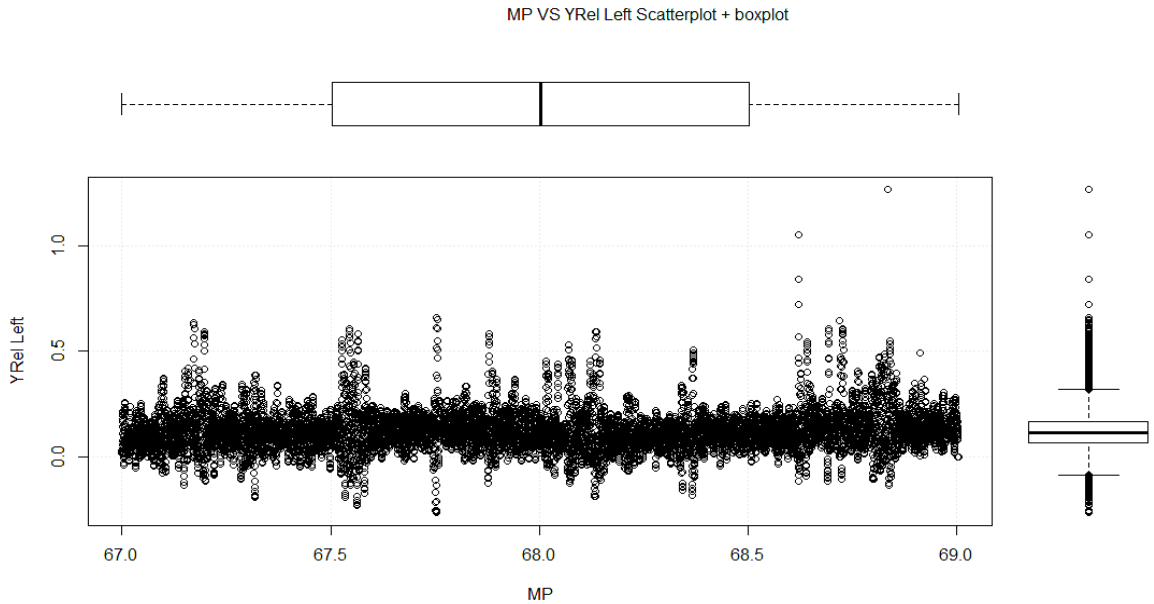


Figure 12: Box and whisker plot and Scatter CSX data; YRel Left and MP

3.1.4 Histograms

Since a histogram provides an accurate visualization of the distribution/frequency/density of a continuous numerical variable in a certain interval, it is an integral part of the EDA. The histogram separates the range of values into different numbers of sections/bins, thus showing the overall distribution of the observations/measurements.

Density vs. Frequency Histogram

Figure 13 and Figure 14 present a standard histogram in two related formats; the density histogram (Figure 13) and the frequency histogram (Figure 14). Note the two are directly related where the density (height) multiplied by the width of the column (bin size) equals the total count or frequency.

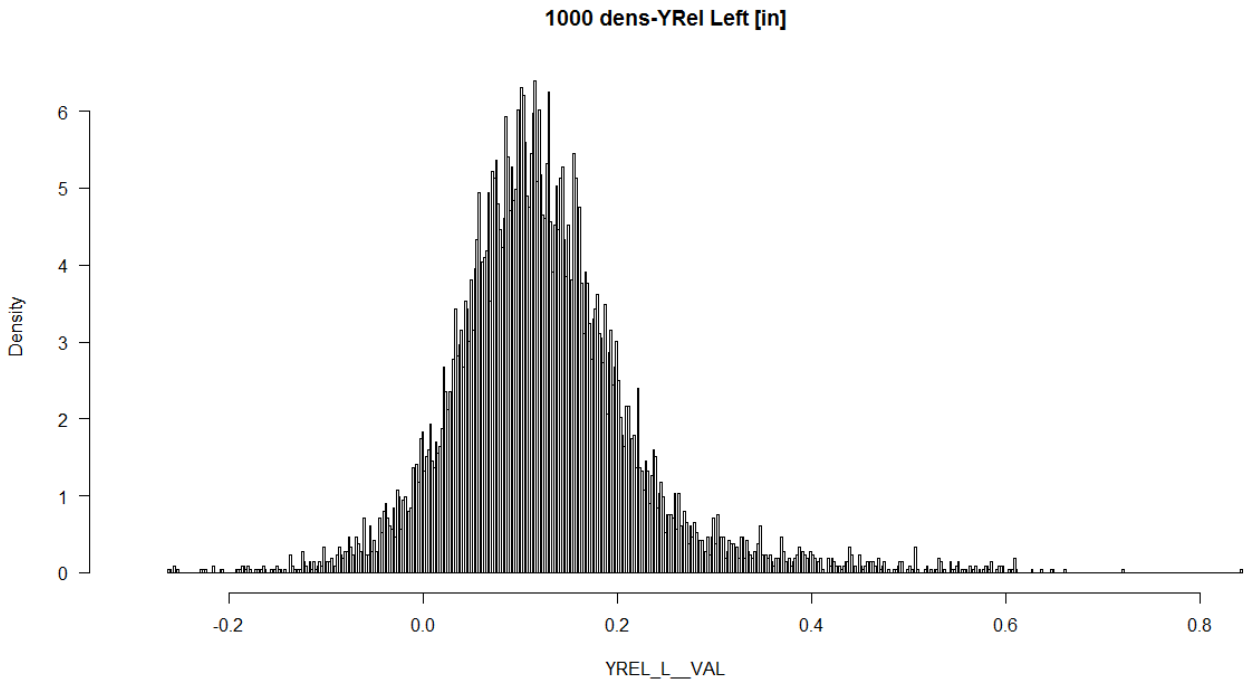


Figure 13: Density histogram, CSX Peninsula Subdivision MP 67–69 data

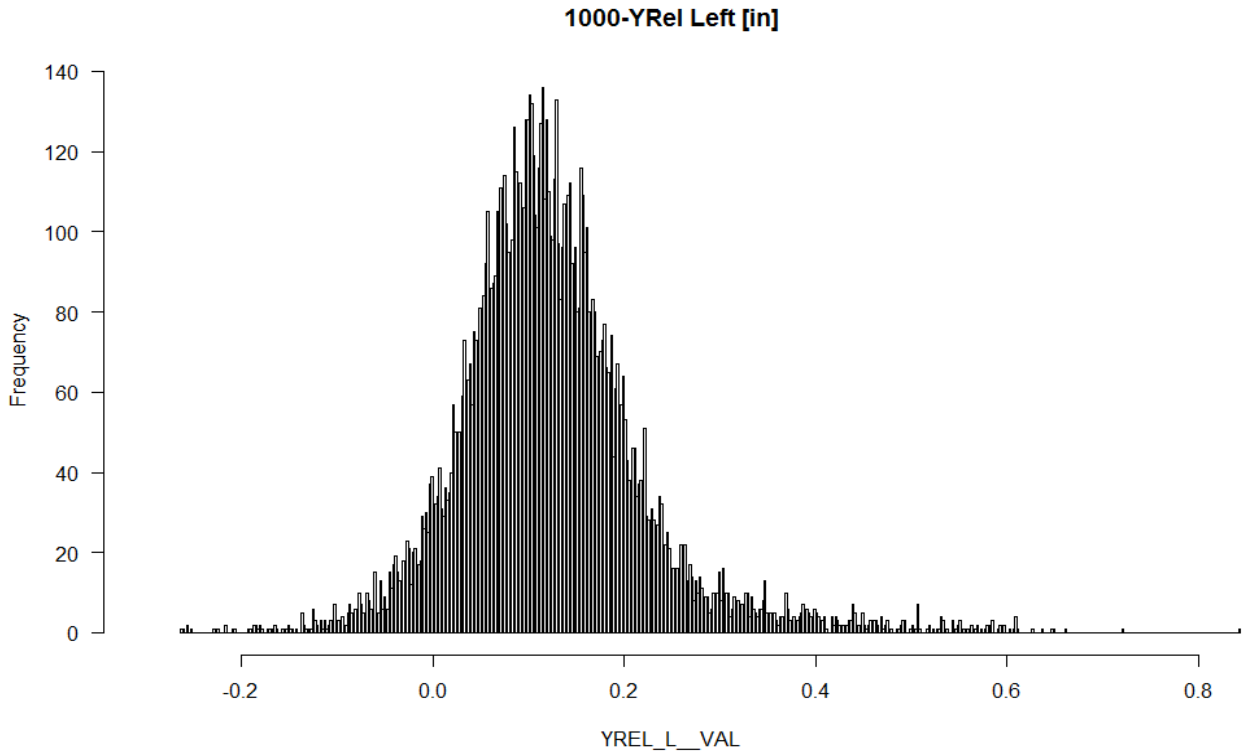


Figure 14: Frequency histogram, CSX Peninsula Subdivision MP 67–69 data

By separating the range of values into different numbers of sections/bins, the histogram can be modified to represent the distribution of the variable observations, as shown in Figure 15 and Figure 16 below. Figure 15 illustrates a 100-bin distribution, while Figure 16 is for 10,000 bins with the same number of observations. (Note, Figure 14 had 1,000 bins.)

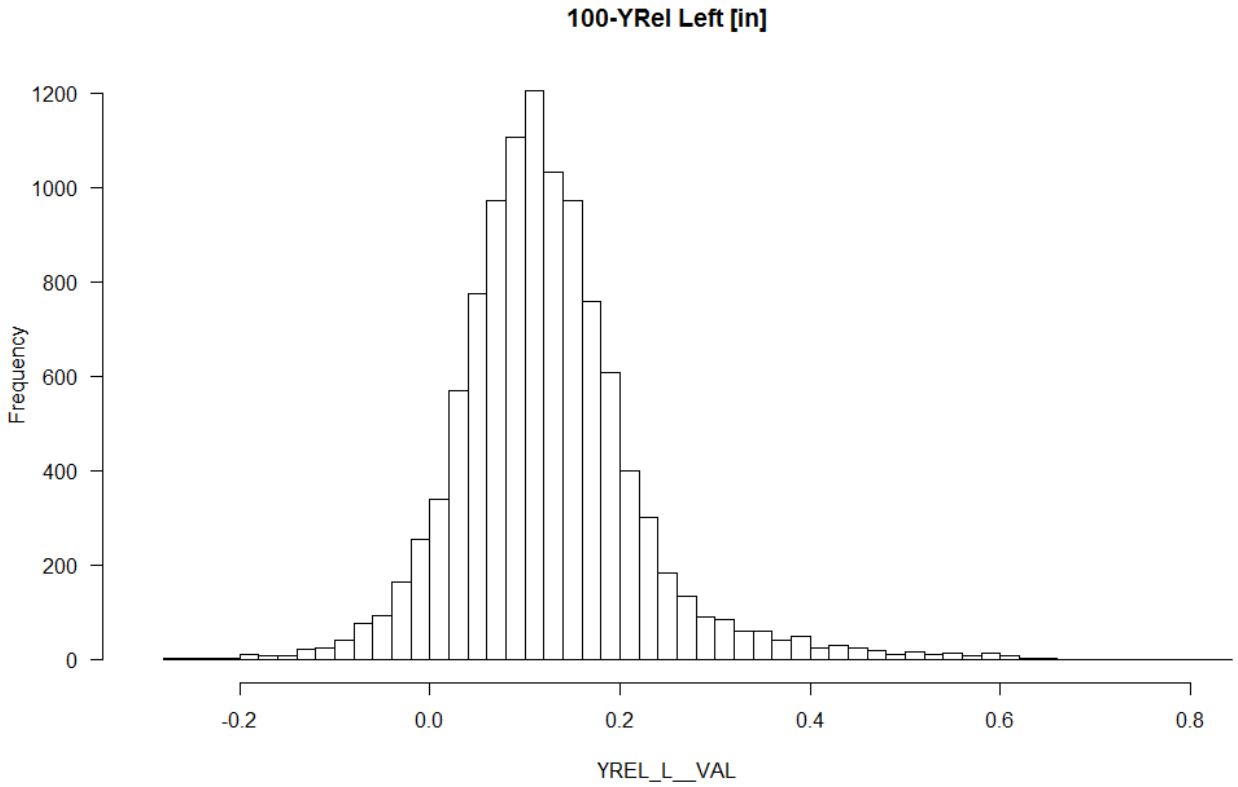


Figure 15: 100 bins histogram, CSX Peninsula Subdivision MP 67–69 data

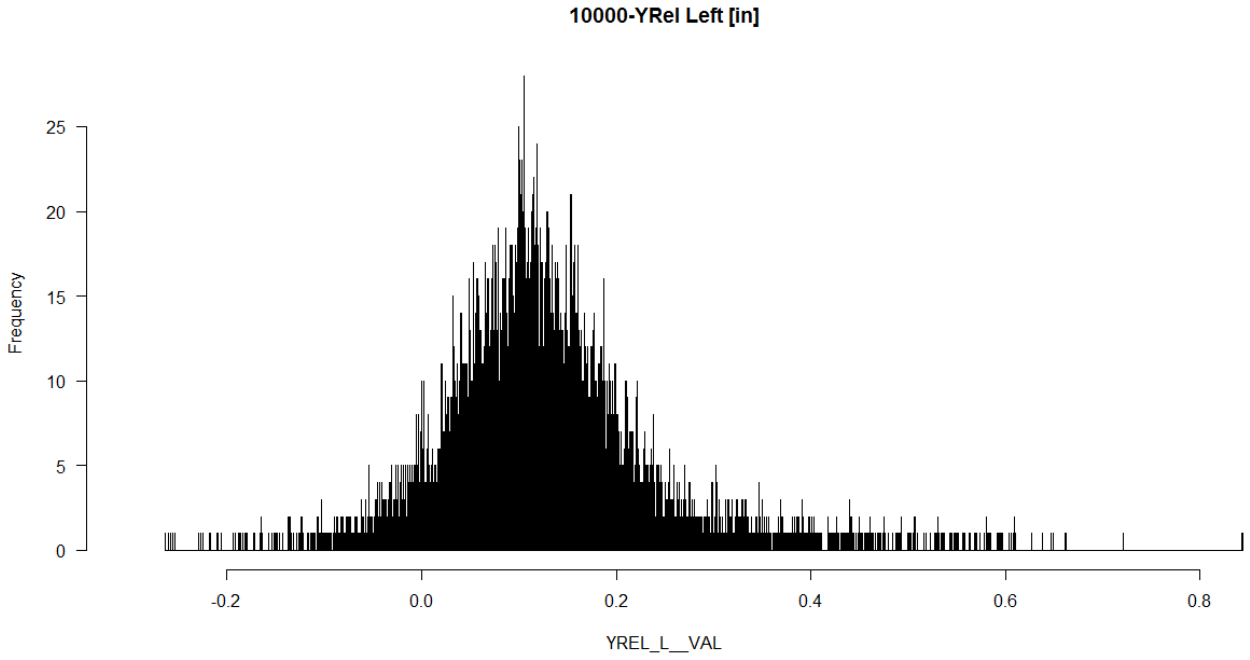


Figure 16: 10,000 bins histogram, CSX Peninsula Subdivision MP 67–69 data

Combination of a Histogram and Nonparametric Density Estimation Line[12]

A combination of histogram and Kernel Density Estimation (KDE) creates a useful illustration of optimally smoothed distribution of random variables in the histogram with high bin numbers. KDE is a non-parametric estimation of the probability density function of a given random variable.

Figure 17 through Figure 20 present the Left and Right Profile62 and YRail data for the CSX Peninsula Subdivision, MP 67–69.

[12] Vertical axis of histogram represents the number of counts per division of horizontal axis

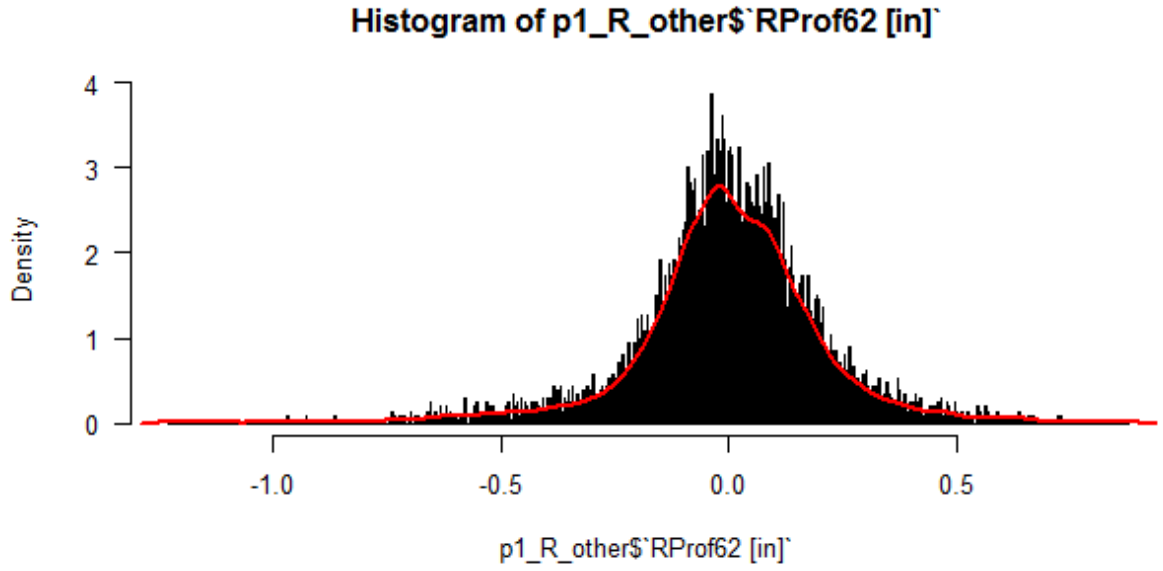


Figure 17: Histogram and KDE of CSX Peninsula Subdivision MP 67–69 inspection data, Right Profile 62

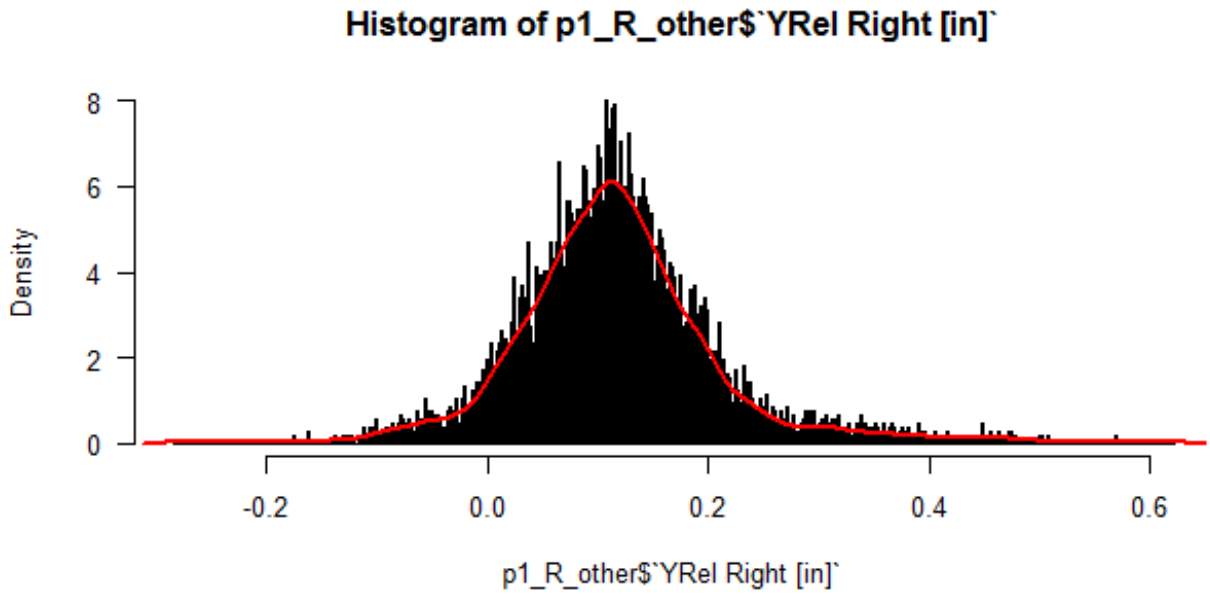


Figure 18: Histogram and KDE of CSX Peninsula Subdivision MP 67–69 inspection data, YRail-Right

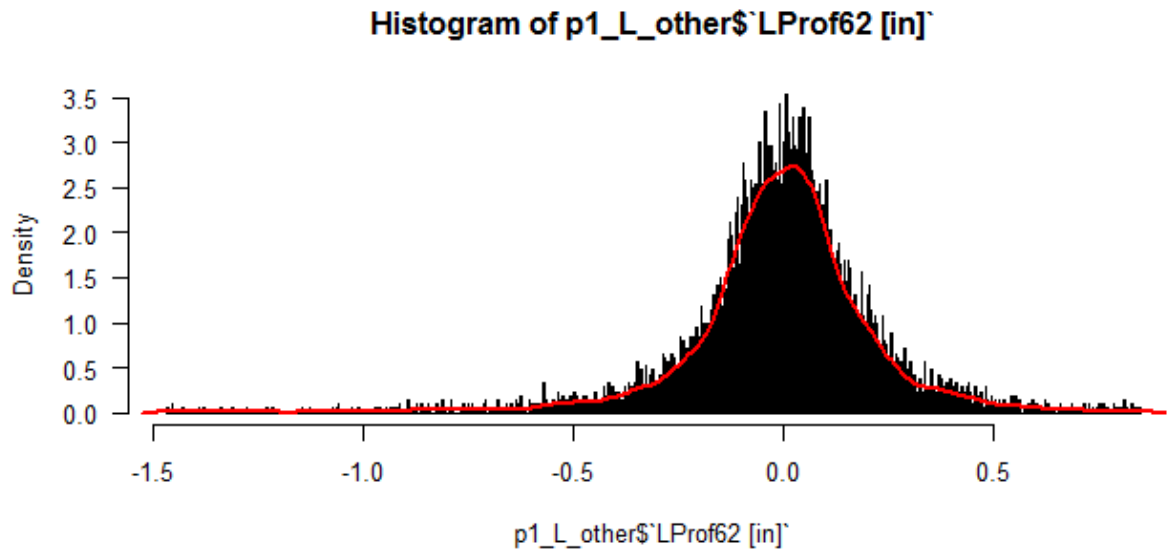


Figure 19: Histogram and KDE of CSX Peninsula Subdivision MP 67–69 inspection data, Left Profile 62

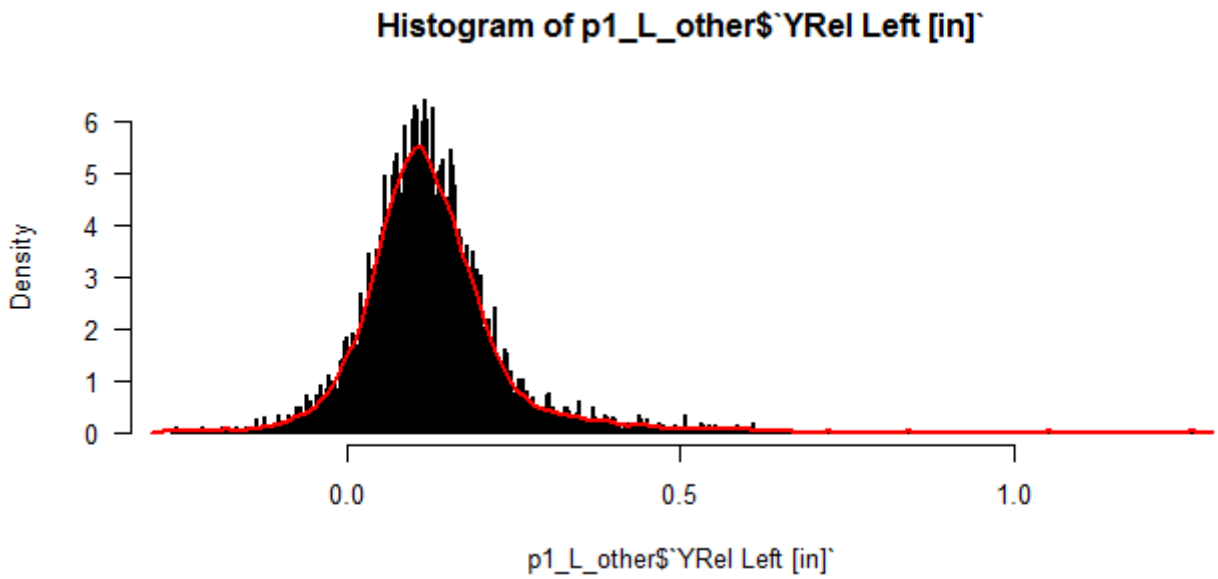


Figure 20: Histogram and KDE of CSX Peninsula Subdivision MP 67–69 inspection data, YRail Left

Quantile-Quantile (QQ) Plot

The Quantile-Quantile (QQ) probability plot is a subsidiary graphical technique for EDA, which compares two probability distributions by their quantiles against one another. The two

probability distributions may be: two variables, an ordered variable and a reference, or a variable and a theoretical distribution.

A typical QQ plot of the variable observations would suggest that they are normally distributed, if observation points on the plot would create an approximate straight line, as shown below, passing through the 1st and 3rd quantiles (the red line on the plots). Thus, for further analysis, as a tool the QQ plots can be used to measure whether variable observations are normally distributed as well as to determine how far from a normal distribution the data set is.

Figure 21 through Figure 24 represent the QQ plots for the CSX Peninsula Subdivision MP 67–69 inspection data (Profile 62 left and right and Yre left and right).

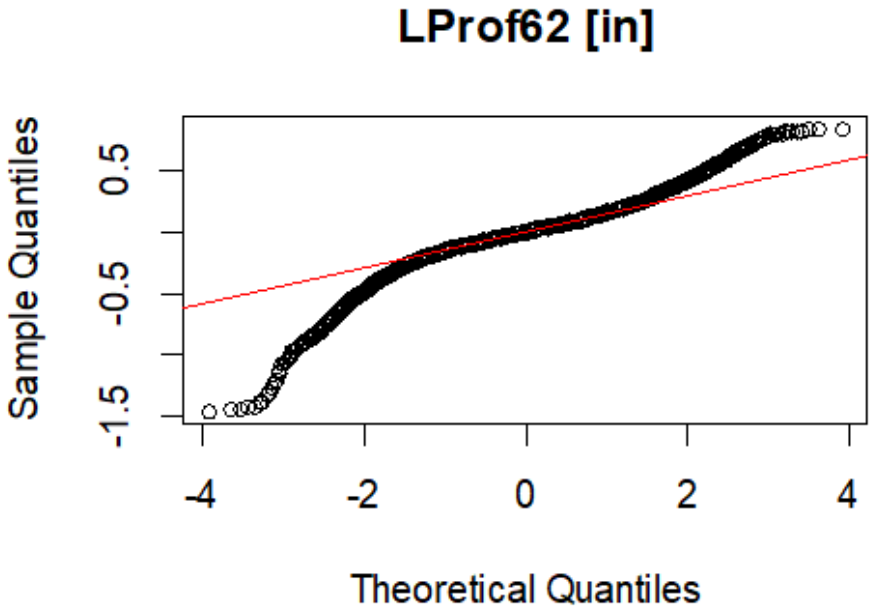


Figure 21: QQ plot, CSX Peninsula Subdivision MP 67–69 inspection data, Left Profile 62

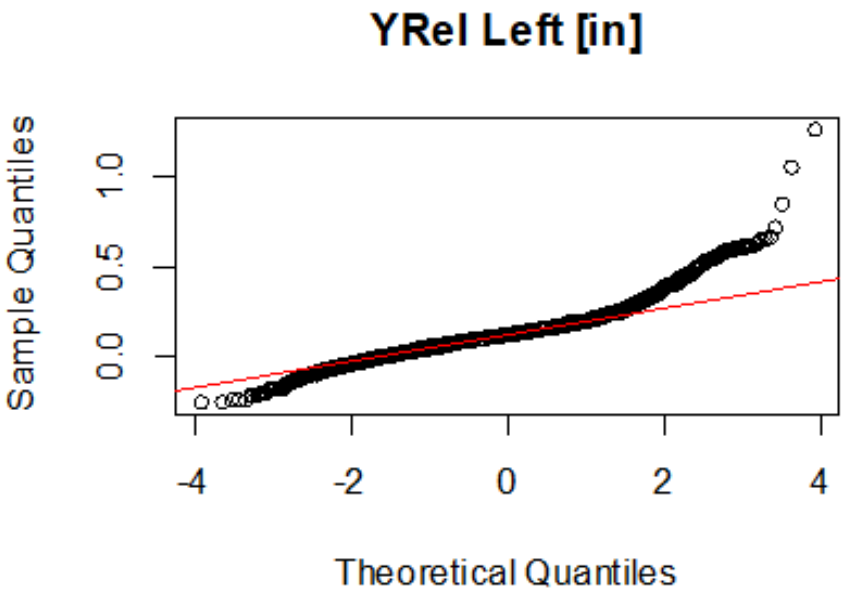


Figure 22: QQ Plot, CSX Peninsula Subdivision MP 67–69 inspection data, YRail Left

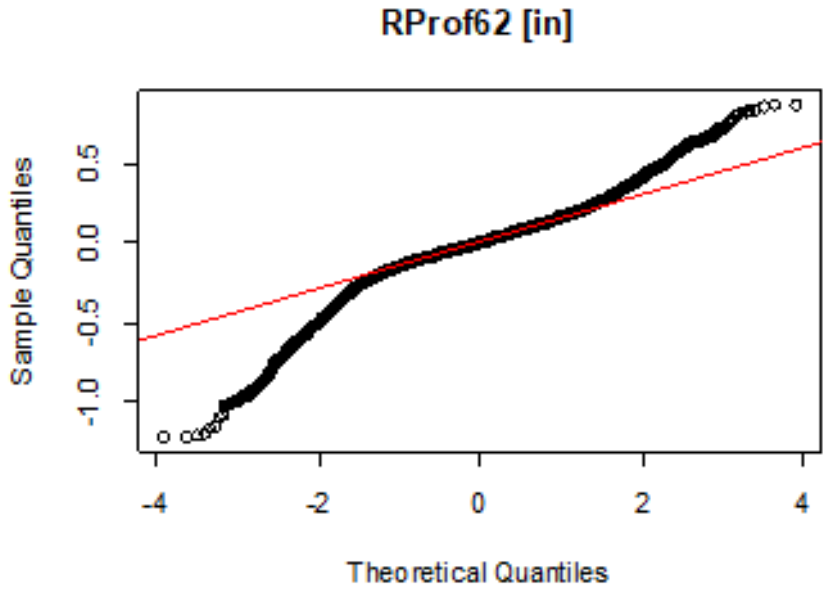


Figure 23: QQ plot, CSX Peninsula Subdivision MP 67–69 inspection data, Right Profile 62

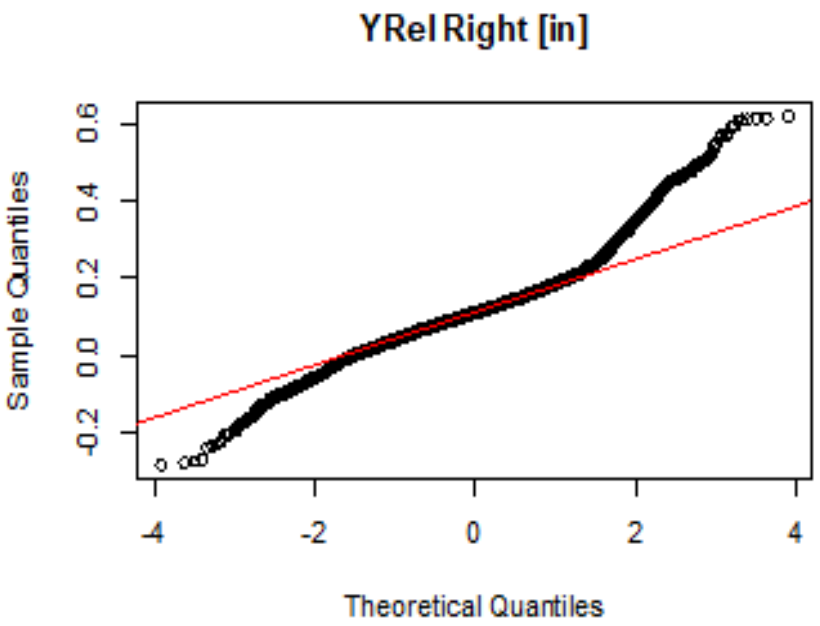


Figure 24: QQ plot, CSX Peninsula Subdivision MP 67–69 inspection data, YRail Right

3.2 Degradation Analysis

Simultaneous with the EDA, an analysis of the degradation of the track geometry data was performed. The focus of this analysis was on the Amtrak continuous track geometry data, where 31 monthly geometry runs over a 4-year period were available for analysis.

Initial analysis of the Oakington Road data focused on overlaying the track geometry data and specifically the Right Profile 62 (profile as measured over a 62-foot chord) and comparing this data with the BFI from the GPR data. This is presented in Figure 25 and Figure 26 which present Right Profile 62 data vs MP as a time series of runs from June 2013 to September 2016. Note Figure 26 shows the absolute value of Right Profile 62. In all the figures, the BFI is presented as red lines, as measured in the center of the track. Note; the BFI is represented as secondary y-axis in the figures (right hand axis) and starts at the BFI= 15 following the Selig Fouling Index

3.2.1 TQI based on Standard Deviation (SD)

To better understand the rate of track geometry degradation, the raw (foot by foot) data presented in Figure 25 and Figure 26 were converted to a TQI based on the standard deviation (SD) of the track geometry over a constant window of 100 ft., or 200 ft. according to the following equation:

$$SD = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}}$$

$\sum_{i=1}^{n}$ – Summation of terms from i = 1 to i = n

$\bar{x}$ – mean value of the measurements in the window

x_i – measurement point value in the window

n - number of measurements in the length of track in the window

The TQI (i.e., SD) was determined for windows of 100 ft. and 200 ft. over all the 31 inspections (time series) from June 2013 to September 2016 on Amtrak's Oakington Road. The analysis was performed on the Right Profile 62 designated in this report as Right Profile 62. The resulting TQI plots are presented in Figure 25 (100 foot window) and Figure 26 (200 foot window). Note, each line represents a different section of track with different ballast conditions, thus the rate of degradation will vary signifciantly from section to section.

In both figures, the degradation of the track, as shown by increasing TQI values with time, can be observed together with the periodic corrective maintenance (surfacing or tamping) which can be identified by the sudden decrease (improvement) in TQI. This can be seen clearly on the top degradation curve in Figure 26 (grey line) where there is a short period of degradation, with a very high TQI (poor track condition), followed in December 2013 by a well-defined tamping activity—where the TQI shows a very significant drop in value (improvement due to the tamping). The track then starts to degrade with time (increasing TQI) for a period of approximately 1 year, until December 2014, when a new tamping activity performed on the track shows an improvement (decrease) in TQI. The cycle continues, but it should be noted that there was a complete rebuilding of the track (and installation of a geocell layer) around September 2015, with a corresponding improvement in track geometry performance. Again, each line represents a different section of track with different ballast conditions, thus the rate of degradation will vary signifciantly from section to section.

It should also be noted that the right BFI antenna value appears to show better correlation to the track geometry degradation as defined by the right surface condition (Right Profile 62).

To better see the relationship between BFI and track degradation (as defined by TQI), sections with differing ballast condition were isolated and examined with regard to standard deviation. The three series in Figure 33 below represents unfouled, moderately fouled and highly fouled ballast sections with corresponding time series of SD100 (SD over 100 feet) from the Right Profile 62 channel of the geometry data. Note, the degradation/maintenance cycles over the time period and the corresponding linear regression fit (with R^2 values of 70 to 87 percent).

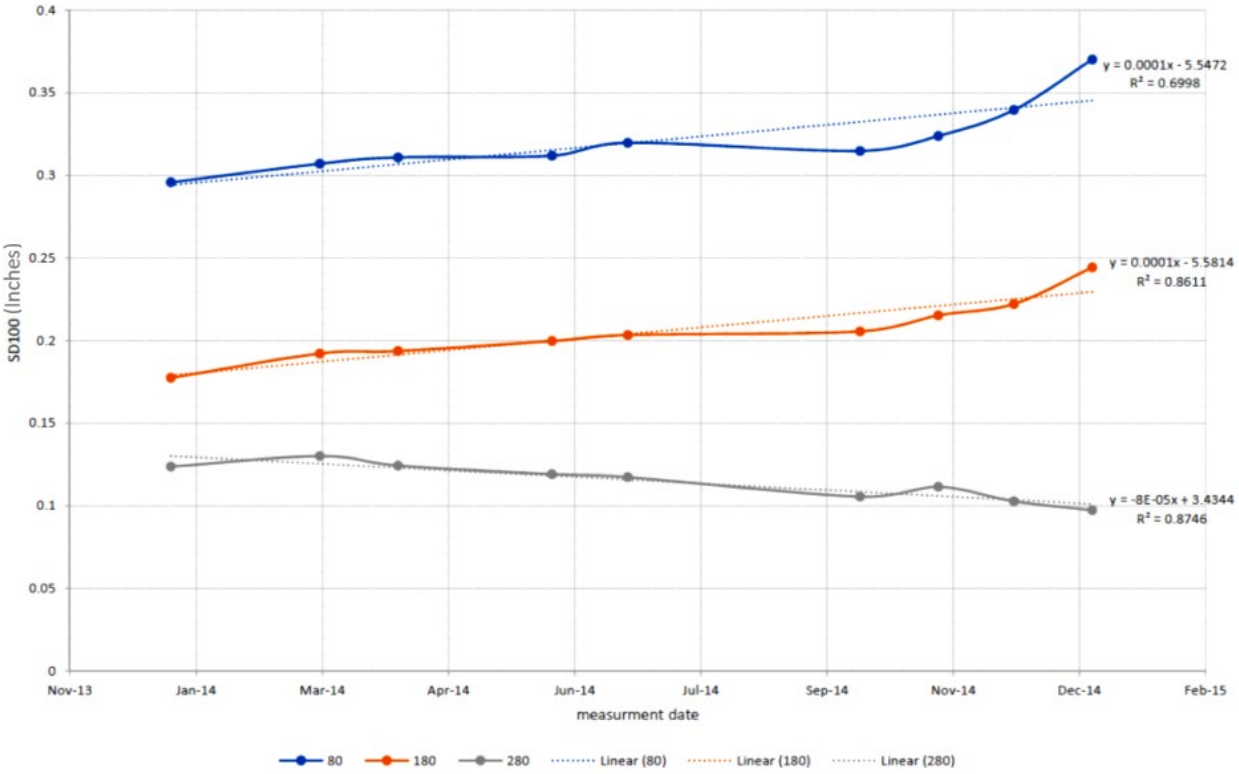


Figure 25: Three sections Right Profile 62 SD100 & linear fit vs. inspection date Amtrak Oakington Road MP 63–64, January 2014 to January 2015

Figure 26 is like Figure 25 with SD calculated on a 200-foot window.

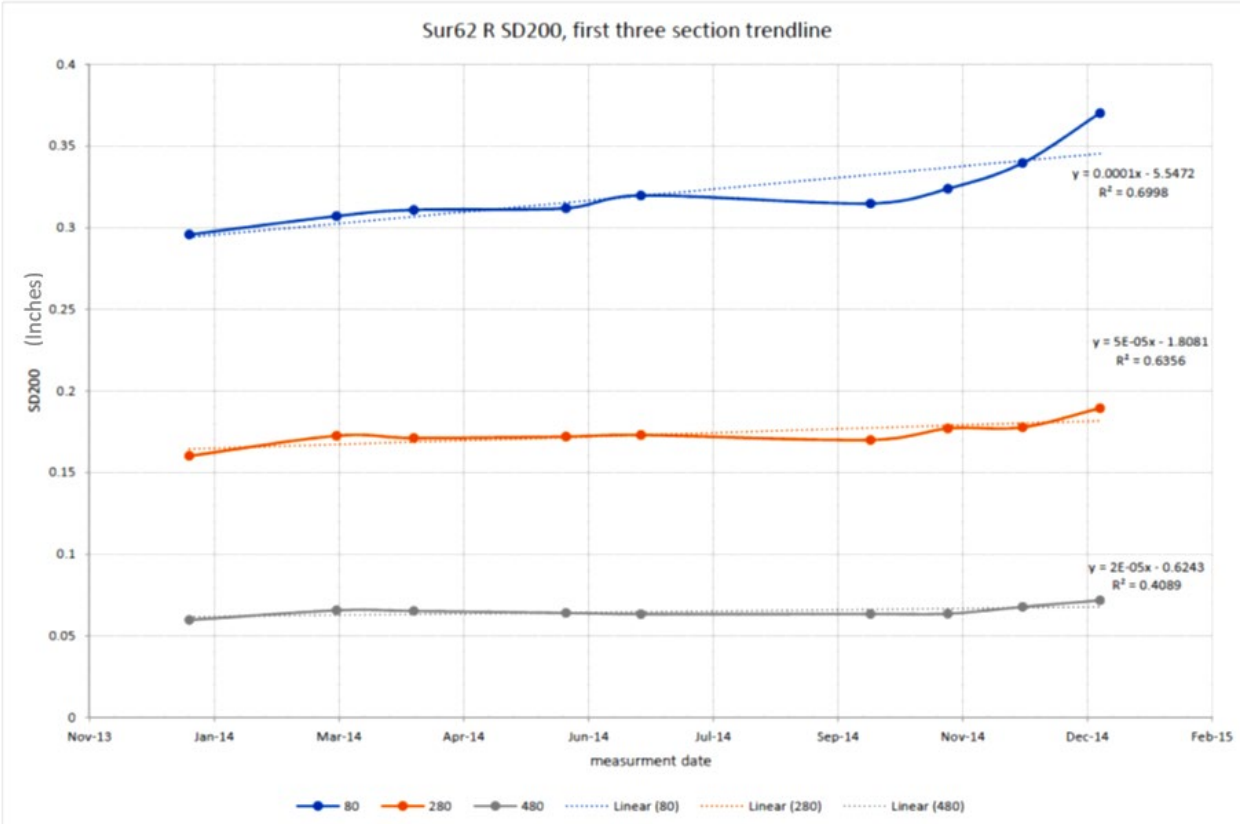


Figure 26: Three sections Right Profile 62 SD200 & linear fit vs inspection date on Amtrak's Oakington Road MP 63–64, January 2014 to January 2015

The rate of degradation of track geometry (using TQIs) can then be directly compared to such GPR parameters as BFI. This is illustrated in Figure 27, which shows three segments of track corresponding to highly fouled (red), moderately fouled (yellow) and relatively clean (green) ballast. The corresponding track degradation relations (using TQI analysis of the three segments over a period of approximately 1 year) show significant rates of degradation for the highly fouled and moderately fouled sections, but no significant degradation for the relatively clean ballast section. Note that the track quality regression curves show very high R^2 values which indicate a good statistical fit.

Figure 28 presents all the sections of Track 2 MP 63–64 of Amtrak's Oakington Road comparing the rate of degradation of the TQI as related to the BFI. Note the significant variation in TQI behavior as a function of the ballast condition as defined by the BFI.

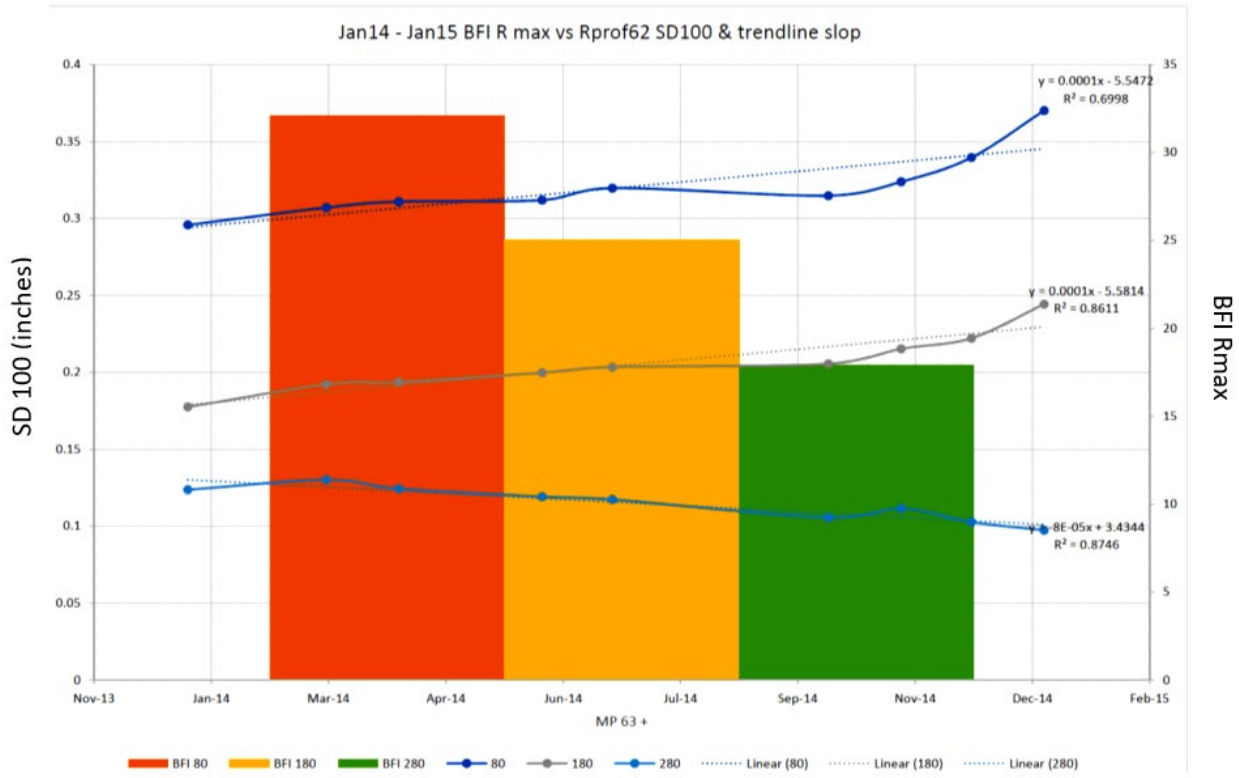


Figure 27: Three sections Right Profile 62 SD100 & linear fit index vs inspection date+ BFI R (sections 80–280) Amtrak Oakington Road, January 2014 to January 2015

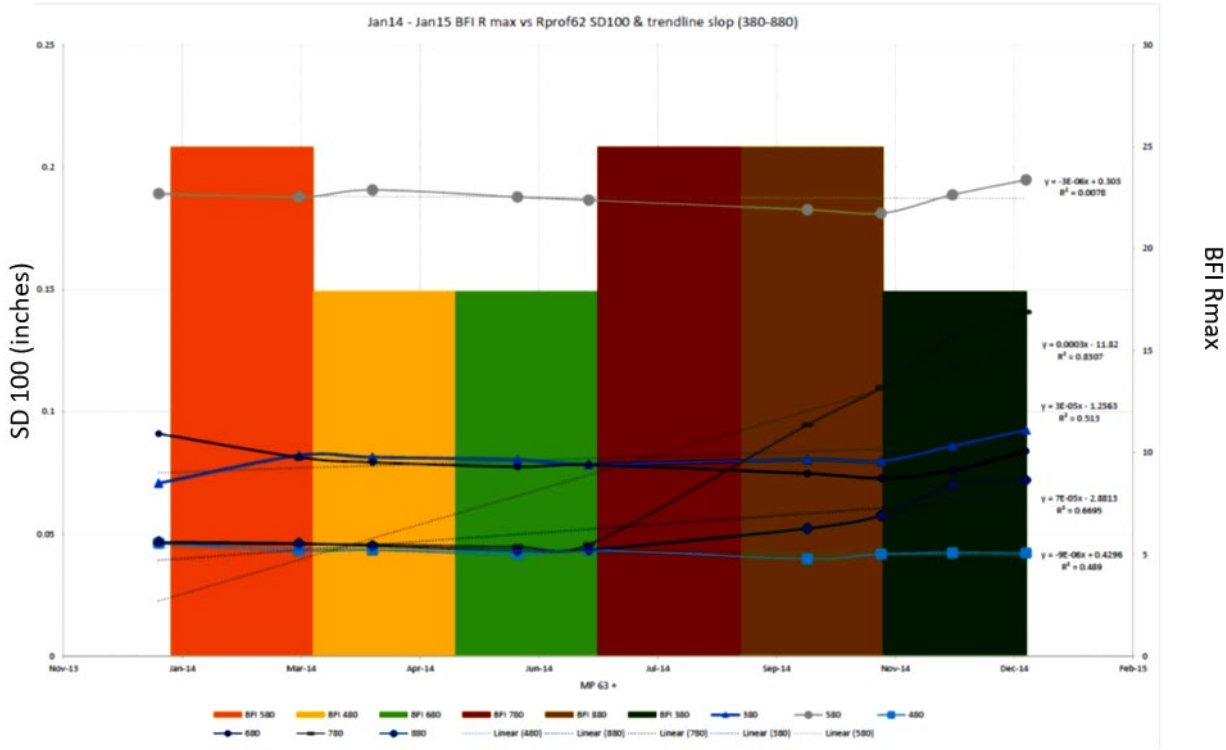


Figure 28: Three sections Right Profile 62 SD100 & linear fit index vs inspection date+ BFI R (sections 380–880) Amtrak Oakington Road, January 2014 to January 2015

The above data suggests that for the higher BFI values (more fouled ballast conditions), the rate of degradation is more rapid than for the less fouled conditions. Furthermore, the data suggests that the December 2013 geometry run represented a "degraded" track geometry condition that could be selected (among the 31 runs) for use in the more detailed analysis.

The results of this EDA and preliminary data analysis stage was a determination that there was a potential relationship between track geometry degradation, and several key GPR measurements to include BFI and BLT. This, in turn led to the application of a next level of data analytics, using LR analysis.

3.3 Preliminary Observations

The goal of this activity to date was to extensively explore the input data sources that may be associated with track geometry degradation, and determine if there was a potential relationship with available subsurface inspection data, namely, GPR and MRail. Particularly, EDA was performed for all the available datasets, along with advanced analytic techniques, parameter combination and data mining activities. As a result, several patterns and relationships began to emerge, which will guide the future analytic techniques to be employed.

The data showed massive and diverse datasets even for the same inspection (repeated, as well as over time), as well as randomness and uncertainty. There was a level of poor-quality data, i.e., missing values, with many outliers. The data is quite variable, as was expected.

The EDA could help identify "bad data" conditions such as seen in the box and whisker plots and correlation analyses (e.g., GPR right channel on CSX was found to be questionable). The EDA results suggest that there appears to be a better correlation between BFI and geometry/profile data, rather than between MRail and geometry. However, there does appear to be some initial correlation between BFI and MRail data. The EDA results also suggest that for the higher BFI values (more fouled ballast conditions), the rate of degradation is more rapid than for the less fouled conditions (e.g., see Figure 35). This EDA analysis also helped identify sections with high probability of having potential subsurface defects or issues.

The EDA analysis sets the stage for and provides guidance to the next level of "Big Data" analysis, which will be discussed in the next section. Among other information, EDA analysis allowed for the identification of specific variables among the extensive array of existing data for further analysis. In addition, it helped identify which is the best source of data, e.g., which chord is preferable to use, what is the best alignment distance to avoid over fit, or under fit (+/- 15, 25, and 50 feet), and how best to align the inspection with varying numbers of observations for the same distance.

Moreover, exploring the data showed the reliability of the inspection variables, which ones should be ignored in the analysis, and which require further investigation. Without performing this type of analysis, results could be very misleading, and the detailed analysis more difficult to implement.

4. Logistic Regression Analysis

The primary analysis tool used in the next phase of the analysis is LR. LR is a regression model where the dependent variable is categorical, i.e., a variable that can take on one of a limited, and usually fixed, number of possible values [5]. A common application is the case of a binary dependent variable, where the output can take only two values, "0" and "1," which represent outcomes such as pass/fail of a defined criterion.

LR is part of the class of algorithms known as Generalized Linear Model (GLM). The main output of the LR model is the probability of an outcome, i.e., the dependent variable. Thus, the binary logistic regression model is used to estimate the probability of a binary response based on one or more predictor (or independent) variables (features), i.e., it models the probability of output in terms of input.

The LR model is trying to find the estimated probability, *P*, where:

$$P = \frac{\text{Outcome of interest}}{\text{all possible outcomes}} \tag{4-1}$$

As part of the process, it calculates the "Odds" that the outcome will occur given a particular exposure:

$$\text{odds} = \frac{\text{Probability of Success}}{\text{Probability of Failure}} = \frac{P(\text{occuring})}{P(\text{not occuring})} = \frac{P(1)}{P(0)} = \frac{P}{1-P} \tag{4-2}$$

The associated odds ratio is:

$$\text{odds ratio} = \frac{odds_1}{odds_2} = \frac{\frac{P_1}{1-P_1}}{\frac{P_0}{1-P_0}} \tag{4-3}$$

Note, the odds ratio for a variable in LR represents how much the odds change with a one-unit increase in that variable holding all other variables constant. In LR we are estimating an unknown P for any given linear combination of the independent variables.

Thus, to estimate the P we take the natural logarithm (ln) of the odds, which is the logit of P

$$\ln(odds) => \ln\left(\frac{P}{1-P}\right) => logit(P) \tag{4-4}$$

$$\text{logit(P)} = \beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k \tag{4-5}$$

Where β = regression coefficients

x = independent variables

k = number of independent variables

Inverting between x and y axis, so y was the probability (also called mean function):

$$logit^{-1}(\alpha) = \frac{1}{1+e^{-\alpha}} = \frac{e^{-\alpha}}{1+e^{-\alpha}} \tag{4-6}$$

α – a linear combination of variables and their coefficients, including regression coefficients and independent variables calculated using maximum likelihood estimation (MLE), which attempts to find the parameter values that maximize the likelihood function, given the observations.

e – constant for the base of the natural logarithm, 2.71828

The LR model estimates the probability of the binary event given the input, therefore solving for P gives:

$$logit(P) = \ln\left(\frac{P}{1-P}\right) = \beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k \tag{4-7}$$

$$\hat{P} = \frac{e^{\beta_0+\beta_1 x_1+\cdots+\beta_k x_k}}{1+e^{\beta_0+\beta_1 x_1+\cdots+\beta_k x_k}} \tag{4-8}$$

This process was applied to the track geometry and GPR data from the previous sections, specifically the Amtrak Northeast Corridor data (Oakington Road) where multiple track geometry data and a complete GPR data set was available. A limited analysis of CSX data to include MRail data was also performed.

The resulting analysis is presented here in three "generations:"

1. Initial generation of the LR model with limited data variables
2. Expanded LR model with additional data variables
3. Higher order hybrid analysis

The specific analysis was performed on two distinct data sets as follows:

- CSX MP 67–69 (limited to initial analysis only)
- Amtrak: MP 62+3,000 to 64+0000

All these analyses are discussed herein.

4.1 Initial Analyses

As noted, an initial analysis was performed of both CSX and Amtrak data sets as follows:

- CSX Analysis
  - Dependent variable: Left Profile, 62-foot chord
  - Independent variables; BFI- L, YRail-L since only Left Profile was being considered
- Initial Amtrak Analysis
  - Dependent variable: Right Profile, 62-foot chord
  - Independent variables
  - BFI Right, BFI Center, and BFI Left

4.2 CSX Analysis

The analysis of the CSX data focused on the one-track geometry run of the DOTX218 car on April 5, 2016, on the Peninsula Subdivision MP 67 to 69. In addition to the foot-by-foot track geometry data, both GPR and MRail data was available for this section. Based on the EDA, the analysis focused on Left Profile (Lprof62), BFI from GPR and YRel Left from MRail.

As noted, the LR analysis uses a binary dependent variable, in this case the probability of a track geometry defect occurring. Thus, the output can take only two values, "0" and "1," which represents a pass/fail of a defined criterion. Analysis of CSX historical maintenance data (when tamping was performed) suggested that Lprof62 value > 0.54 inches represented degraded conditions requiring maintenance. As such, for the analysis of CSX data, the pass/fail binary criterion was the probability that the absolute value of the left profile (62 ft.) was 0.54 inches; thus:

- P(Lprof62) < 0.54 = 0 No Defect
- P(Lprof62)>0.54 = 1 Defect

Thus, the resulting equation was to be of the form

$$P(Lprof62(0.54{:}0.9))=f(BFI, MRail) \quad (4\text{-}9)$$

Where the independent variables were BFI Left and YRel L and the dependent variable is Lprof62.

The resulting LR equation was

$$\hat{P}_{geometry} = \frac{e^{-4.31784+0.04136\cdot BFI-0.58408\cdot Yrel}}{1+e^{-4.31784+0.04136\cdot BFI-0.58408\cdot Yrel}} \quad (4\text{-}10)$$

Note, the exponential nature of the LR equation; this is typical of the LR models.

This relationship is clearly visible in Figure 29, which fixes the YRail value [blue curve at YRel = 0.239 inches] and shows probability of having a profile defect LProf 62 > 0.54 inches vs. BFI Left [orange curve] as a function of increasing BFI (to a maximum probability of 0.42).

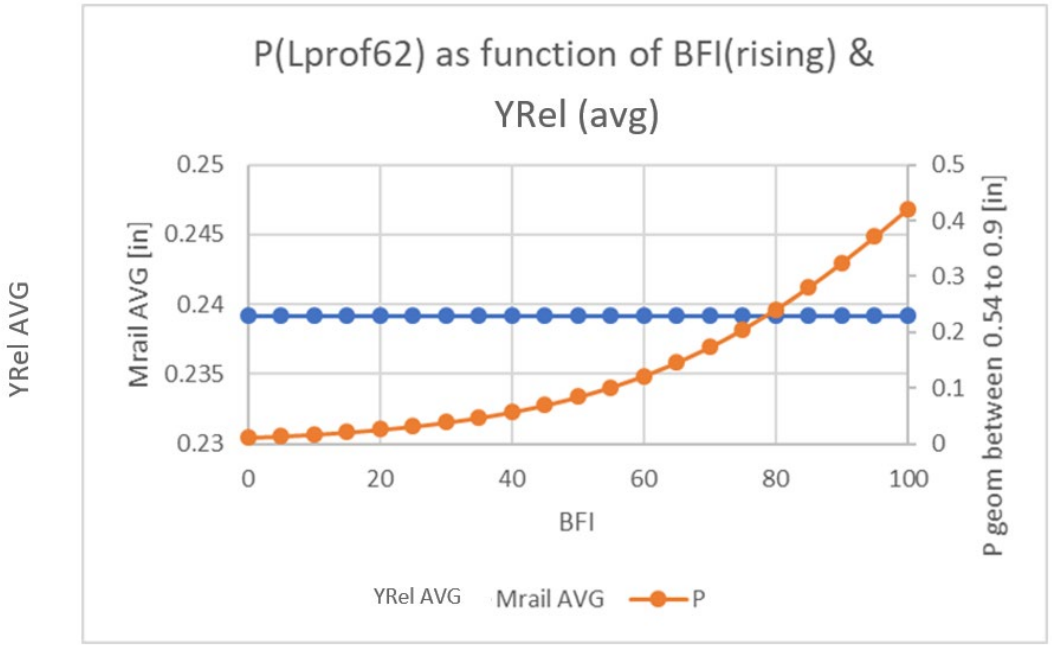


Figure 29: Two-dimensional graph of LR model for CSX MP 67–69; MRail fixed at 0.239 inches

Figure 30 fixes the BFI value [Blue curve at BFI = 22.5] and shows probability of having a profile defect LProf 62 > 0.54 inches vs. YRel (left) [orange].

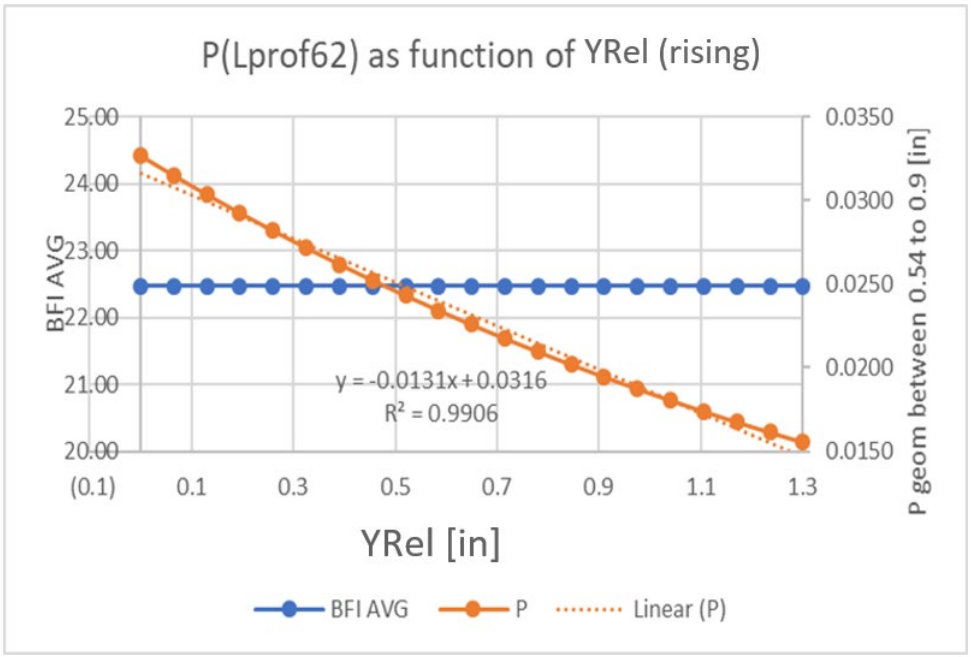


Figure 30: Two-dimensional graph of LR model for CSX MP 67–69; BFI fixed at 22.5

The behavior exhibited in Figure 30 is opposite to what was expected, since increasing YRel is supposed to correlate with poor ballast/subgrade conditions which would result in a high probability of having a geometry defect. Subsequent discussions with FRA indicated that there

was a problem with the MRail measurements in this test, and as a result the MRail data was suspect and not used further in the analysis activity.

The results of the analysis of the CSX data showed a direct relationship between increasing BFI values (increased ballast fouling) and increased probability of having surface/profile deviation in the Surface/Profile 62-foot chord measurement. However, the CSX data was limited in that there was only one geometry run, which was not a "worst case" run (as opposed to the Amtrak data such as illustrated in Figure 27 and Figure 28, where multiple geometry runs allow for the selection of a "poor" or "degraded" geometry condition.) While CSX did provide track geometry exception report data, this exception report data was not sufficiently detailed to allow for the level of analysis required here.

Based on this, the remaining analysis focused on the Amtrak geometry data, though the CSX correlation with BFI was a forerunner of the more detailed relations developed further in this report. Again, as noted, because of the poor correlation with YRel, and the suspect YRel data, no further MRail analysis was performed.

4.3 Preliminary Amtrak Analysis

The analysis of the Amtrak data included the 31 geometry runs between 2013 and 2016, however as a result of the track degradation analysis discussed previously, a "degraded" track condition was evident in the track geometry run of December 2013 between Track 2, MP 62.6–64.0. In addition to the foot-by-foot track geometry data, GPR data was available for this section. Based on the EDA, the analysis focused on Right Profile (Rprof62) and all three BFIs from GPR:

- BFI Left
- BFI Center
- BFI Right

Again, as noted, the LR analysis uses a binary dependent variable, in this case P(abs(Rprof62)), probability of a track geometry exceedance occurring. Thus, the output can take only two values, "0" and "1," which represent a pass/fail of a defined criterion. For the analysis of Amtrak data, the pass/fail binary criterion was the probability that the absolute value of the Rprof62 was greater than 0.4 inches. This threshold was based on analysis of the 31 Amtrak track geometry runs, and when maintenance was performed. Note that this is less than the value of 0.54" used for CSX, which operates at slower speeds and thus can allow for larger values of profile.

- P(abs(Rprof62))< 0.4 = 0 No Defect
- P(abs(Rprof62))>0.4 = 1 Defect

Thus, the resulting equation was to be of the form

$$\text{P(abs(Rprof62(>0.4))=f(BFI L, BFI C, BFI R)} \tag{4-11}$$

Where the independent variables were BFI Left, BFI Center, and BFI Right and the dependent variable is Rprof62. The resulting LR equation was

$$\hat{p}_{geometry} = \frac{e^{-11.474563-0.003287\cdot BFI_{Left}+0.093458\cdot BFI_{\text{Center}}+0.256443\cdot BFI_{Right}}}{1+e^{-11.474563-0.003287\cdot BFI_{Left}+0.093458\cdot BFI_{\text{Center}}+0.256443\cdot BFI_{Right}}} \tag{4-12}$$

Analysis of the equation exponent values shows that BFI Left was not significant (very small weight), and in fact was behaving in the wrong direction (negative sign). Figure 31 presents this as a two-dimensional graph where it fixes the BFI Left and Right values and shows probability of having a profile defect abs(RProf 62)> 0.4inches vs. BFI Center. Note the strong relationship shown between geometry defect and BFI. Furthermore, observe that the curve shows both the mean relationship and the mean +/- one standard deviation. Based on this, the analysis of the Amtrak data was expanded as discussed in the next section.

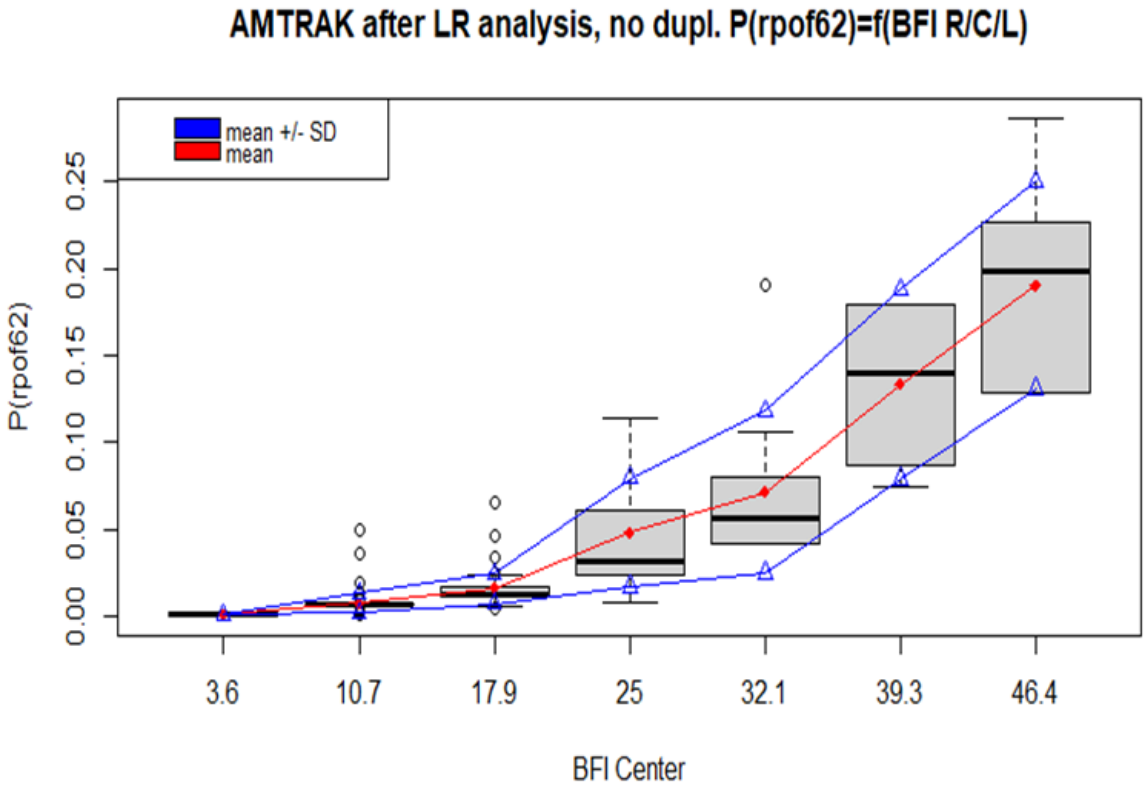


Figure 31: Two-dimensional graph of LR model for Amtrak; BFI Right and Left fixed

5. Expanded Logistic Regression Analysis of Amtrak Data

As noted above, examination of the data set, to include the EDA, suggested that a LR model could be developed to determinate a relationship between the probability of generating a track geometry defect and key GPR inputs (i.e., BFI). The follow up analysis (see below) expanded on this to also include BLT, as measured by GPR.

Again, as in the previous initial analysis, the two key data sets used were GPR data and track geometry data. Since the objective was to develop a relationship between GPR data measurements of ballast and subgrade condition and the probability of generating a track geometry defect; the GPR data served as the primary independent variables and the track geometry as the dependent variable. As such there was extensive data preparation, as noted in the previous chapters to include

- Expanded GPR data analysis, preparation, and filtering
- Geometry data preparation to include correlating the geometry data with GPR data and create single database

Figure 32 presents the track geometry input used, specifically, the right profile data as taken from the track geometry car (specifically the Rprof62 channel which is the right profile as measured over a 62-foot chord). Note, the specific geometry run used in this analysis was dated December 2013 which represent a "worst case" condition. Note, this data was provided in digitized format with data points at one foot intervals. The GPR output for MP 62+3100 (62.6) to 63+3000 (63.58) matches the track geometry data section.

Note, the GPR data encompasses the three sections of Track 2 (Left, Center, and Right) though as noted earlier, the focus will be on the Right rail, and thus only the center and right GPR measurements are used. The key GPR input values used in this analysis are as follows:

- BLT (as determined from the top and bottom of ballast layer for center and right of Track 2)
- BFI for
  - Center
  - Right

As noted earlier, this data was digitized manually, based on 16.7 intervals. Note, the BFI digitization used the color-BFI index relationship shown previously in Table 6. In general, moderately fouled ballast has a BFI value greater than 15.

As noted, this data set was consolidated into a unified database, which included matching between GPR and geometry inspection measuring points and creation of a mutual data frame of reference, taking into account different data sampling rate. After alignment of the inspection data by shifting the signal to match the peaks, the signals were consolidated into a common reference MP, noting that each inspection has different sampling steps and corresponding different number of measurements in the approximately 2 miles of data.

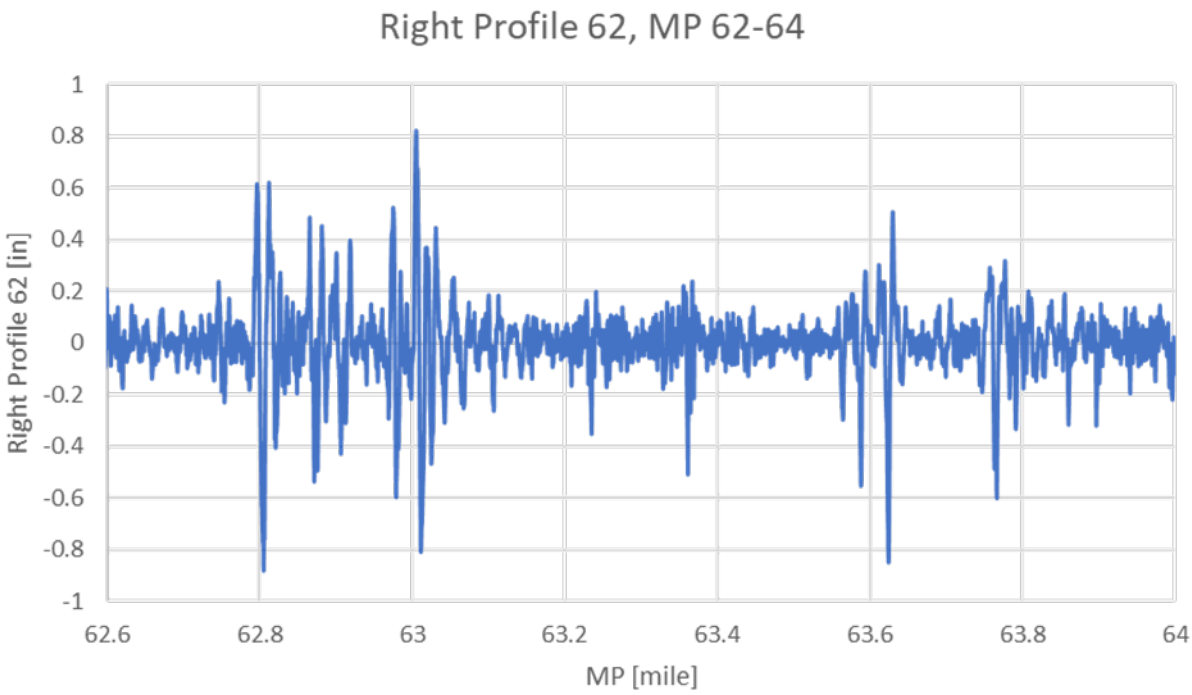


Figure 32: Right Profile 62 (December 2013) by MP

After performance of EDA as illustrated in Figure 33 it was determined that the potentially most useful GPR data channels were ballast thickness, as determined at the center of the track, and BFI. Thus, for the right profile, BFI Right and Center were both useful.

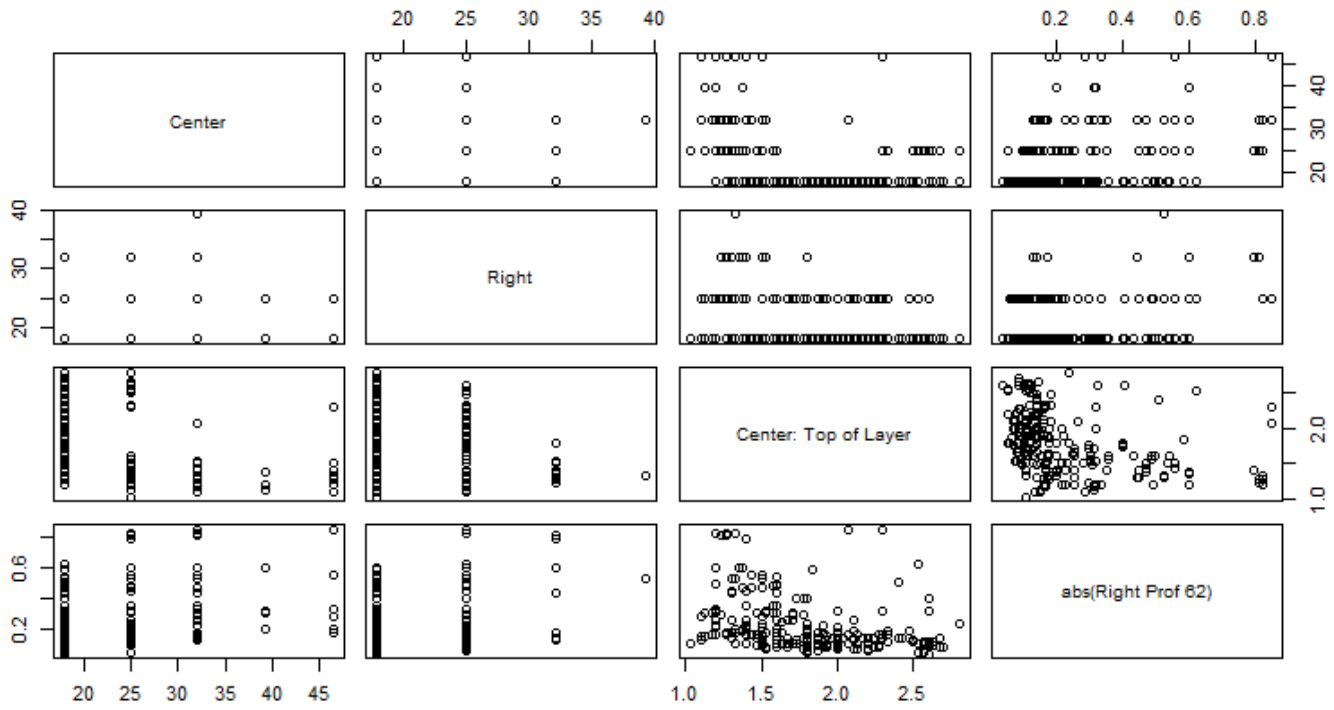


Figure 33*:* Amtrak MP 62+3,000-64+0000 relationship between BFI, Thickness C[13]

Thus, using the right rail profile (62-foot chord) as the dependent variable and the GPR values of BLT Center, and BFI Right and Center as the independent variables, the LR model was developed.

LR modeling determines a probabilistic relationship between the independent variables (GPR measurements) and the dependent variables (track geometry defects). As noted earlier, LR is a regression model where the dependent variable is categorical, i.e., a variable that can take on one of a limited, and usually fixed, number of possible values [5]. A common application is the case of a binary dependent variable, where the output can take only two values, "0" and "1," which represent outcomes such as pass/fail of a defined criterion.

As before, the LR model estimates the probability P of the binary event given the input, therefore solving for P gives:

$$logit(P) = \ln\left(\frac{P}{1-P}\right) = \beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k \tag{5-1}$$

$$\hat{P} = \frac{e^{\beta_0+\beta_1 x_1+\cdots+\beta_k x_k}}{1+e^{\beta_0+\beta_1 x_1+\cdots+\beta_k x_k}} \tag{5-2}$$

Where β = regression coefficients

x = independent variables

k = number of independent variables

[13] BLT Center

Using the right rail profile as measured over a 62-foot chord as the dependent variable and the GPR values of BLT Center and BFI (Right and Center) as the independent variables, the LR model was developed. However, since the output of a LR model must be binary in nature, it was necessary to convert the profile data into a binary data set. This was done by defining the absolute value of the profile data as being

- P(abs(Rprof62)) < 0.4 = 0 No Defect
- P(abs(Rprof62))>0.4 = 1 Defect

The threshold of 0.4 was selected based on the analysis of the track geometry (profile). This further allowed for a statistically significant number of data points to be used in the LR analysis.

The database used in the analysis included:

- Right Profile 62 measured on December 2013; abs(Rprof62)
  - Note, this was the worst measured profile condition among the approximately 31 set of inspections available.
- BFI Right and BFI Center
  - Digitized every 16.7 feet
  - After removal of values of BFI smaller than 15
- Ballast Layer Thickness as measured in the track BLT Center

The resulting LR models thus calculated the probability that the Right Profile 62 exceeded a value of 0.4 ("defect") as a function of BFI Center and Right, and BLT center. The resulting relations, in generalized form is given as:

$$P(abs(Rprof62) > .4) = f[BFI_{center}, BFI_{right}, BLT_{center}] \quad (5\text{-}3)$$

The resulting LR equation calculated from this dataset utilizing the R software is:[14]

$$logit(P) = \ln\left(\frac{P}{1-P}\right) = -4.98 + 0.04 \cdot BFI_{center} + 0.18 \cdot BFI_{right} - 0.92 \cdot BLT_{center} \quad (5\text{-}4)$$

$$\hat{P}_{geometry} = \frac{e^{-4.98+0.04 \cdot BFI_{center}+0.18 \cdot BFI_{right}-0.92 \cdot BLT_{center}}}{1+e^{-4.98+0.04 \cdot BFI_{center}+0.18 \cdot BFI_{right}-0.92 \cdot BLT_{center}}} \quad (5\text{-}5)$$

As can be seen in this equation, the model coefficients can be defined in terms of an increase in logit score in one-unit as follows:

- Increase in BFI Right is 0.18
- Increase in BFI Center is 0.04
- Decrease in BLT Center is 0.92

As a result, the probability of having a profile "defect" increase as a function of BFI (with BFI Right more significant than BFI Center) and decreases with ballast thickness. This corresponds with engineering expectations and was discussed further under sensitivity analysis.

[14] Note the coefficients of the equation come directly from the analysis performed with the R software

Looking at the statistical significance of the resulting LR model, according to the z-test, there is a strong likelihood for a relation between BFI Right and target variable, i.e., abs (Rprof62)>0.4 (i.e., presence of as profile defect). Furthermore, the test suggests that BFI Right is a very significant variable in predicting the variable abs(Rprof62)>0.4 (presence of a profile defect). The results of BFI Center and BLT Center p-values are not as strong and implies a relationship that is not as strong or well defined as the BFI Right relationship.

It should be noted that several different LR models were developed as part of this activity. Appendix B presents three of these, where Model 3.1 in Appendix B corresponds to the LR model presented in this report. The other models did not have the same level of performance and are presented in the appendix for completeness.

5.1 Sensitivity Analysis

This section presents the sensitivity of the LR model to the three independent variables used as discussed previously. Noting that the model has three independent variables, there are six permutations when presented as two-dimensional sensitivity graphs and three permutations when presented as a three-dimensional graph. Effective illustration of the sensitivity was achieved by preparing plots as follow:

- One independent variable was presented as a continuous function
- Second independent variable as a set of three values (Minimum, Average, and Maximum)
- Third independent variable held constant

Using this approach, it is possible to observe the influence of each parameter in predicting the probability of having a profile exceedance [P(abs(Rprof62)>0.4)]. The results are illustrated in two- and three-dimensions charts.

Figure 34 shows the probability of having a profile defect [P(abs(Rprof62)>0.4] as a function of BFI Right for three cases:

- Ballast Layer Thickness (BLT) = Minimum
- Ballast Layer Thickness (BLT) = Average
- Ballast Layer Thickness (BLT) = Maximum

Note, BFI Center is held constant at its average value for this graph.

As can be seen from this graph, the probability of having a defect is very sensitive to BFI Right, with increasing ballast fouling (higher BFI value) causing a greater probability of having a defect. Note, the inverse sensitivity to ballast thickness, with decreasing ballast thickness increasing the probability of having a defect, as expected from engineering experience.

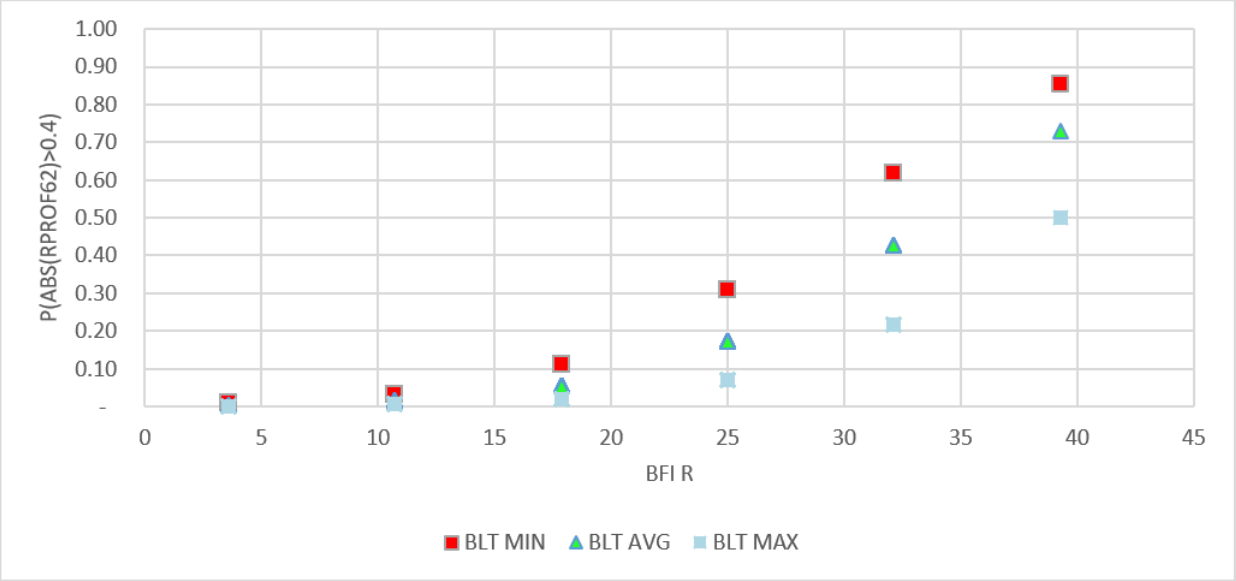


Figure 34: Probability of geometry defect as a function of BFI Right and BLT

Figure 35 shows the same behavior in three-dimensional. Again, BLT has significant importance and influence on the likelihood of having a profile defect, the thicker the layer of the ballast in the center of the track the lower the probability of having a profile defect. BFI Right also has a strong influence on the profile defect likelihood, but it is significantly lower when the ballast layer in the center of the track is thick.

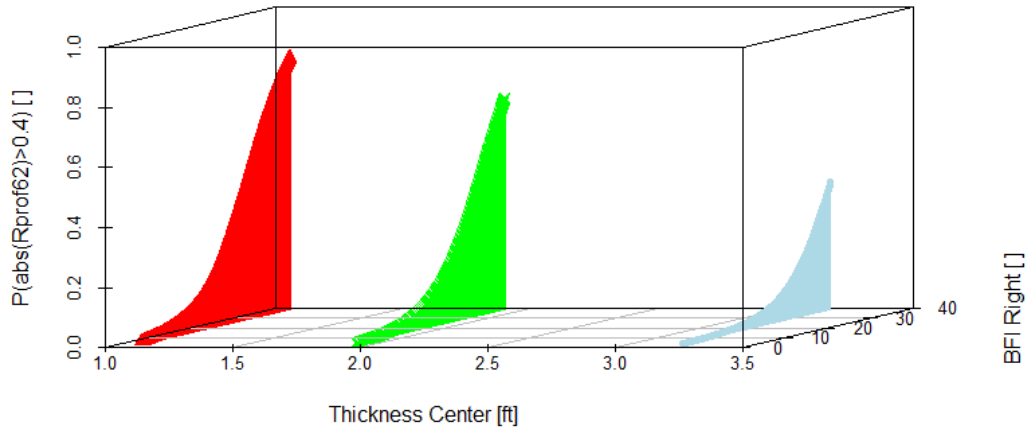


Figure 35: Probability of geometry defect as a function of BLT and BFI Right

Figure 36 and Figure 37 show the same data in an inverted format, with the probability of having a profile defect (P(abs(Rprof62)>0.4) as a function of BLT, for three cases:

- BFI Right = Minimum
- BFI Right = Average

- BFI Right = Maximum

Note, BFI Center is held constant at its average value for this graph.

As can be seen in these graphs, the probability of having a defect is inversely sensitivity to ballast thickness, with decreasing ballast thickness increasing the probability of having a defect, as expected from engineering experience. Note, the strong sensitivity to BFI Right.

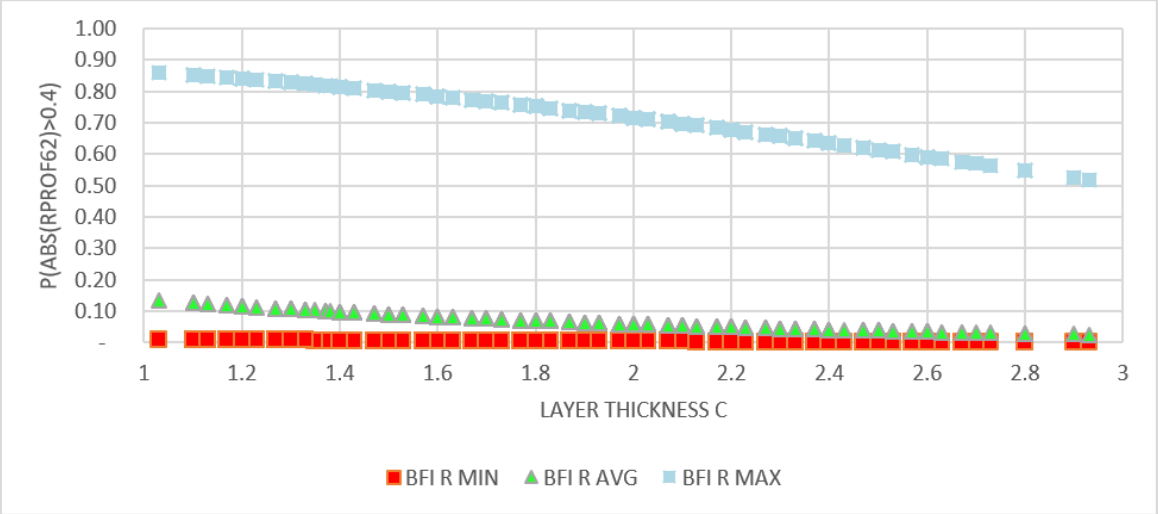


Figure 36: Probability of geometry defect as a function of BLT and BFI Right

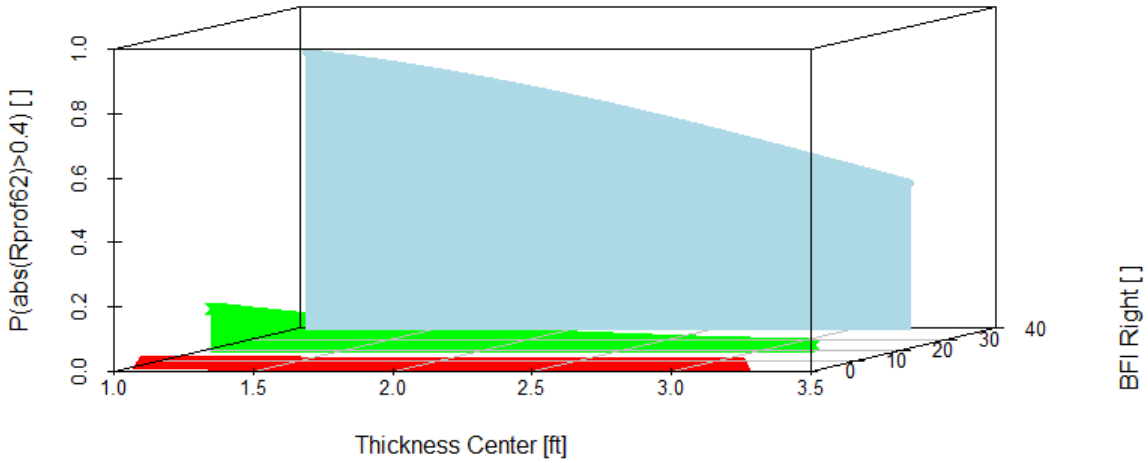


Figure 37: Probability of geometry defect as a function of BLT and BFI Center (alternate view)

Figure 38 and Figure 39 show the probability of having a profile defect [P(abs(Rprof62)>0.4] as a function of the two BFI parameters, holding BLT constant. In this case, BFI Right, is continuous and BFI Center is presented for three values:

- BFI Center (BFI C) = Minimum

- BFI Center (BFI C) = Average
- BFI Center (BFI C) = Maximum

Note, BLT Center is held constant at its average value for this graph.

As can be seen from this graph, the probability of having a defect is sensitive to both BFI Right and BFI Center, however, the sensitivity to BFI Center is not as great as that observed for BFI Right.

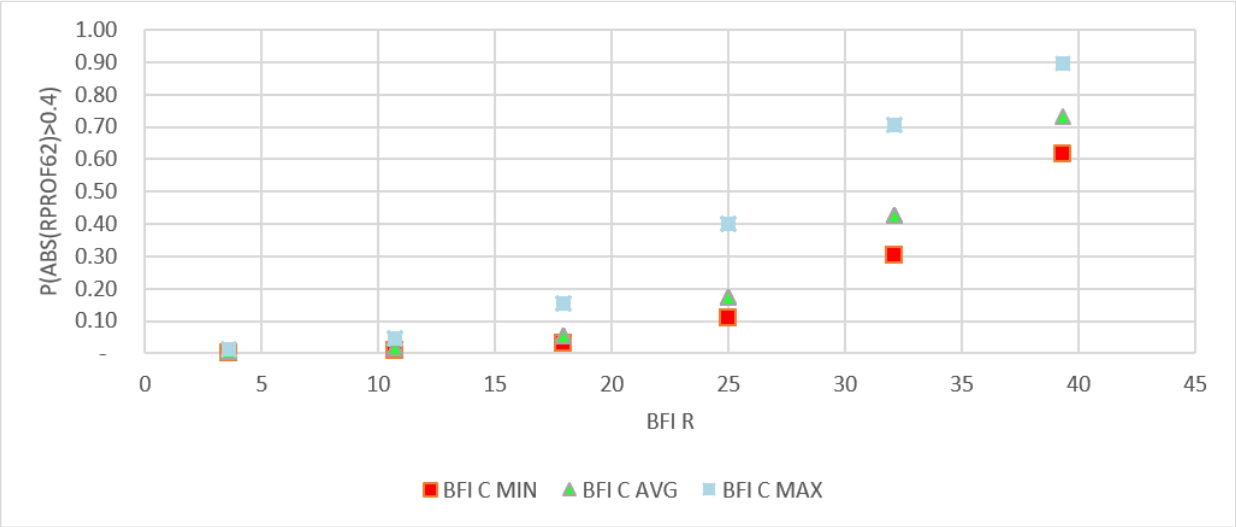


Figure 38: Probability of geometry defect as a function of BFI Right and BFI Center

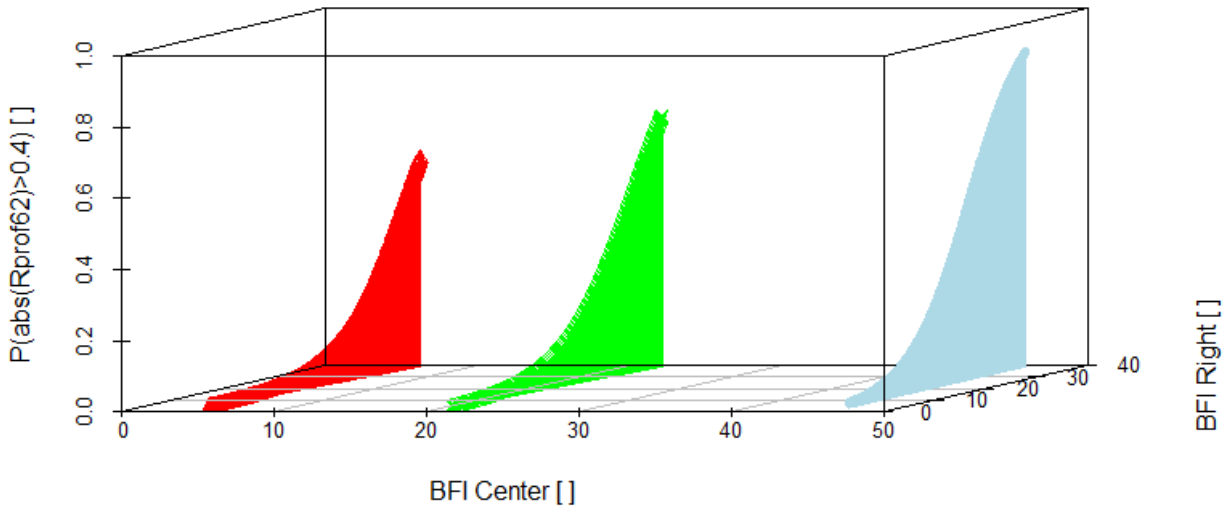


Figure 39: Probability of geometry defect as a function of BFI Center and BFI Right

Figure 40 presents a complementary view that shows BFI Center has low influence on the probability, especially when the BFI Right values are low. Thus, when BFI Right is less than 10 (no fouling) the probability of a defect is close to 0. When BFI Right is 40, fouled ballast, there

is a definite effect, but the effect of increasing BFI Center is not as significant as compared to BFI Right.

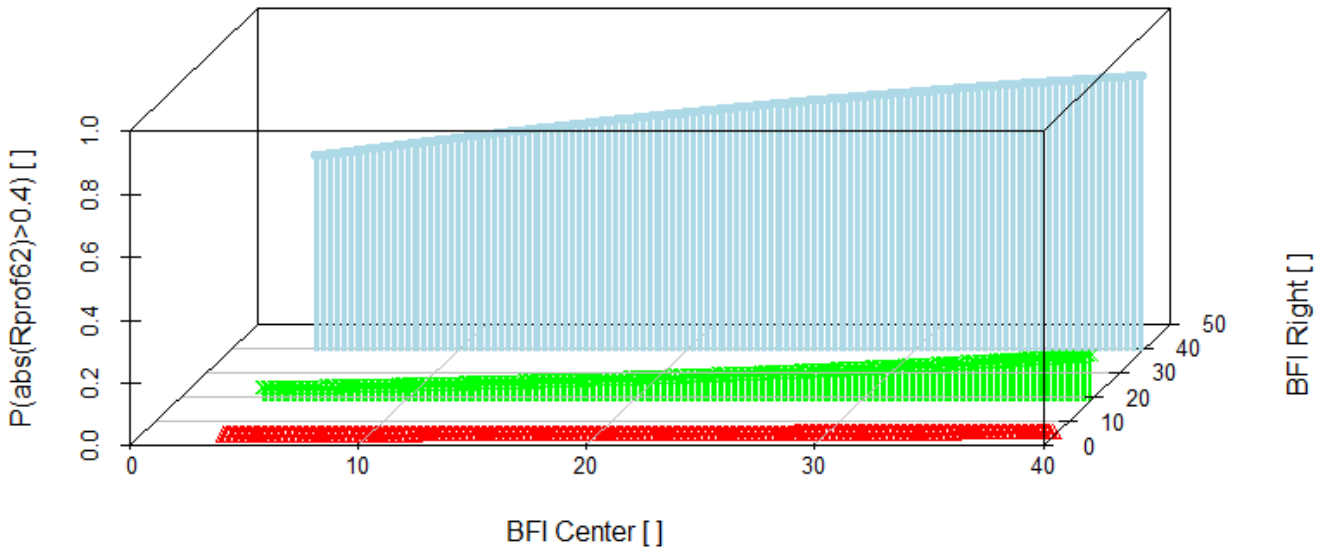


Figure 40: Probability of geometry defect as function of BFI Center and BFI Right (alternate view)

Finally, in Figure 41, the sensitivity of BFI Center and BLT shown, with BFI Center a continuous value, and BLT presented as discrete values. From the figure, it is shown that BFI Center has little influence on the probability of having a profile defect and BLT has relatively significant influence on the probability. Having sufficient thickness ballast layer, e.g., higher than 2 ft., will dramatically reduce the likelihood of having a profile defect even when the BFI Center is very high. As noted previously, the sensitivity to BFI Right is greater than that of BFI Center.

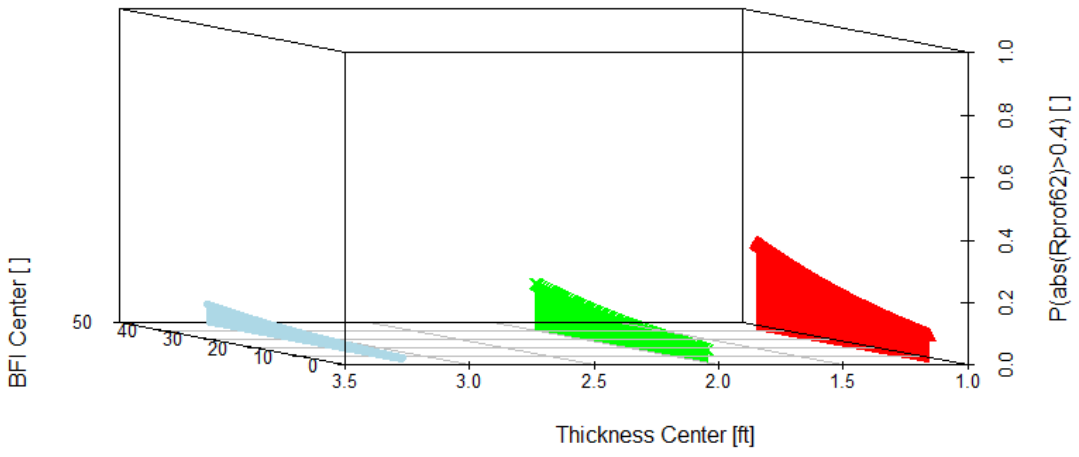


Figure 41: Probability of geometry defect as a function of BFI Center and BLT

While summarizing the results of the sensitivity analysis, the following is clear:

- BFI Right has the highest influence on the probability for having a profile defect.
- The two highest probabilities for having a profile defect is a combination of:
    - High BFI Right and BFI Center, the probability of having a profile defect is 90 percent. (BLT equals its average value of 23 inches)
    - High BFI Right and low BLT, the probability of having a profile defect is 84 percent. (BFI Center equals its average value of 17)
- For a thick ballast layer, the highest probability is of the order of 42 percent, even for highly fouled conditions (high BFI values). For a thin ballast layer, this rises to 84 percent for highly fouled conditions.

5.2 Statistical Validation

To statistically validate the model, an error matrix or "confusion matrix" approach [6] was used as illustrated in Table 13. The confusion matrix validation process constructs the model from the entire training dataset, then separates the dataset into parts, and applying the model on the subdivided datasets. By counting the predicted versus actual observations, a 2x2 matrix can be constructed showing the number of correct and incorrect predictions based on the reference and the corresponding performance of the model.

Table 13 presents the results of the confusion matrix application to the LR model and associated training dataset. As can be seen in Table 13, of the 253 predictions, 220 (213+7) or 87 percent match the actual (true condition) value, i.e., predicted positive = actual condition positive and predicated negative = actual condition negative. Of the remaining 13, 29 or 11.5 percent are false positive and 4 or 1.5 percent are false negatives. The analysis showed that the true positive rate (TPR), i.e., sensitivity of the model is 98.16 percent, [213/217] which is very good and the false negative rate (FNR), "miss" rate of 1.86 percent [4/217] is very low. The other statistical values also support the validity of the LR model in this application.

Table 13: "Confusion" matrix for logistic regression analysis

			True condition	
	Total population		**Condition positive**	**Condition negative**
Predicted condition	**Predicted condition positive**		**213**	**29**
	Predicted condition negative		**4**	**7**

Accuracy
86.96%

	True positive rate (TPR), Sensitivity	False positive rate (FPR), probability of false alarm
	98.16%	80.56%
	False negative rate (FNR), Miss rate	True negative rate (TNR), Specificity (SPC)
	1.84%	19.44%

6. Hybrid Analysis

Following up on the initial LR analysis, a more extensive data analytics approach using combinational hybrid analysis was implemented and applied to the dataset. The objective was a higher order polynomial LR model, with increased accuracy, for the determination of the probability of a rail profile defect occurring at locations with measured ballast fouling and measured ballast thickness.

As part of this more comprehensive analysis approach EDA was again used to map the initial correlation between the GPR and profile data, and to identify the GPR parameters that appeared to be most influential for the profile degradation analysis. As noted previously, EDA is an approach that allows a first insight into data by means of a variety of analytic techniques, many of them graphical [4]. EDA helps characterize the data where there are anomalies in the variables (outliers), or if there are complex relationships within the variables, patterns, etc.

The results were then used in a combinational hybrid analysis of emerging and well-established data analysis techniques consisting of hierarchical clustering analysis of histogram-valued data, a corresponding application of higher degree polynomial functions on the defined parameters, and the generation of the LR model based on these higher order polynomials.

Figure 42 presents the data analysis steps and their order of application.

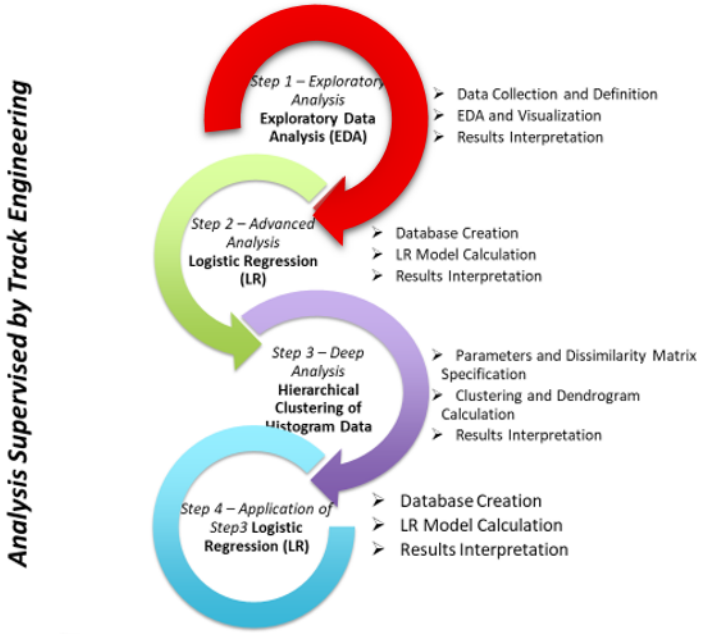


Figure 42: Hybrid analysis steps

6.1 Hierarchical Clustering Analysis of Histogram-Valued Data

Hierarchical clustering analysis of histogram-valued data is an unsupervised classification technique that combines both the clustering analysis and the symbolic data analysis. Both methods are well established and widely used in data analysis. The advantage of using these combinations allows for better classification of massively large datasets.

6.1.1 Introduction

The clustering method is a machine learning technique whose main objective is to automatically group a dataset, such that similar data objects (samples) are within one cluster [4]. The objects are grouped together into a cluster based on a pre-defined and selected measures to determine the underlying structure within a dataset. This analysis used the clustering method as a first step before the LR modeling since its goal is to group the data by dividing it to similar parameters, thus bringing meaningful sense of their behavior.

Clustering analysis belongs to the category of unsupervised learning, i.e., where examples are not labeled [4], therefore, the model tries to understand the patterns of data in it. According to Milligan [5], the steps for cluster analysis are:

- Choosing the objects to be clustered
- Choosing the measurements or variables
- Standardization of variables
- Choosing a (dis-)similarity measure
- Choosing a clustering method
- Finding the number of clusters

The clustering analysis used here is polythetic, unlike monotectic, where each category members are alike, as well as hierarchical. The roots of the hierarchical clustering methods go back to 1960s, and suggest that the clustering will have paternal structure ordering from top to bottom [7]. However, the polythetic hierarchical clustering, as used in this section, is done by an agglomerative approach, which groups the members from the bottom up. This way, each cluster's observation is considered as its own cluster, i.e., node, and new members, will join iteratively to the existing cluster until all members are joined to the most common clusters and only a single cluster, i.e., root remains. The result of that process is a hierarchical pattern structure of the clustered members on distance and nodes, i.e., a tree plotted as a dendrogram. Various clustering schemes share this procedure as a common definition, but differ in the way in which the measure of inter-cluster dissimilarity is updated after each step.

The proximity between each of the other members that define the clusters are determined by the linkage type, distance, and the method of its calculation. Many different types of methods are used to establish for the linkage between clusters, to include the more commonly used: single linkage, complete linkage, average linkage, and centroid methods. The method to calculate the distance between clusters for these methods will be presented in the next section.

The clustering analysis groups symbolic data, rather than classical data. Symbolic data analysis (SDA) is an extension of standard data analysis where symbolic data tables are used as input and symbolic objects are made output as a result [7]. The data units are called symbolic since they are more complex than standard ones, as they not only contain values or categories, but also include internal variation and structure. The symbolic representation of a variable is not a single value, e.g., mean of dataset, but a new variable that stands for variability of the entire dataset. Symbolic data as used in this research refers to distributional data representation, i.e., histogram-valued variables.

6.1.2 Theory of Wasserstein-Distance Based Mean

For computing the distance between the distributions, the Wasserstein distance method is used, which gives the possibility of defining a single center in the form of a distribution. There are different formulations of the Wasserstein distance in the literature, but the L2 version of the Wasserstein distance as defined by Reference 10 was used. This formalization defines the distance (dw_p) between two densities Φ_i and $\Phi'_{i'}$ using the quantile functions Φ_i^{-1} and $\Phi_{i'}^{-1}$ that associates with cumulative distribution functions (CDF) ϕ_i and $\phi_{i'}$ [9] as follows:

$$d_{W_P}(\phi_i, \phi_{i'}) = \left(\int_0^1 \left|\Phi_i^{-1}(t) - \Phi_{i'}^{-1}(t)\right|^P dt\right)^{\frac{1}{P}} \quad (6\text{-}1)$$

Where, P is the probability, as defined in Equation 5-2.

To avoid multiple solution for the distance, the following formula was proposed for the L2 Wasserstein distance between two probability distributions:

$$d_W(\phi_i, \phi_{i'}) = \sqrt{\int_0^1 [\Phi_i^{-1}(t) - \Phi_{i'}^{-1}(t)]^2 dt} \quad (6\text{-}2)$$

In the equation below Y stands for the Fréchet mean, i.e., a single representative point for a group of points, with respect to dw and assuming equal weights w_i. The mean quantile function (Mw) below solves optimization problem for distribution variable (Y). Note that arg min returns the combination of parameters that minimizes the function.

$$M_W(Y) = \arg\min(x) \sum_{i=1}^{n} d_W^2(\Phi_i, x) \quad (6\text{-}3)$$

6.1.3 Benefits and Motivation for Application of HCA of Histogram-Valued Data

Clustering methodology analysis is a very useful tool to reveal the relationship between datasets parameters and is well established for classical data [8]. However, with extremely large datasets, such as the railway track data used in these analyses, they are difficult to implement using traditional approaches [4]. Aggregation of classical data as into symbolic data, e.g., histograms data, is one of the tools used when confronted with large, excessively large, datasets [8]. It reduces the number of parameters by the creation of symbolic objects, i.e., categories, thus reducing the computational complexity associated with the large datasets.

In addition, track inspections generate multivariable data, which is very complex to analyze for relationship development and classification. Thus, unsupervised machine learning algorithms like hierarchical clustering analysis with symbolic data are used to provide enhanced analyses.

Histogram-valued data defines each variable as a histogram, as presented in the following formula.

$$Y_u = (\{[a_{jk}, b_{jk}], P_{ujk}; k_j = 1, \ldots, s_{uj}\}, j = 1, \ldots, p), u = 1, \ldots, m \quad (6\text{-}4)$$

Where,

Y_u = histogram of random variable Y of observation u

P_{ujk} = relative frequencies associated with the subinterval $[a_{jk}, b_{ujk})$

S_{uj} = number of subintervals

These two techniques, hierarchical clustering analysis and histogram-valued data analysis, can then be combined into a composite hierarchical clustering analysis with histogram data. The resulting hierarchical clustering analysis with histogram data analysis approach can shed light on deep relationships between track structure components and conditions, which are difficult to see and to prove. As these non-obvious relationships are identified, a more accurate model, or equation for the relationship, can be generated, e.g., a model following LR analysis. Due to this, hierarchical clustering analysis with histogram data is one of the methods used for finding the most representative relationship between the key variables from the track geometry inspection and the substructure condition as measured by GPR.

Data Preparation

A key part of the hierarchical clustering analysis of histogram-valued data is dataset preparation. During that process, the variables of geometry inspection Rprof62 (Right Profile 62) and GPR inspection. BLT, and BFI Center and BFI Right were accurately aligned using the inspections MP. Missing values in the data were either removed or filled-in using linear interpolation.

For example, while GPR data was recorded continuously, the data sampling rate used was 100 ft. for BLT Center and 16.67 ft. for BFI Center and BFI Right. Furthermore, the track geometry data recording rate was foot by foot so that there was a 1-foot sample rate for Rprof62. Noting that the BLT did not vary abruptly but rather varied gradually, the chosen interval rate for all the variables was selected as 16.67 feet. The Rprof62 data were matched to this interval, while BLT Center linear interpolation obtained this data interval.

Data Normalization

To ensure that the data could be used for the analysis, they had to be "standardized," through either normalization or scaling. Standardization of variables is a form of transformation, but with a different ration. Standardization turns the focus of a distance measurement of the clustering analyses between given variables, rather within each variable. That is because the normalization process allows the analyst to ignore the relative scale of the variable observations as compared to other variables observations, and reducing the observations scale's influence on any further analysis. This transformation has significant impact on the clustering, particularly if the variables are measured on different scales. Thus, in the datasets used here, the BFI variable scale is between 0 to 100, the moisture variable scale is between 0 to -5, the thickness variable scale is between 0 to 5, and surface/profile variable scale between -2 to 2. Normalization adjusts the scales of the observation in such a way that all variable means are zero and the standard deviation one.

The normalization/standardization is applied to the variables in the prepared dataset and each observation uses the following equation for each variable

$$O_{N,j,i} = \frac{(O_{j,i} - \mu_j)}{\sigma_j} \qquad (6\text{-}5)$$

$O_{N,i}$ – normalized observation i of a variable j

O_i – observation i of a variable j

μ_j – the mean value of the variable j

σ_j – the Standard Deviation (SD) value of the variable j

The resulting normalized means and standard deviations are presented in Table 14

Table 14: Variable mean and SD values used for normalization

	BFI Center	BFI Right	BLT Center	BLT Right	abs(Right Prof 62)
Variables Mean	21.61	20.2	1.81	1.86	0.22
Variables SD	6.82	4	0.45	0.34	0.17

Table 15 and Table 16 present the summary statistics for these dataset variables before and after normalization. As can be seen in Table 15, the variables have different scaling and distribution.

Table 15: Variables statistical summary before normalization

BFI Center	BFI Right	BLT Center	BLT Right	abs(Right Prof 62)
Min.: 17.90	Min.: 17.90	Min.: 1.030	Min.:1.130	Min.: 0.03845
1st Qu:[15] 17.90	1st Qu:[9] 17.90	1st Qu:[9] 1.470	1st Qu[9]: 1.600	1st Qu:[9] 0.10925
Median: 17.90	Median: 17.90	Median: 1.800	Median: 1.800	Median: 0.16052
Mean: 21.61	Mean: 20.23	Mean: 1.811	Mean: 1.858	Mean: 0.22461
3rd Qu.: 25.00	3rd Qu.: 25.00	3rd Qu.: 2.10	3rd Qu.: 2.00	3rd Qu: 0.29816
Max.: 46.40	Max.: 39.30	Max.: 2.800	Max.: 2.600	Max.: 0.85022

In Table 16, which represents the statistical summary of the variables after normalization, the impact and influence of the normalization process can be clearly observed in the form of the scaling of the statistical values.

Table 16: Normalized variables statistical summary

BFI Center	BFI Right	BLT Center	BLT Right	abs(Right Prof 62)
Min.: -0.5442	Min.: -0.5829	Min.: -1.7445	Min.: -2.1518	Min.: -1.0742
1st Qu.:-0.5442	1st Qu.: -0.5829	1st Qu.: -0.7615	1st Qu.: -0.7627	1st Qu.: -0.6657
Median: -0.544	Median: -0.583	Median: -0.024	Median: -0.172	Median: -0.37
Mean: 0.00	Mean: 0.00	Mean: 0.00	Mean: 0.00	Mean: 0.00
3rd Qu.: 0.4975	3rd Qu.: 1.1936	3rd Qu.: 0.6461	3rd Qu.: 0.4195	3rd Qu.: 0.4244
Max.: 3.6375	Max.: 4.7717	Max.: 2.2100	Max.: 2.1928	Max.: 3.6099

The impact of the normalization can also be seen in the scatter plots presented in Figure 43 below. Note the first chart represents variables behavior before normalization and the bottom chart represents after normalization. The difference in the inter-variable behavior can clearly be seen in the change in the axes scale. In the before normalization, the scale of the variables (BFI, BLT, RProfil62) are all different, corresponding to the defined units, e.g., profile is defined in fractions of an inch. After the normalization, the scale is the same for all the variables.

[15] Quantile

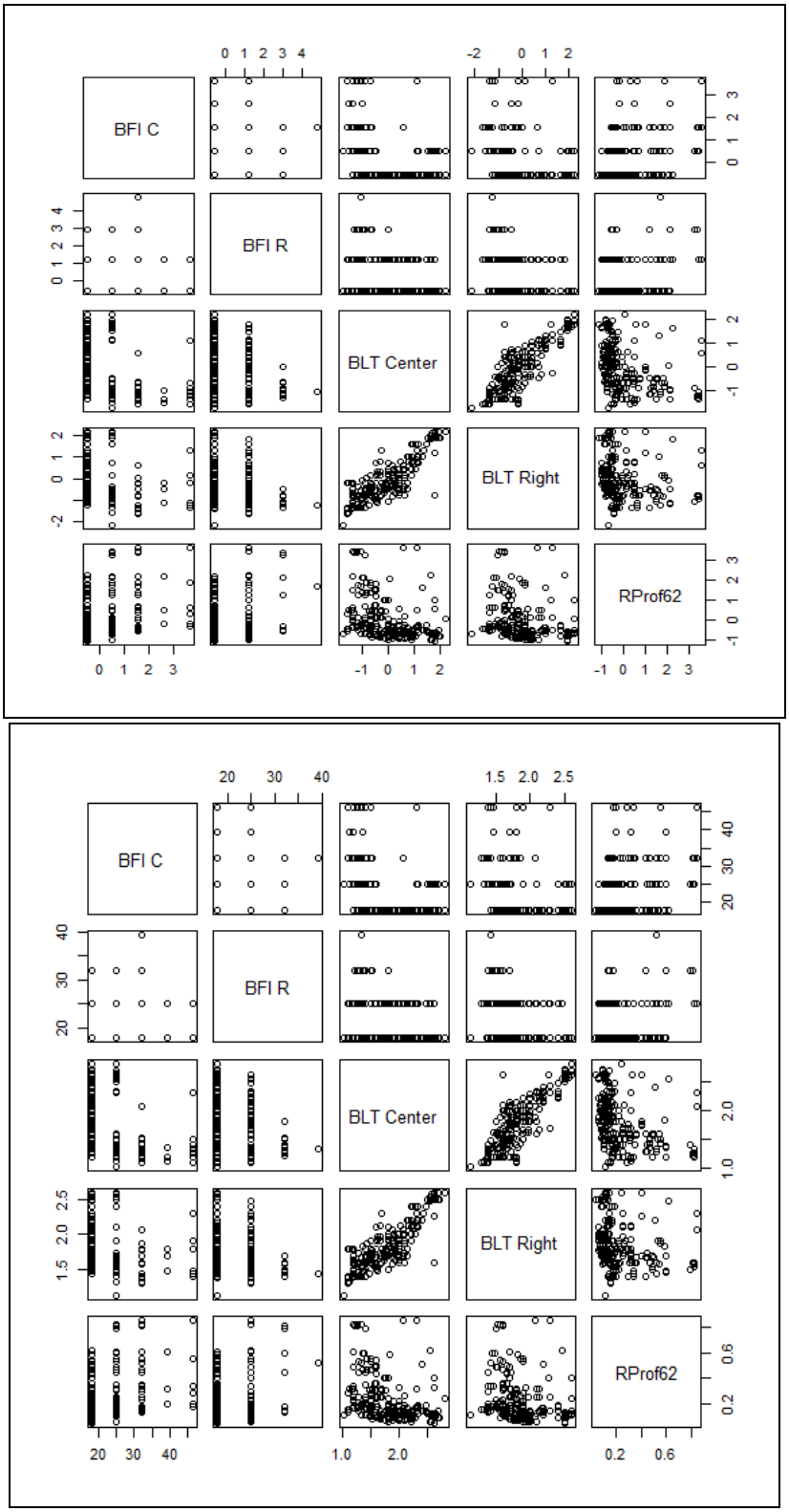


Figure 43: Inter-variable relationship chart before (bottom chart) and after (upper chart) normalization

6.1.4 Converting Classical Data to Symbolic Data

The next step in the analytical process was converting classical histogram-valued data (Figure 44) to symbolic data. This was accomplished using the 'data2hist' function of 'HistDAWass' package.[16] As presented below, each variable is valued as a histogram, so that real-valued data is aggregated by means of intervals and the corresponding distribution is not considered.

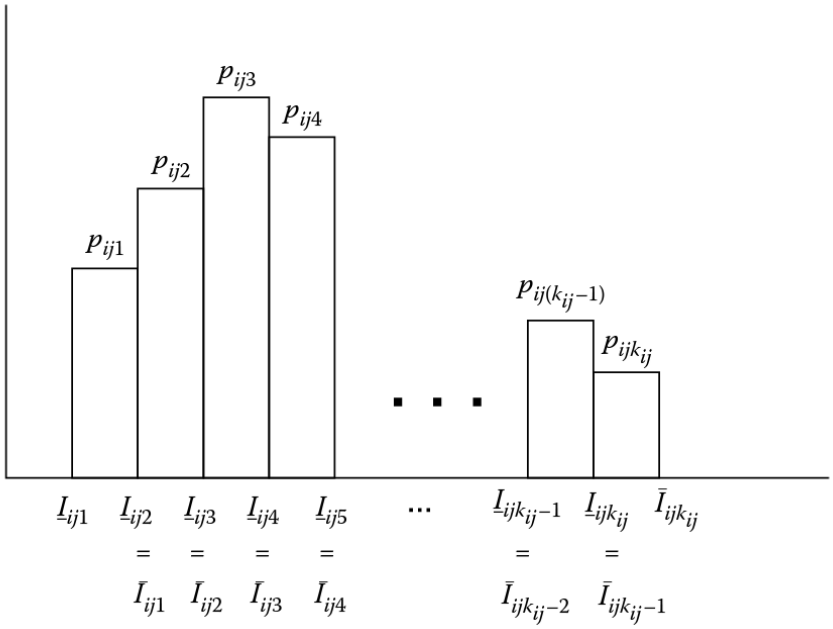


Figure 44: Histogram representation of real-valued data

The real-valued data is aggregated by means of intervals into n classes with the histogram-valued variable defined as follows:

$$S = (\{[\underline{I}_{j1}\overline{i_{i1}}], p_{i1}; \dots; [\underline{I}_{ik_i}, \overline{I_{ik_i}}], p_i k_i = 1, \dots, s_j\} j = 1, \dots, p) \quad (6\text{-}6)$$

Where,

S = histogram of random variable

p_{i1} = relative frequencies associated with the subinterval $[I_{j1}, i_{i1}]$

Thus, for the profile track geometry data, Rprof62, a histogram of data observation is presented in Figure 45 below. The corresponding summary statistics are presented in Table 17.

[16] 'HistDAWass' is an analysis package available in the R software

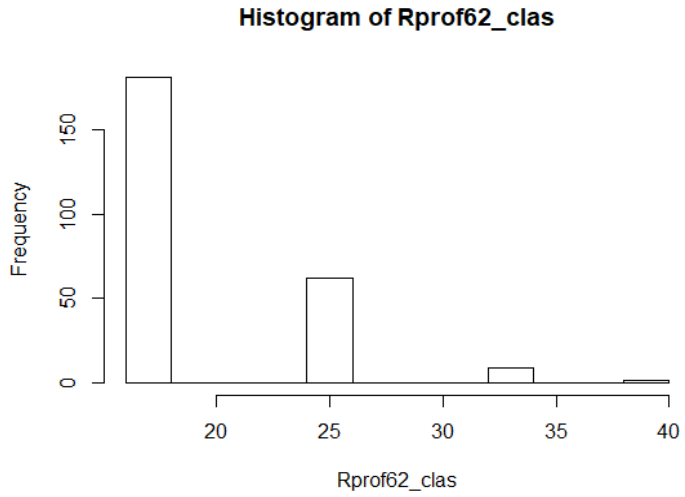


Figure 45: Histogram of real-valued variable absolute Rprof62

Table 17: Summary of real-valued variable absolute Rprof62

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
17.9	17.9	17.9	20.23	25	39.3

The results of converting absolute Rprof62 variable from classical data to histogram-valued data (Figure 46), as stored in the software database, are presented in the Table 18 below.

Table 18: Description of histogram-valued variable absolute Rprof62

	Bins intervals	Probability
Bin_1	[-1.0742--0.81397)	0.09881
Bin_2	[-0.81397--0.55374)	0.2253
Bin_3	[-0.55374--0.29351)	0.2569
Bin_4	[-0.29351--0.033283)	0.07905
Bin_5	[-0.033283-0.22695)	0.07115
Bin_14	[2.3088 ; 2.569)	3.95E-06
Bin_15	[2.569 ; 2.8292)	3.95E-06
Bin_16	[2.8292 ; 3.0895)	3.95E-06
Bin_17	[3.0895 ; 3.3497)	0.003953
Bin_18	[3.3497 ; 3.6099)	0.02767
Mean = 0.0043		
SD = 0.997		

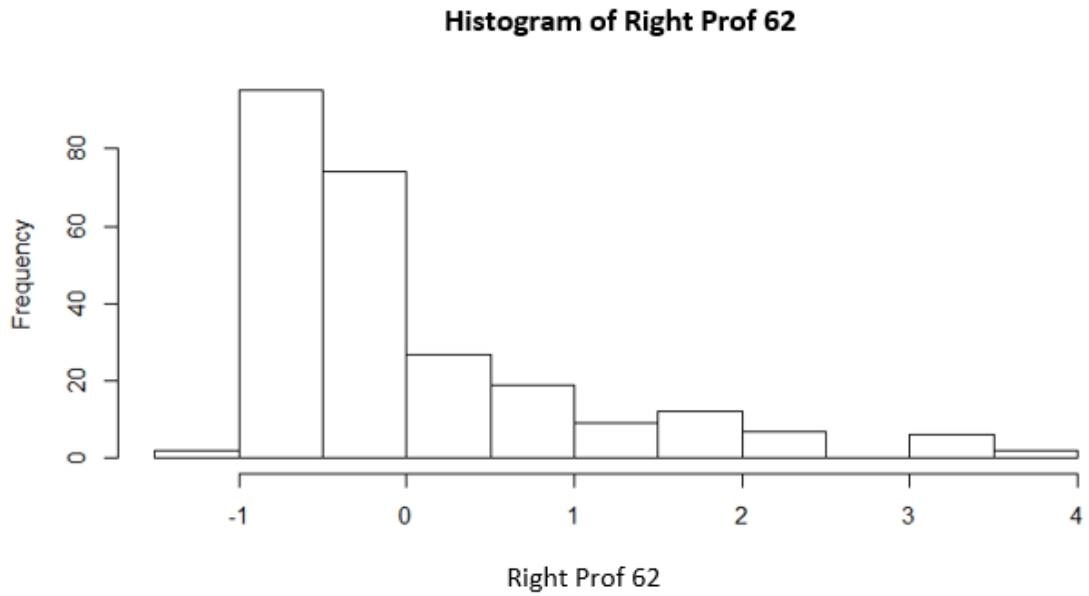


Figure 46: Histogram of histogram-valued variable absolute Rprof62

As noted above, the normalization process compensates for the relative scale of the variable observations and the reduction in observation's scale influence on the follow-up analysis. This is important for those cases where most of the data is of low value with a limited number of high value occurrences. This can be seen in Figure 46 where only a few high value profile exceptions are found but which represents key input data into the model analysis. This normalization adjusts the scales of the observation so that all variable means was zero and the standard deviation one. Using these histograms for the independent and dependent variables, a matrix of distributions was created, where all the parameters in the matrix cells are stored as histograms. Appendices C. 1 through C.3 present a more comprehensive set of this histogram data.

6.1.5 Creating a Matrix of Distributions by Rows

The last stage of data preparation for the hybrid analysis i.e., before using hierarchical clustering analysis of histogram-valued data, is the creation of a matrix of distributions with the model parameters. All parameters in the matrixes cells are stored as distributions and represented by the following information (see Table 19):

- Parameter name
- Mean
- Standard deviation

Table 19: Matrix of distributions description

Variable	Mean/SD
abs_RProf_62	[m= 0.0042916 ,s= 0.99696]
BFI Center	[m= -0.13211 ,s= 0.82406]
BFI Right	[m= -0.2243 ,s= 0.8209]
C_Layer	[m= 0.023813 ,s= 0.99858]
R_Layer	[m= -0.0015034 ,s= 0.99479]

Figure 47, Figure 48, and Figure 49 below are respectively: histogram, density approximation and box-plot presentations of the data for the key variables; track geometry (RProf_62), BFI Right and BFI Center, and BLT (R_Layer and C_Layer). Each plot is a comparative plot to describe graphically the parameters in the matrix of distributions.

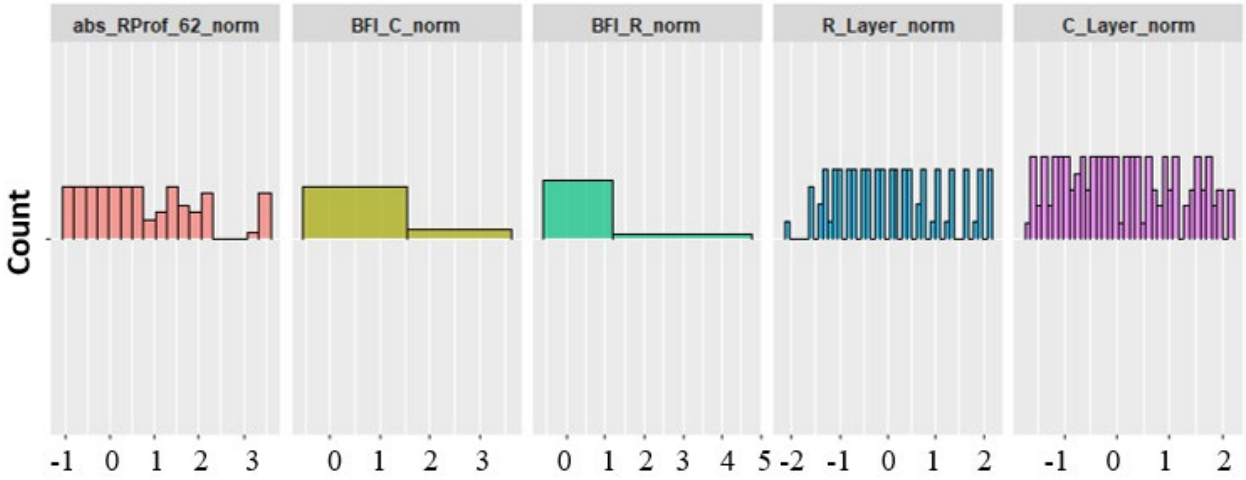


Figure 47: Comparative histogram plot of matrix of distributions

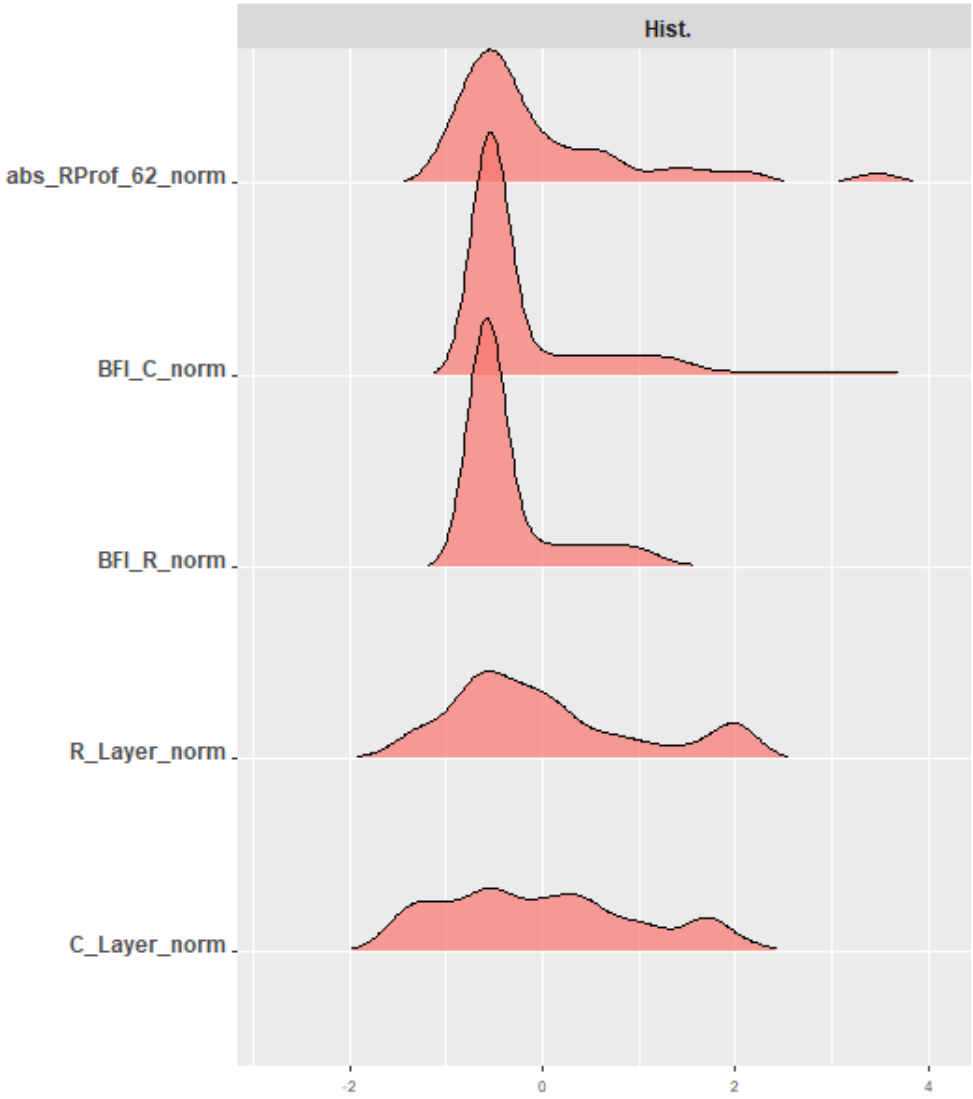


Figure 48: Comparative density approximation plot of matrix of histograms

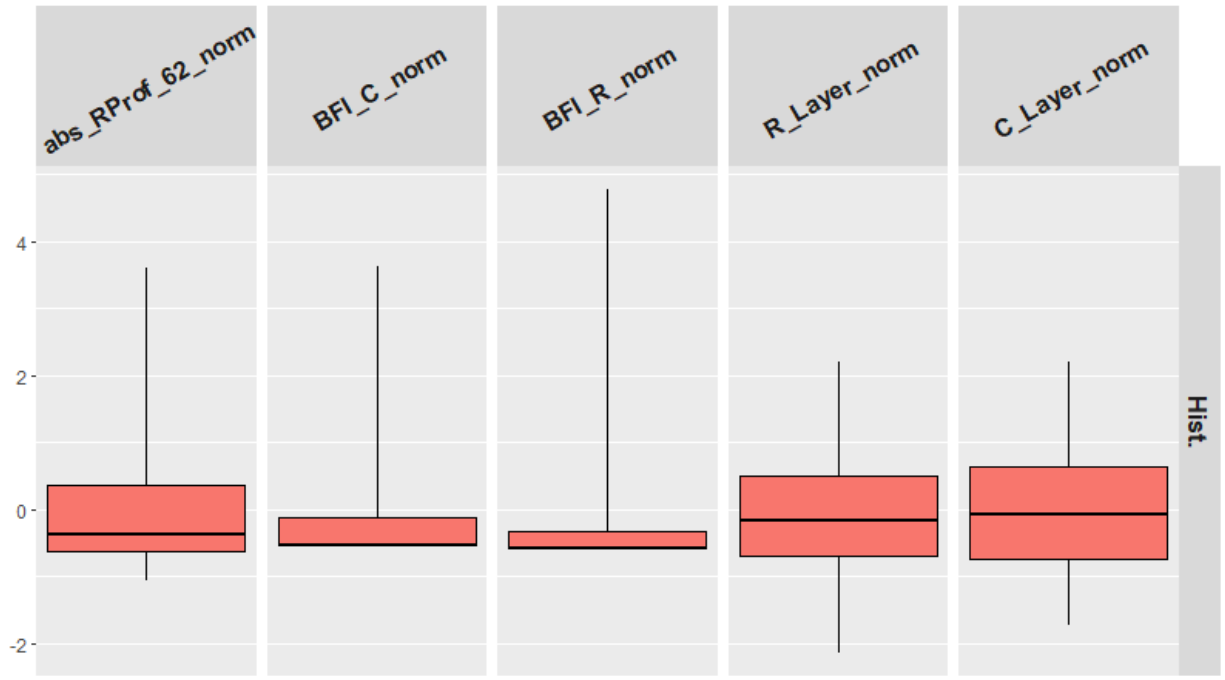


Figure 49: Comparative box-plot of matrix of histograms

Examination of these charts gives a useful picture of the variability, and distribution of the variables. For example, the following can be seen:

- BFI Center is much spread out than BFI Right, showing a wider range of variability in the data with high peak values.[17]
- There is no true symmetricity in the data.
- BLT has a broad distribution of data with a low peak value and two peaks.
- Absolute Rprofile62 is the most diverse variable.
- Most of the charts are skewed to the right, i.e., to the positive direction, as clearly seen in the box-plot and density approximation plot.

6.1.6 Hierarchical Clustering for a Set of Histogram-Valued Data Results and Interpretations

Once the matrix of histogram data was developed, a hierarchical clustering analysis of the histogram data was performed. As already noted, hierarchical clustering analysis of histogram-valued data is an unsupervised classification technique that combines both clustering analysis and symbolic data analysis [8] [9]. Both methods are well established and widely used in data analysis. The advantage of using their combination is it allows for classification of very large datasets.

The main objective of the clustering analysis is to group the dataset so that similar data objects are within one cluster. Clustering methods rely on a distance matrix where the d_{ij} value is the

[17] BFI Right has outlier data indicated by the large maximum

distance between pairs of observations [4]. The goal is to minimize the distances within clusters and maximize the distance between clusters. In this analysis, several clustering analyses were performed to include K-mean and adaptive K-mean clustering analyses [4] [8] [10] [11].

Hierarchical clustering organizes the data into hierarchical structures of partitions starting from singleton clusters (each data point is its own cluster) and progressing until one cluster covers the entire data [4]. Eight different linkage methods were used here, based on different approaches to the measurement distance, but all using L2 Wasserstein method for distance calculation between two probability distributions (ϕ_i, $\phi_{i'}$), as defined by the equation [9] [10] [11] [12]:

$$d_W(\phi_i, \phi_{i'}) = \sqrt{\int_0^1 [\Phi_i^{-1}(t) - \Phi_{i'}^{-1}(t)]^2 dt)} \quad (6\text{-}7)$$

Where, d_w = Wasserstein distance

Φ_i^{-1}, $\Phi_{i'}^{-1}$ = quantile functions

For each linkage method, a cluster dendrogram was prepared.[18] The linkage methods used included [4] [11] [12]:

- The complete linkage method, also called "farthest neighbor," is the proximity between two parameters, or clusters is the distance between their two most distant objects (see Figure 50 for a graphical example of the dendrogram).
- The average linkage method where the distance between two clusters is defined as the average of distances between all pairs of objects, where each pair is made up of one object from each group (see Figure 51 for a graphical example of the dendrogram).
- The single linkage method also referred to as "nearest neighbor" where the distance between two points is the minimum distance (see Figure 52 for a graphical example of the dendrogram).
- The Ward.D linkage method also called "Ward's method," or minimal increase of sum-of-squares (MISSQ) is the proximity between two clusters as the magnitude by which the summed square in their joint cluster was greater than the combined summed square in these two clusters (see Figure 53).
- The centroid linkage method where the distance between two clusters is the distance between the two mean vectors of the clusters. The proximity between two clusters is the proximity between their geometric centroids: [squared] Euclidean distance between those clusters (see Figure 54).
- The Ward.D2 linkage method uses squared Euclidean distances in Ward.D linkage (see Figure 55).
- The median linkage method uses Euclidean distance as the distance metric (see Figure 56).

[18] Eight cluster Dendrograms were developed, one for each link; however only one is presented in this paper to illustrate the Cluster Dendrogram concept.

- The McQuitty linkage method also called the equilibrium centroid method (WPGMC) measures the proximity between two clusters using their geometric centroids applying: squared Euclidean distance between those two clusters (see Figure 57).

6.1.7 Dendrogram for Different Linkages Methods

The result of the clustering analysis using different clusters methods are concentrated below by the linkage methods.

Figure 50 through Figure 57 below are a dendrogram representation of the hierarchical structure of the data. The "Y" axis stands for the distance between the parameters, or merged clusters, while “X” axis shows the parameters that are clustered. (Also, see Appendix C.4 for the details of the calculation of the clusters and dendrograms.)

Linkages Method = "complete"

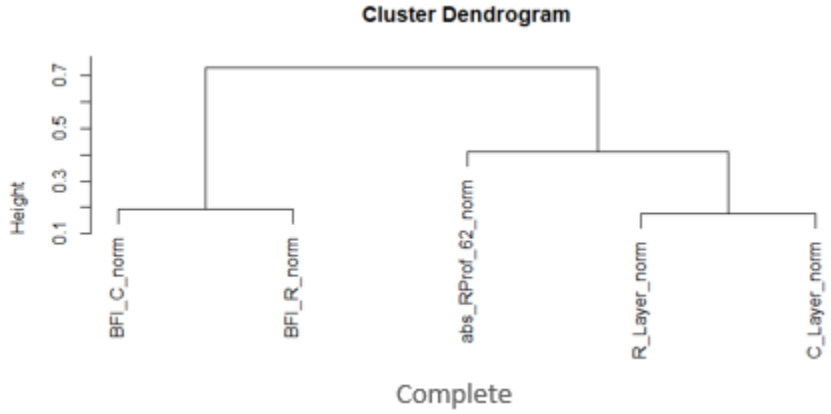


Figure 50: Cluster dendrogram with maximum dissimilarity (complete linkage)

Linkages Method = "average"

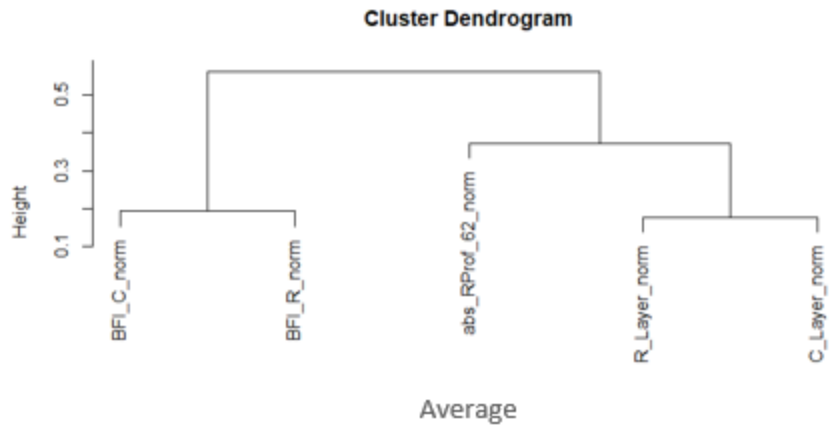


Figure 51: Cluster dendrogram with average dissimilarity (average linkage)

Linkages Method = "single"

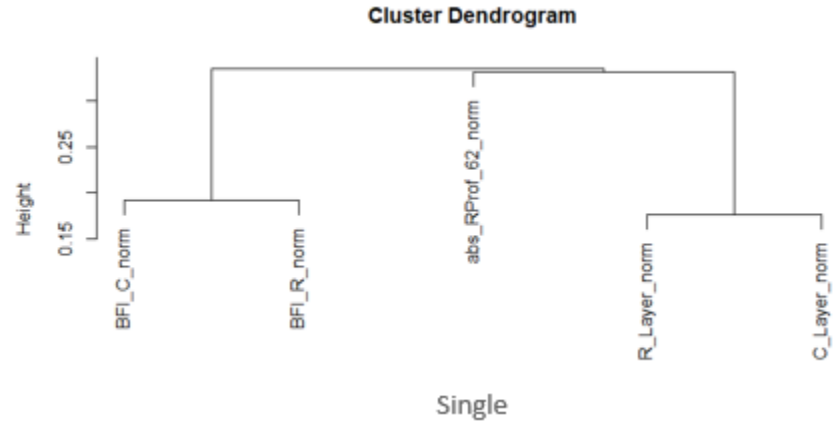


Figure 52: Cluster dendrogram with minimum dissimilarity (single linkage)

Linkages Method = "Ward.D"

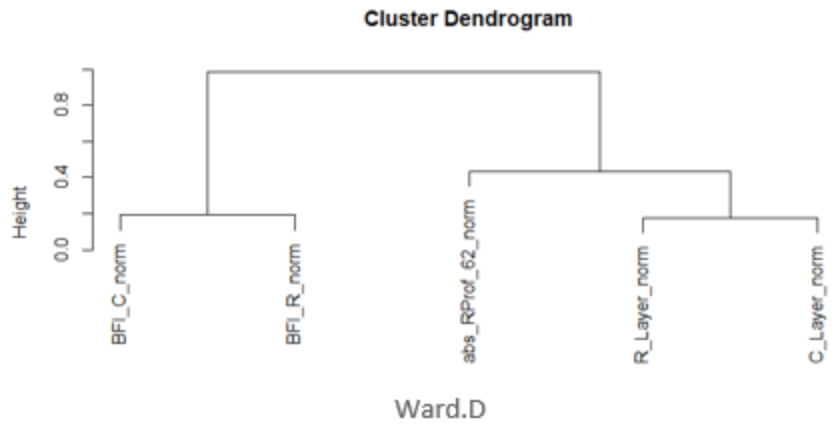


Figure 53: Cluster dendrogram with Ward method

Linkages method = "centroid"

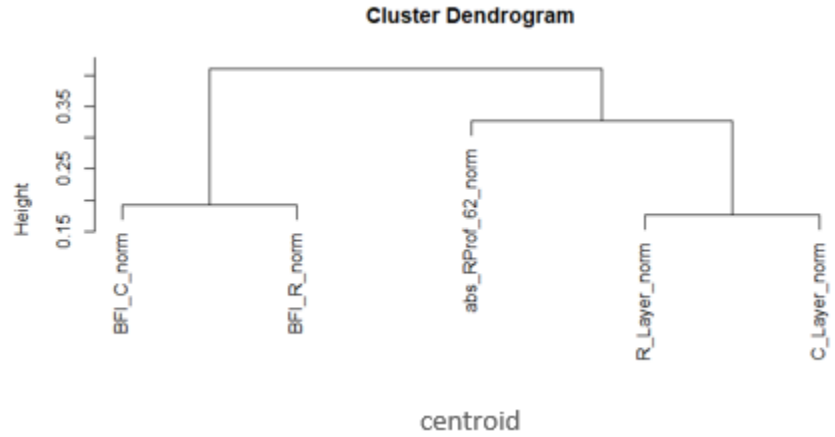


Figure 54: Cluster dendrogram with centroid method

Linkages Method = "Ward.D2"

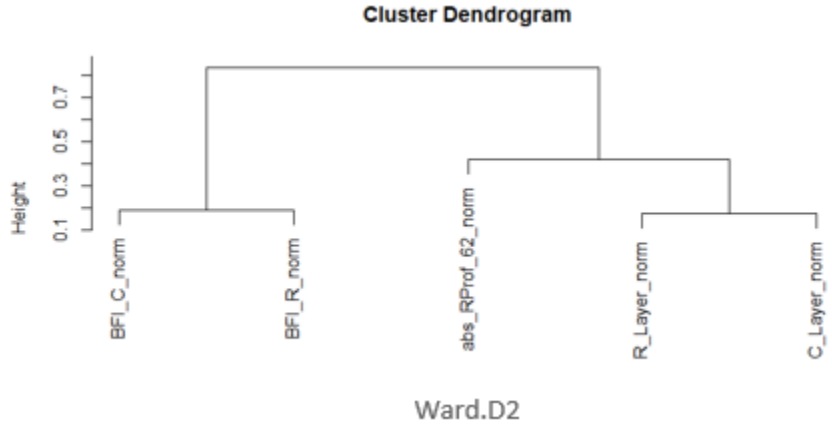


Figure 55: Cluster dendrogram with Ward.D2 method

Linkages Method = "median"

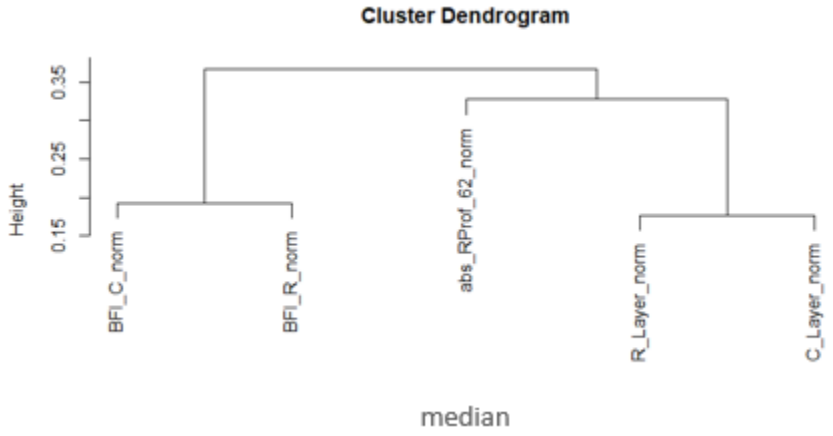


Figure 56: Cluster dendrogram with median method

Linkages method = "McQuitty"

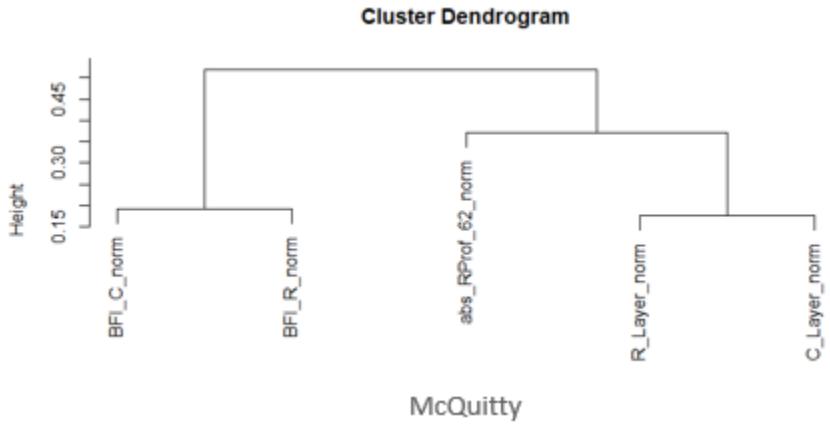


Figure 57: Cluster dendrogram with McQuitty method

Note the difference in the height scale for the eight different dendrograms.

The resulting dendrograms were then analyzed using the "cut the tree" method which is a function that "cuts" the dendrogram, i.e., the "tree," of hierarchical clustering, for a set of histogram-valued data based on the L2 Wasserstein distance, to several groups either by the desired number of groups or the desired cut height. Due to having five parameters, cut the tree generates four groups, which is the optimal number of depended variables in the model equation. Appendix C.5 shows the details of the cut the tree approach used here.

The results for the clustering analysis for the eight different linkage types are presented in Table 20 below:

Table 20: Summary of cut the tree for four groups

	HCA for a HD data based on the L2 Wasserstein distance for 4 clusters							
	Linkage methods							
	complete	average	single	ward.D	centroid	ward.D2	median	McQuitty
abs_RProf_62	1	1	1	1	1	1	1	1
BFI_Center	2	2	2	2	2	2	2	2
BFI_Right	3	3	3	3	3	3	3	3
BLT Center	4	4	4	4	4	4	4	4
BLT Right	4	4	4	4	4	4	4	4

Table 20 shows, for each of the eight linkage methods, the clustered parameters belonging to groups after the hierarchical dendrogram is "cut" creates four clusters. The results in the table suggest that in any linkage method BLT Center and BLT Right are grouped together, while the rest of the parameters are identified as independent clusters. These results suggest that modeling the relationship between the parameters BLT Right, or BLT Center separately should correlate

with other parameters. Based on this, the LR models (see below) measure the two main parameter combinations as follows:

- $f_i(\text{abs\_RProf\_62}) = [BFI\ C, BFI\ R, BLT\ C]$
- $f_j(\text{abs\_RProf\_62}) = [BFI\ C, BFI\ R, BLT\ R]$

6.2 Logistic Regression Analysis with Higher Order Polynomials

Based on the results of the hierarchical clustering analysis with histogram data, the key LR parameters were identified and incorporated into four different LR models. These parameters were:

- Right Profile 62 (dependent variable)[19]
- BFI Center
- BFI Right
- BLT Center
- BTL Right (not used in all LR models)

However, unlike the original analysis presented previously [15] higher order polynomial combinations of these variables were included in the LR analysis. Application of higher order polynomials was used here to create a more accurate model given the variables to approximate the complex nonlinear relationship between the variables [12] [13]. The purpose of the application of the higher order polynomial parameters is the improvement of the model performance by improving the function approximation. The order of the polynomial model used here is that of first and second order polynomials e.g. x, x^2, y, y^2, xy, etc. Table 21A, Table 21B, and Table 21C define the variables and the associated higher order polynomials used in the LR analysis. These tables are a map to the variables in the LR model for ease of display.

Table 21A: Parameter definitions

Parameter	Description
Rprof 62	Right Profile 62 (dependent variable-binary
BFI C	Ballast Fouling Index Center
(BFI C)^2	BFI C Squared
BFI R	Ballast Fouling Index Right
(BFI R)^2	BFI R Squared
BLT C	Ballast Layer Thickness Center
(BLT C)^2	BLT C Squared
BLT R	Ballast Layer Thickness Right
(BLT R)^2	BLT R Squared

[19] As noted previous, the output is presented as a binary value equal to 1 if the absolute value of Right Profile 62 is greater than 10 mm [0.4 inch], otherwise 0.

Table 21B: Parameters representation 1

parameter	BFI C	BFI R	(BFI C)^2	(BFI R)^2	BLT C	BLT R
representation	C	R	C2	R2	LC	LR

Table 21C: Parameters representation 2

parameter	(BLT L)^2	(BLT R)^2	(BFI C)^2	(BFI R)^2	BLT C	BLT R
representation	C	R	C2	R2	LC	LR

As noted, four LR models were developed with the following formats (per Table 21A, Table 21B, and Table 21C) and are schematically presented below (probability of a Right Profile 62 exceedance is a function of the associated independent variable per the above table):

- $f_1(\text{abs Right Profile 62}) = [C, R, C^2, R^2, LC^2, C \cdot LC, R \cdot LC]$
- $f_2(\text{abs Right Profile 62}) = [C, R, C^2, R^2, LR, LR^2, C \cdot LR, R \cdot LR]$
- $f_3(\text{abs Right Profile 62}) = [C, R, LC]$
- $f_4(\text{abs Right Profile 62}) = [C, R, LR]$

The resulting final models are the following equations, where the probability of exceedance is defined as a function various independent variables and associated weighting factors:

$$\hat{P}_{1,\,\text{defect}} = \frac{e^{1.72+0.02\cdot BFI_{center}+0.11\cdot BFI_{right}-8.17\cdot BLT_{center}-0.005\cdot (BFI_{center})^2+0.001\cdot (BFI_{right})^2+0.69\cdot (BLT_{center})^2+0.2\cdot (BFI_{center}\cdot BLT_{center})+0.013\cdot (BFI_{right}\cdot BLT_{center})}}{1+e^{1.72+0.02\cdot BFI_{center}+0.11\cdot BFI_{right}-8.17\cdot BLT_{center}-0.005\cdot (BFI_{center})^2+0.001\cdot (BFI_{right})^2+0.69\cdot (BLT_{center})^2+0.2\cdot (BFI_{center}\cdot BLT_{center})+0.013\cdot (BFI_{right}\cdot BLT_{center})}} \quad (6\text{-}8)$$

$$\hat{P}_{2,\,\text{defect}} = \frac{e^{-4+0.33\cdot BFI_{center}+0.57\cdot BFI_{right}-0.002\cdot BLT_{right}-0.01\cdot (BFI_{center})^2+0.007\cdot (BFI_{right})^2-2.53\cdot (BLT_{right})^2+0.2\cdot (BFI_{center}\cdot BLT_{right})+0.25\cdot (BFI_{right}\cdot BLT_{right})}}{1+e^{-4+0.33\cdot BFI_{center}+0.57\cdot BFI_{right}-0.002\cdot BLT_{right}-0.01\cdot (BFI_{center})^2+0.007\cdot (BFI_{right})^2-2.53\cdot (BLT_{right})^2+0.2\cdot (BFI_{center}\cdot BLT_{right})+0.25\cdot (BFI_{right}\cdot BLT_{right})}} \quad (6\text{-}9)$$

$$\hat{P}_{3,\,\text{defect}} = \frac{e^{-7.36+0.06\cdot BFI_{center}+0.19\cdot BFI_{right}+0.05\cdot BLT_{right}}}{1+e^{-7.36+0.06\cdot BFI_{center}+0.19\cdot BFI_{right}+0.05\cdot BLT_{right}}} \quad (6\text{-}10)$$

$$\hat{P}_{4,\,\text{defect}} = \frac{e^{-4.98+0.04\cdot BFI_{center}+0.18\cdot BFI_{right}-0.92\cdot BLT_{center}}}{1+e^{-4.98+0.04\cdot BFI_{center}+0.18\cdot BFI_{right}-0.92\cdot BLT_{center}}} \quad (6\text{-}11)$$

Note, Models 1 and 2 were more complex with the higher order polynomial effect. Models 3 and 4 were first order, like the second-generation model presented previously [15]. Appendix C.6 presents additional details of the LR models.

6.3 Statistical Validation

To compare the four models as well as to statistically validate the models, an error matrix or "confusion matrix" approach was used [6]. The confusion matrix validation process examines the number of correct and incorrect predictions based on the reference and the corresponding performance of the model. Table 22 summarizes the models' statistics, validation and performance:

Table 22: Hybrid LR models' results comparison

						Confusion Matrix and Statistics				
							Ref.			
	AIC	Null Deviance	Residual deviance	Accuracy	Kappa	***Predict***	***0***	***1***	Accuracy	AUC
Model 1	176.7	207.01	158.73	0.873	0.24	***0***	215	27	0.885	0.7951
						1	2	9		
Model 2	173.8	207.01	155.88	0.877	0.28	***0***	213	24	0.889	0.8154
						1	4	12		
Model 3	180.6	207.0	172.61	0.862	0.2	***0***	213	27	0.877	0.7537
						1	4	9		
Model 4	177.7	207.01	169.71	0.866	0.25	***0***	213	29	0.869	0.7799
						1	4	7		

Examination of the statistical performance of the four models showed that the higher order models (Models 1 and 2) have greater accuracy than the lower order polynomial models (Models 3 and 4). That is because of the non-linear relationships between the input parameters and "cross-influence" it generates.

In comparing Models 1 and 2, while Model 2 has a slightly higher accuracy, Model 1 has a better prediction behavior as shown in the confusion matrix statistics. Thus, it has more—Predictive Positive (Pred 0)—Actual Positive (Ref 0) "hits" (215 vs 213) and a better miss rate with fewer false negatives (Pred 1 – Ref 0) with 2 false negatives as opposed to 4 for Model 2 (miss rate of 0.9% vs 1.8%).

Comparing model 1, developed here, with the previously developed second-generation LR model [15], in the confusion matrix of Table 23, again shows that Model 1 has a measurably higher accuracy, higher Predictive Positive (Pred 0)—Actual Positive (Ref 0) "hits" (215 vs 213) and a better miss rate.

Table 23: Comparison of hybrid Model 1 to second-generation LR model

	Model performance			cross validation		Confusion Matrix and Statistics				ROC
							Ref.			
Model identifier	AIC	Null Deviance	Residual deviance	Accuracy	Kappa	*Pred.*	*0*	*1*	Accuracy	AUC
previous Model	177.71	207.01	169.71	0.874	0.27	*0*	213	29	0.869	0.7799
						1	4	7		
Model 1	176.7	207.01	155.73	0.873	0.24	*0*	215	27	0.885	0.7951
						1	2	9		

See Appendix C.7, Appendix C.8, and Appendix C.9 for additional details.

6.4 Sensitivity Analysis

Noting that model 1 is the most statistically accurate of the models, the sensitivity of Model 1 prediction of the probability of having a track geometry profile defect to the three independent variables (BFI-C, BFI-R, and BLT-C)[20] can be examined. Noting that the model has three independent variables, there are six permutations when presented as two-dimensional sensitivity graphs and three permutations when presented as a three-dimensional graph. To effectively display these sensitivities, one independent variable was presented as a continuous function, a second as a set of three values (Minimum, Average, and Maximum) and the third held constant. Using this approach, it is possible to observe the influence of each parameter in predicting the probability of having a profile defect (P(abs(Rprof62)>10 mm [0.4 inches]). The results are illustrated in two and three-dimensions charts as follows. (see Appendix C.10 for additional plots.)

Figure 58 show the probability of having a profile defect (P(abs(Rprof62)>10 mm [0.4 inches]) as a function of BFI Right (BFI R), for three cases:

- Ballast Layer Thickness (BLT) = Minimum
- Ballast Layer Thickness (BLT) = Average
- Ballast Layer Thickness (BLT) = Maximum
- Note, BFI Center (BFI C) is held constant at its average value for this graph

As can be seen from this graph, the probability of having a defect is very sensitive to BFI R, with increasing ballast fouling (higher BFI value) causing a greater probability of having a defect. Note, the inverse sensitivity to ballast thickness, with decreasing ballast thickness increasing the probability of having a defect, as expected from engineering experience.

[20] BLT-R drops out in Model 1

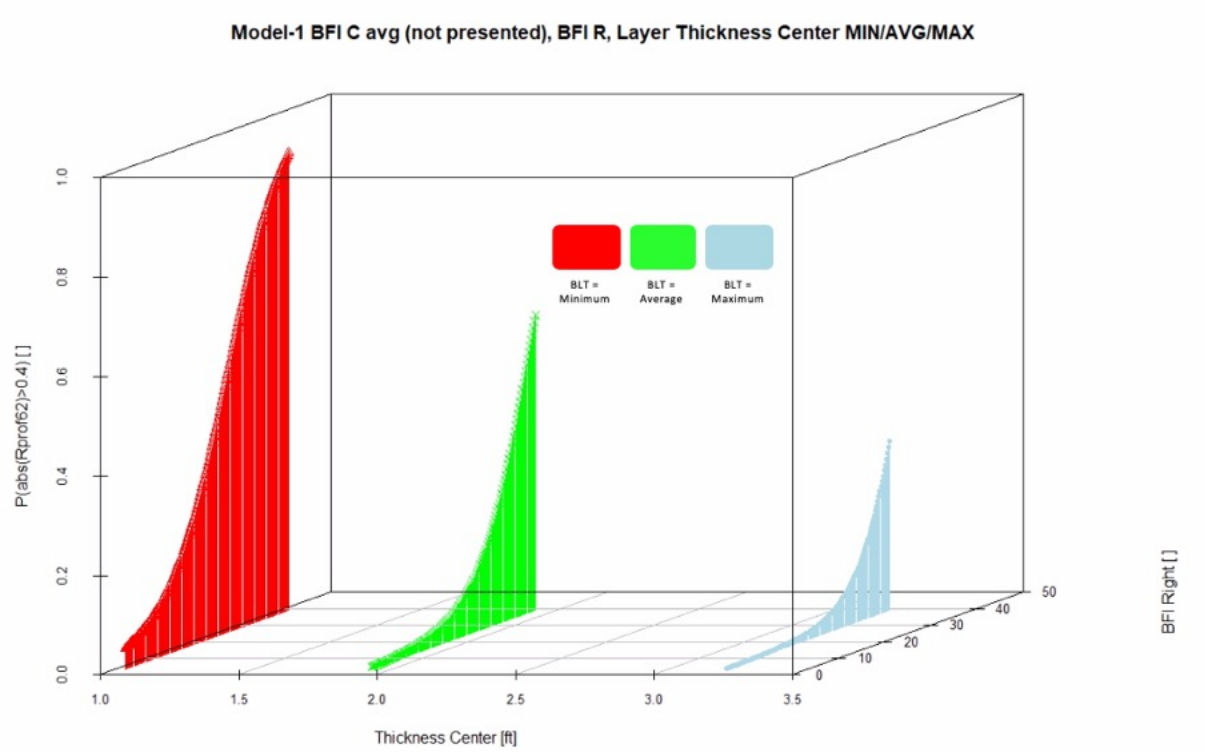


Figure 58: Probability of geometry defect as a function of ballast fouling index (BFI-Right) and ballast layer thickness (BFI Center held constant)

Figure 59A and Figure 59B show the probability of having a profile defect (P(abs(Rprof62)>10 mm [0.4 in]) as a function of BLT Center for three cases:

- Ballast Fouling Index-Right (BFI-R) = Minimum
- Ballast Fouling Index-Right (BFI-R) = Average
- Ballast Fouling Index-Right (BFI-R) = Maximum
- Note; BFI Center (BFI C) is held constant at its average value for this graph

As in the case with Figure 58, the probability of having a defect is very sensitive to BFI R, with increasing ballast fouling (higher BFI value) causing a greater probability of having a defect. Again, note, the inverse sensitivity to ballast thickness, with decreasing ballast thickness increasing the probability of having a defect, as expected from engineering experience.

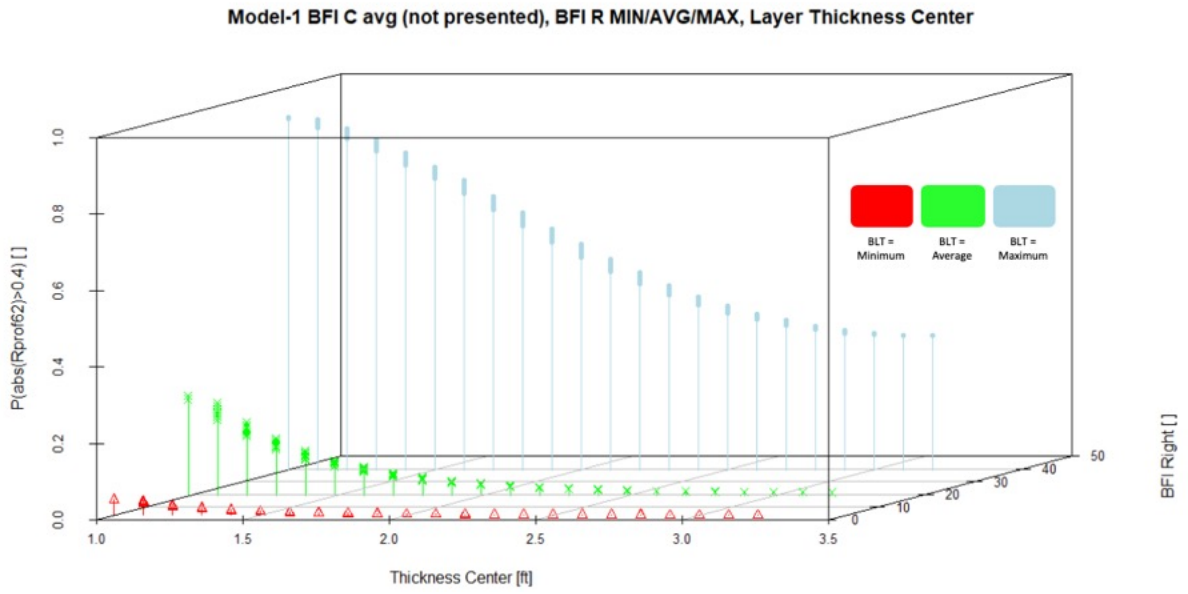


Figure 59A: Probability of geometry defect as a function of BFI Right and BLT (BFI-Center held constant)

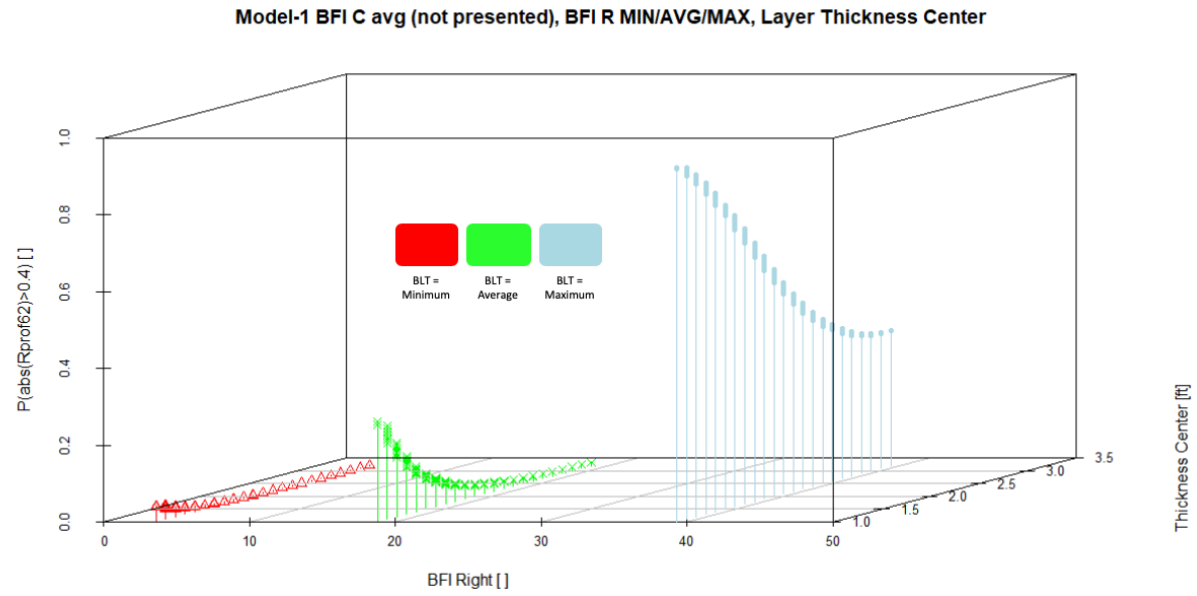


Figure 59B: Probability of geometry defect as a function of BFI Right and BLT (BFI-Center held constant)–axes reversed

As can be seen from the above figures, BLT has significant importance and influence on the likelihood of having a profile defect, the thicker the ballast layer in the center of the track, the lower the probability of having a profile defect. BFI Right also has a strong influence on the profile defect likelihood, but it is significantly lower when the ballast layer in the center of the track is thick.

Figure 60 presents a three-dimensional plot of the probability of a track geometry profile defect as a function of BFI Right and BLT.[21] Thus, the maximum probability of a geometry (profile) defect occurs when the ballast is fouled (high BFI) and there is a thin ballast layer (low BLT). This probability is of the order of 92 percent. Furthermore, as can be seen in this figure, as ballast fouling decreases (lower BFI) the probability of a profile defect decreases. Likewise, as ballast thickness increases (high BLT), the probability of a profile defect decreases. However, sensitivity to BFI-Right appears to be greater than that to BLT (this can be seen more clearly in Figure 59A and Figure 59B).

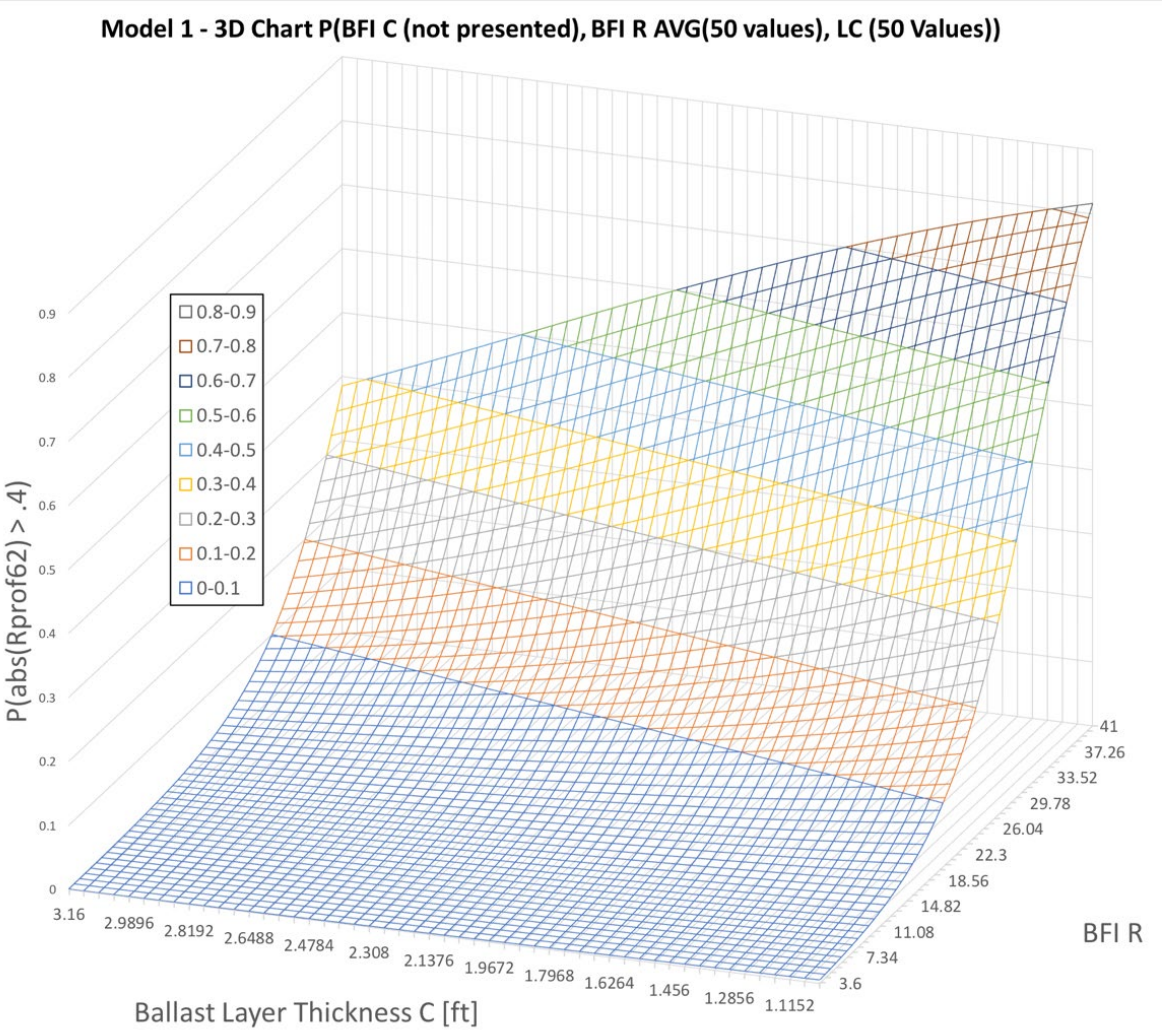


Figure 60: Three-dimensional plot of probability of profile defect as a function of BFI Right and BLT (BFI Center held constant)

[21] Note the BLT axis is reversed, going from maximum to minimum (decreasing magnitude). This allows for a better visualization of the three-dimensional model.

Figure 61A and Figure 61B show the probability of having a profile defect (P(abs(Rprof62)>10 mm [0.4 in]) as a function of the two BFI parameters, holding BLT constant. In Figure 61A, BFI Right (BFI R), is continuous and BFI Center is presented for three values (Minimum, Average, and Maximum). In Figure 61B, BFI Center is continuous and BFI right is given as three values. Note: BLT Center is held constant in both cases at its average value for these figures.

As can be seen from this graph, the probability of having a defect is sensitive to both BFI Right and BFI Center, however, the sensitivity to BFI Center is not as great as that observed for BFI Right.

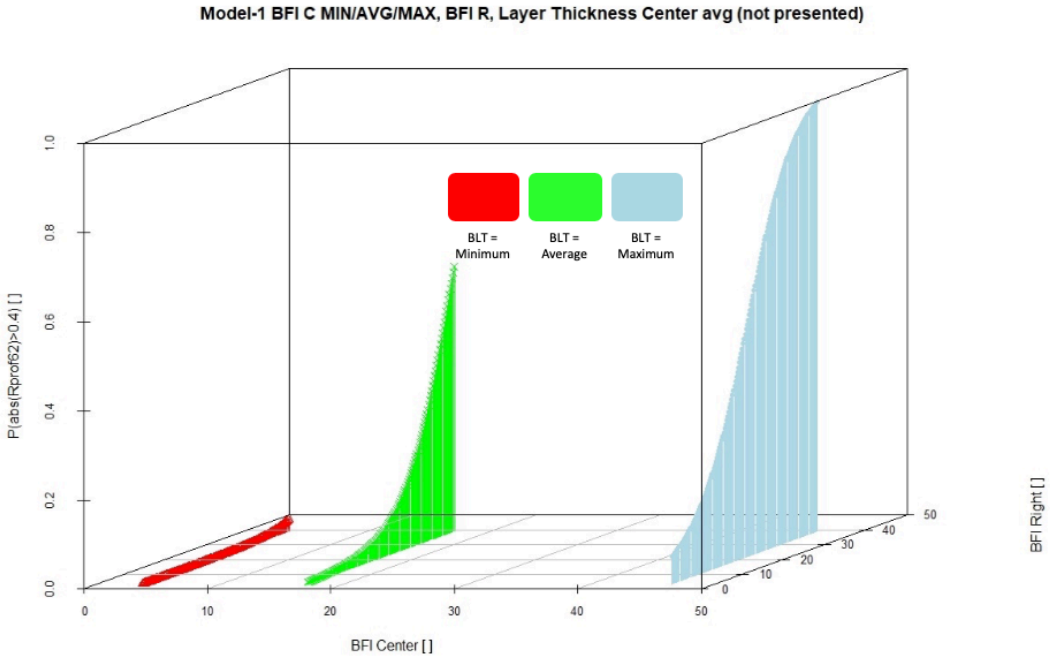


Figure 61A: Probability of profile defect as function of BFI Right and BFI Center (BLT constant)

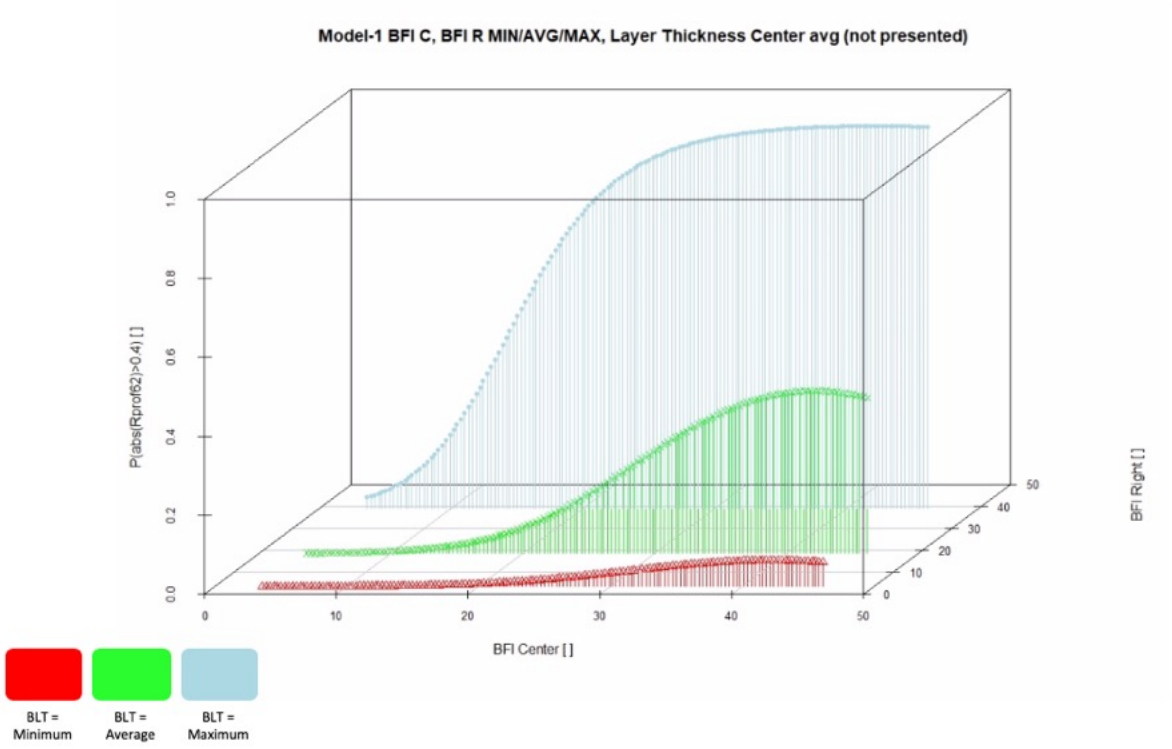


Figure 61B: Probability of profile defect as function of BFI Right and BFI Center (BLT constant)

Finally, in Figure 62 the sensitivity of BFI Center and BLT is shown, with BFI Center presented as a continuous value, and BLT presented as discrete values. Note, the behavior for a thick ballast section. It appears that a thin ballast layer does not have consistent expected behavior with respect to BFI Center. It appears that the BFI Center value introduces some sensitivities that are not consistent with expected engineering behavior.

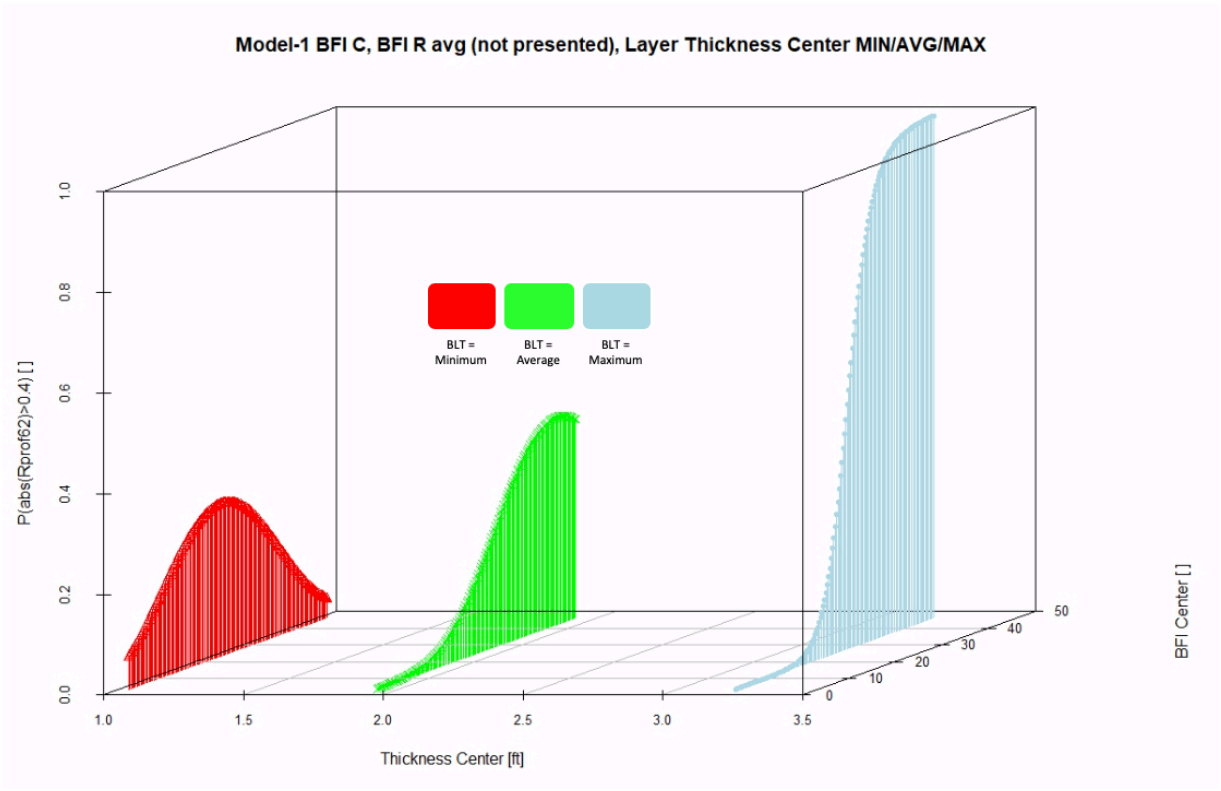


Figure 62: Probability of profile defect as function of BLT and BFI Center (BFI Right constant)

Summarizing the results of the sensitivity analysis as the following:

- BFI Right has the highest influence on the probability for having a profile defect with increasing fouling condition resulting in an increased probability of having a defect develop at that location.
- BLT has a significant influence on the probability for having a profile defect with decreasing BLT resulting in an increased probability of having a defect develop at that location.
- The two highest probabilities for having a profile defect are from the following combinations:
    - High BFI Right and BFI Center, the probability of having a profile defect is 96.5 percent. (BLT equals its average value of 584 mm [23 inches])
    - High BFI Right and low BLT, the probability of having a profile defect is 91.8 percent. (BFI Center equals its average value of 17)

7. Conclusion

This report presented the results of an FRA sponsored study on the relationship between track geometry defects and track subsurface conditions as measured by GPR. The analysis made use of multiple track geometry runs and the associated track geometry degradation behavior combined with data, specifically BFI and BLT. Correlation and statistical analyses were performed looking at the relationship between probability of significant geometry degradation (i.e., the development of a profile defect) and measured GPR parameters (e.g., BFI, BLT)

A LR model was used to develop the relationship between the probability of having a geometry (profile) defect, defined in the analysis as having a profile value greater than 0.4 inches, and the GPR based independent variables: BFI Right and Center and BLT. The resulting model showed that both BFI Right and BLT had significant influence on the likelihood of having a profile defect. In the case of BLT, the thicker the layer of the ballast in the center of the track the lower the probability of having a profile defect. In the case of BFI Right, the greater the ballast fouling (high BFI value) the greater the probability of having a defect. However, while BFI had a strong influence on the probability of developing a profile defect it was significantly lower when the ballast layer in the center of the track was thick. Furthermore, when the BFI index is low, the probability of having a defect was low even with thin ballast layers. Overall, BFI Right had the highest influence on the probability for having a profile defect (on the right rail), with BLT also having a significant influence. BFI Center, likewise effects the probability of having a profile defect, but not as significant as BFI Right.

Figure 63 presents these results in an alternative way; by showing lines of constant "probability" of developing a defect. Thus, for example, looking at the dark blue line at the right of the graph, which corresponds to a 90 percent defect probability, Figure 63 presents the combination of BFI Right and BLT Center that will give a 90 percent probability of developing a profile defect in the right rail (BFI Center is held at a constant at the "average" value). Any combination to the right of that dark blue (90%) limit line will have a greater than 90 percent probability of having a profile exceedance. Conversely, for the grey line at the left of the graph, which corresponds to a 10 percent defect probability, Figure 63 presents the combination of BFI Right and BLT Center that will give a 10 percent probability of developing a profile defect in the right rail. Any combination to the left of that grey (10%) line will have a less than 10 percent probability of having a profile exceedance. Similar lines are presented for 25, 50, and 75 percent levels of probability of having a profile exceedance.

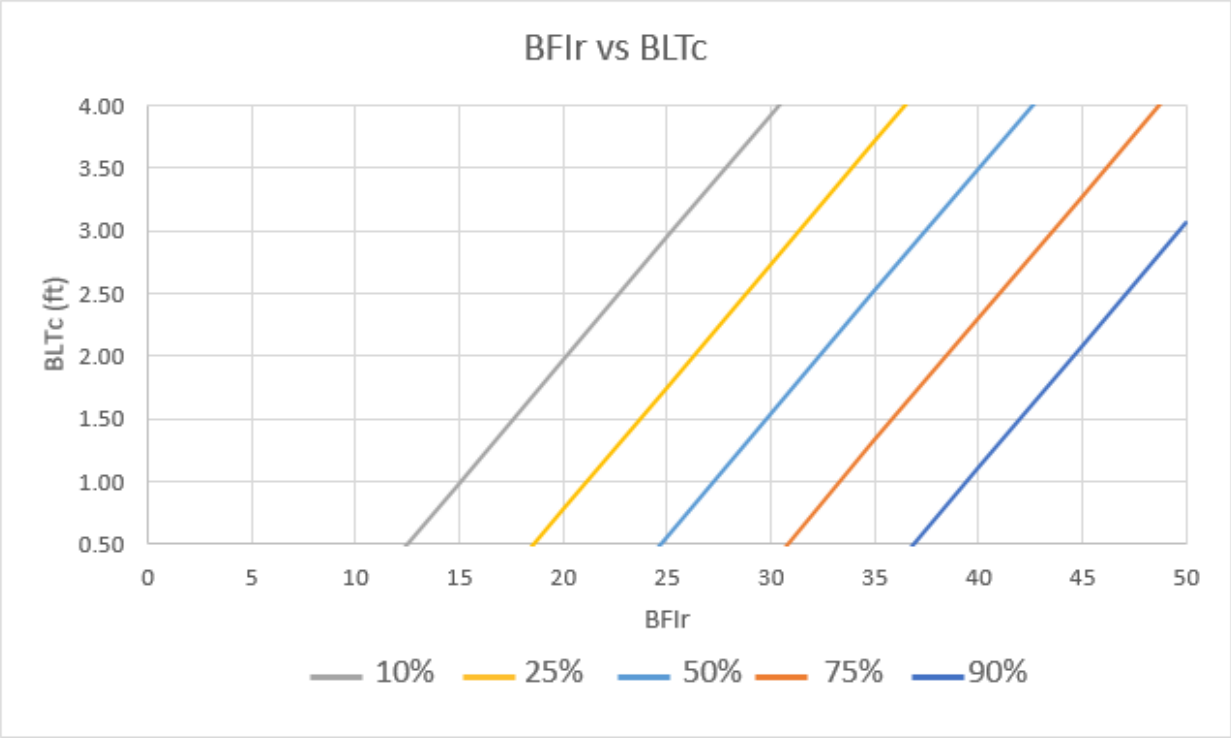


Figure 63: Probability of a profile defect as a function of BFI/BLT combinations

Figure 63 can be particularly useful to track engineers who know the condition of their BFI and BLT so they can ascertain the risk of developing a profile exceedance. In addition, trade-off analyses can be performed to look at what is the benefit of cleaning the ballast or increasing the depth of the ballast layer.

Building upon the initial LR model, a higher order data analytics approach using combinational hybrid analysis to include hierarchical clustering analysis with histogram data, LR analysis, and application of higher degree polynomials was performed. The result was a higher order polynomial LR model for determination of the probability of a rail profile defect occurring at locations with measured ballast fouling (as defined by the BFI) and measured BLT (as measured by GPR as the BLT).

The resulting higher order LR model developed for the relationship between the probability of having a track geometry (profile) defect, defined in the analysis as having a measured value greater than 10 mm [0.4 inches], and the GPR based independent variables: BFI Right and BFI Center and BLT. The resulting model showed that both BFI Right and BLT had significant influence on the likelihood of having a profile defect. In the case of BLT, the thicker the layer of the ballast in the center of the track the lower the probability of having a profile defect. In the case of BFI Right, the greater the ballast fouling (high BFI value) the greater the probability of having a defect. However, while BFI had a strong influence on the probability of developing a profile defect it was significantly lower when the ballast layer in the center of the track was thick. Furthermore, when the BFI is low, the probability of having a defect was low even with thin ballast layers. Overall, BFI Right had the highest influence on the probability for having a profile defect (on the right rail), with BLT also having a significant influence. BFI Center,

likewise effected the probability of having a profile defect, but introduced some engineering inconsistencies, particularly at low ballast thickness values.

Figure 64A and Figure 64B present the models sensitivities in a manner that can be utilized in practice; by showing lines of constant "probability" of developing a defect. Thus, for example, in Figure 64A looking at the blue zone at the bottom-right of the graph, this corresponds to an 80 percent to 100 percent probability that a profile defect greater than 10 mm [0.4 inches] will develop at a location with a BFI Right-BLT combination as shown in this figure. Similarly, the yellow zone in Figure 64A corresponds to a 60 percent to 80 percent probability that a profile defect greater than 0.4 inches will develop at a location with a BFI Right–BLT combination as shown. Similar zones are presented for 0–20 percent, 20–40 percent, and 40 to 60 percent levels of probability of having a profile defect. Figure 64B presents the same date with the axes reversed.

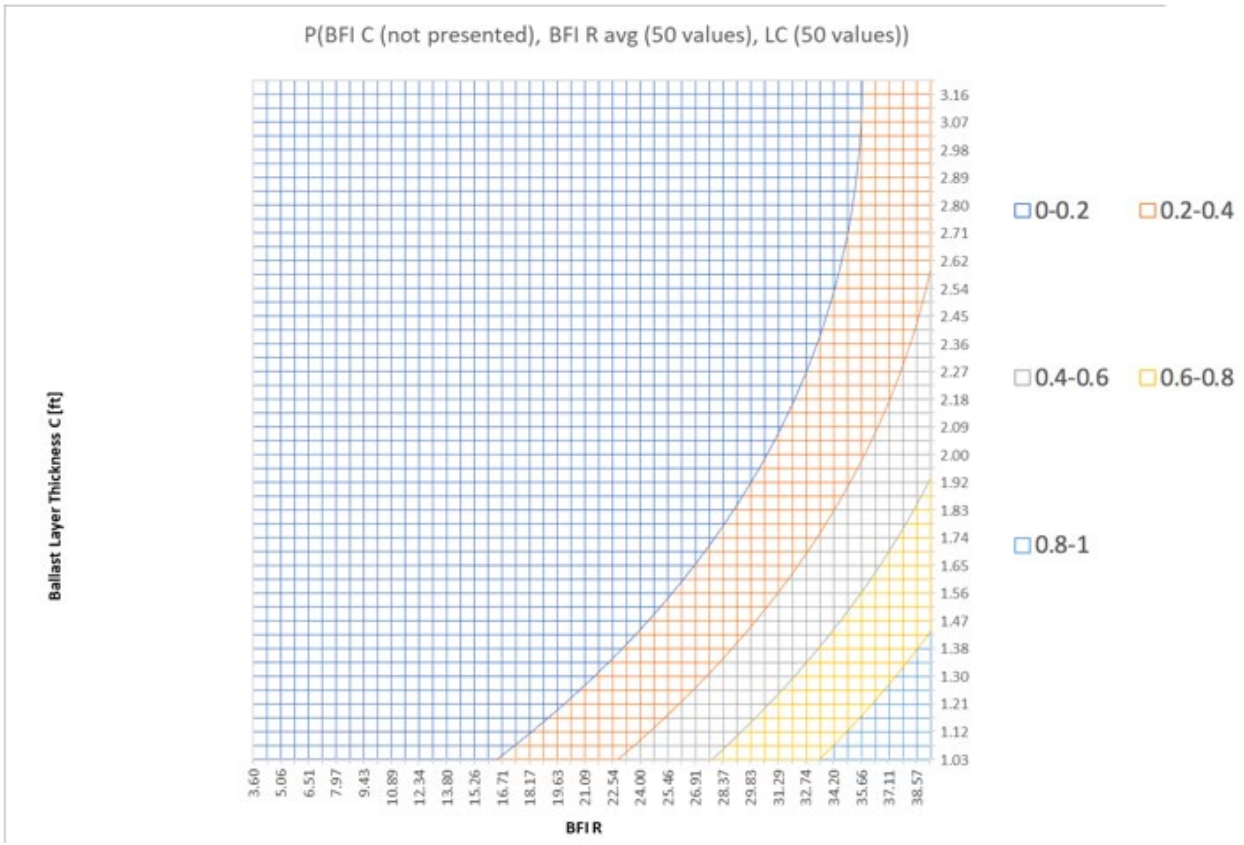


Figure 64A: Probability of a profile defect as a function of BFI/BLT combinations

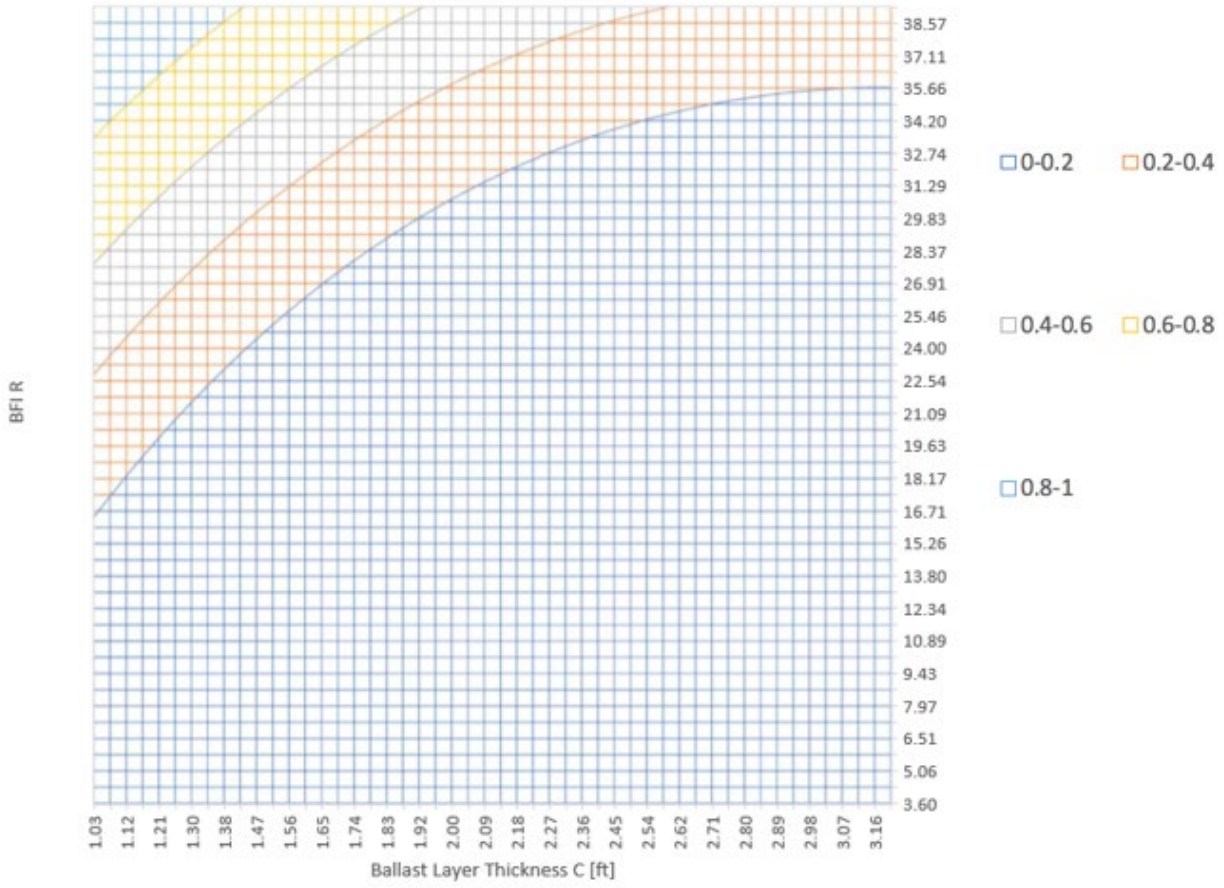


Figure 65B: Probability of a profile defect as a function of BFI/BLT combinations (axes reversed)

These figures can likewise be particularly useful to track engineers who know the condition of their BFI and BLT so they can ascertain the risk of developing a track geometry profile value of greater than 10 mm [0.4 inches].

Overall, the results of the study showed that there was a statistically significant relationship between high rates of geometry degradation and poor subsurface conditions as defined by GPR parameters, particularly the BFI and the BLT. The result was a predictive model that was developed to determine the probability of a track geometry defect developing as a function of these key GPR parameters. Moreover, this probability model can be used by track design and maintenance engineers to look at the effect of such important ballast parameters as level of ballast fouling and ballast thickness on the development of track geometry defects and to further look at different actions (and associated trade-offs) that can be performed to improve the performance of the track.

The results of this study showed a statistically significant relationship between the probability of occurrence of a defined measurement value of track geometry profile/surface and poor subsurface conditions as defined by GPR parameters, particularly the BFI and the BLT. Specifically, a high degree of ballast fouling and a thin ballast layer combine to result in an increased probability of a significant track geometry profile measurement for high speed track (> 10 mm [0.4 inches]). The result was a predictive model that was developed to determine the

probability of a track geometry profile measurement greater than 10 mm [0.4 inches] developing as a function of these key GPR parameters. This probability model provides a methodology by which track design and maintenance engineers can evaluate the effect of such important ballast parameters (level of ballast fouling and ballast thickness) on the development of track geometry defects. This in turn can be used to further evaluate different maintenance actions (and associated trade-offs) that can be implemented to improve the long-term performance of the track.

In addition, the analyses approach presented here, particularly the higher order data analytics approach using combinational hybrid analysis to include hierarchical clustering analysis with histogram data, lends itself to other analyses involving large datasets. Particularly, the benefit of this approach is that it is an unsupervised learning analysis approach where the model tries to understand the patterns of data in the dataset and identifies the key variables that influence the analysis. Thus, several independent variables can be initially identified, and the model (using the dendrogram approach) will cluster the independent variables together in similar groups and then identify the key independent variables that influence the dependent variable (which in this analysis was the probability of having a track geometry defect). This is a powerful tool with broad application to the railway (and other) industries which is experiencing order of magnitude increases in data collected, without necessarily having the tools to analyze this data beyond the most simplistic means (e.g., threshold analysis, etc.)

7.1 Recommendations

The results of this model and the analytical tools used in its development represent only the starting point of this research. This study clearly shows a relationship between GPR measurement of the ballast condition and one (key) track geometry parameter profile (surface).

The first extension of this research should be to widen its scope beyond the initial dataset and to add additional data, both from Amtrak (where significant additional data is available) and other Class I railroads, particularly freight railroads, where the maintenance thresholds for track geometry maintenance are much lower than on the Northeast Corridor for high-speed rail.

Continued research should also be extended to include the other track geometry parameters, particularly cross-level, warp, and twist, all of which are related to profile variation.

The initial study included MRail data (as recorded by the FRA DOTX218 car on CSX), but the validity of this data was questionable. It would be of great value to obtain useful MRail data and combine it with the GPR data to determine if improved modeling analysis can be performed with a better developed model.

Likewise, other ballast and subgrade measurement techniques can be added and correlated with the GPR data analyzed here.

Finally, the higher order data analytics approach using combinational hybrid analysis to include hierarchical clustering analysis with histogram data, lends itself to other analyses involving large data sets. This unsupervised learning analysis approach can assist in the analysis of very large-scale datasets where no clear engineering or superficially apparent relationships exist, but where there may be significant underlying relationships.

8. References

Citation	References
[1]	Palese, J. W., Hartsough, C. H., Zarembski, A. M., Thompson, H., Ling, H. L., and Pagano, W., "Life Cycle Benefits of Subgrade Reinforcement Using Geocell on a Highspeed Railway - A Case Study," in *American Railway Engineering Association Annual Conference*, Indianapolis, IN, September 2017.
[2]	Zarembski, A. M., Grissom, G. T., and Euston, T. L., "On the Use of Ballast Inspection Technology for the Management of Track Substructure," *Journal of Transportation Infrastructure Geotechnology, 1*(1), pp. 83–109, 2014.
[3]	Zarembski, A. M., Palese, J. W., Hartsough, C. M., Ling, H. I., and Thompson, H., "Application of Geocell Track Substructure Support System to Correct Surface Degradation Problems Under High-Speed Passenger Railroad Operations," *Journal of Transportation Infrastructure Geotechnology, 4*(4), pp. 106–125, December 2017.
[4]	Attoh-Okine, N. O, Big Data and Differential Policy: Analysis Strategies for Railroad Track Engineering, Hoboken, NJ: John Wiley & Sons, 2017.
[5]	Freedman, D. A., "Selecting and interpreting measures of thematic classification accuracy," in *Statistical Models: Theory and Practice*, Revised Edition: 2009.6 ed., vol. 62, S. V. Stehman, Ed., Cambridge University Press, 1997, pp. 77–89.7.
[6]	Milligan, G. W., and Cheng, R., "Measuring the influence of individual data points in a cluster analysis," *Journal of Classification, 13*(2), pp. 315–335, 1996.
[7]	Billard, L., and Diday, E., Symbolic Data Analysis: Conceptual Statistics and Data Mining, John Wiley & Sons, 2012.
[8]	Billard, L., and Kim, J., "Hierarchical clustering for histogram data," *Wiley Interdisciplinary Reviews: Computational Statistics, 9*(5), pp. e1405, 2017.
[9]	Irpino, A., and Verde, R., "Basic statistics for distributional symbolic variables: a new metric-based approach," *Advances in Data Analysis and Classification, 9*(2), pp. 143–175, 2015.
[10]	Müllner, D., "Modern hierarchical, agglomerative clustering algorithms," *IEEE Signal Processing Letters, 19*(4), pp. 231–234.12, 2011.
[11]	Punj, G., and Stewart, D.W., "Cluster Analysis in Marketing Research: Review and Suggestions for Application," *Journal of Marketing Research, 20*(2), pp. 134, 1983.
[12]	Irpino, A., Verde, R., and De Carvalho, F. D. A. T., "Dynamic Clustering of histogram data based on adaptive squared Wasserstein distances," *Expert Systems with Applications, 41*(7), pp. 3351–3366, 2014.

Citation **References**

[13] Hosmer. D., Lemeshow, S., and Sturdivant, R. X., "Logistic Regression Modelsfor the Analysis of Correlated Data," in *Applied Logistic Regression*, Third Edition ed., John Wiley & Sons, 2013, pp. 313–376.

[14] Boryga, M., and Graboś, A., "Planning of manipulator motion trajectory with higher-degree polynomials use," *Mechanism and Machine Theory, 44*(7), pp. 1400–1419, 2009.

[15] Zarembski, A. M., Yurlov, D., Palese, J. W., Attoh-Okine, N., and Thompson, H., "Relationship between Track Geomery Degradation and Subsurface Condition as Measured by GPR," in *American Railway Engineering Annual Conference*, Chicago, IL, September 2018.

Appendix A:
Exploratory Data Analysis (EDA)

A.1. EDA Dataset Description

Table A.1: Variables summary output, rows description

Row Number	Information given about the variable
1	Variable name
2	Number of observation and min value
3	Class of the variable and its 1st quartile value
4	Provide the Mode of the character and its median
5	Provide the mean
6	Provide the 3rd quartile value
7	Provide the max value

A.1.1 Multivariable Plot

Variables of the CSX Peninsula Subdivision data MP 67 to 69; YRel Right (from MRail and Right Profile 31 (track geometry car).

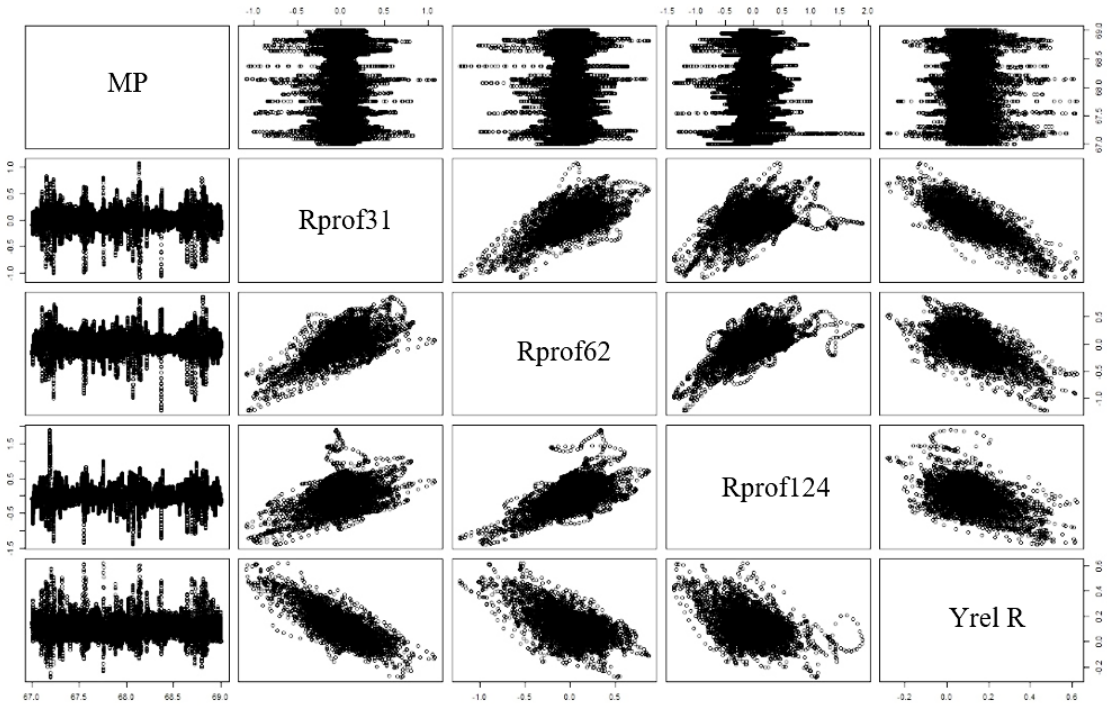


Figure A.1: CSX Peninsula Subdivision MP 67–69 multivariable plot-2

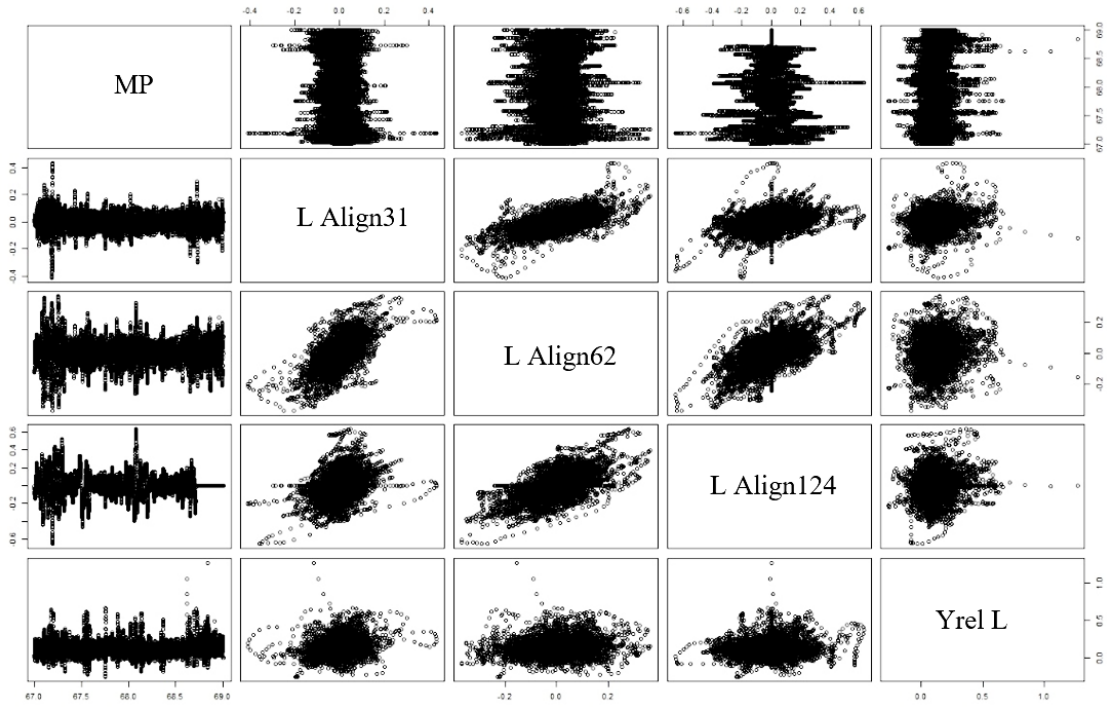


Figure A.2: CSX Peninsula Subdivision MP 67–69 multivariable plot-3

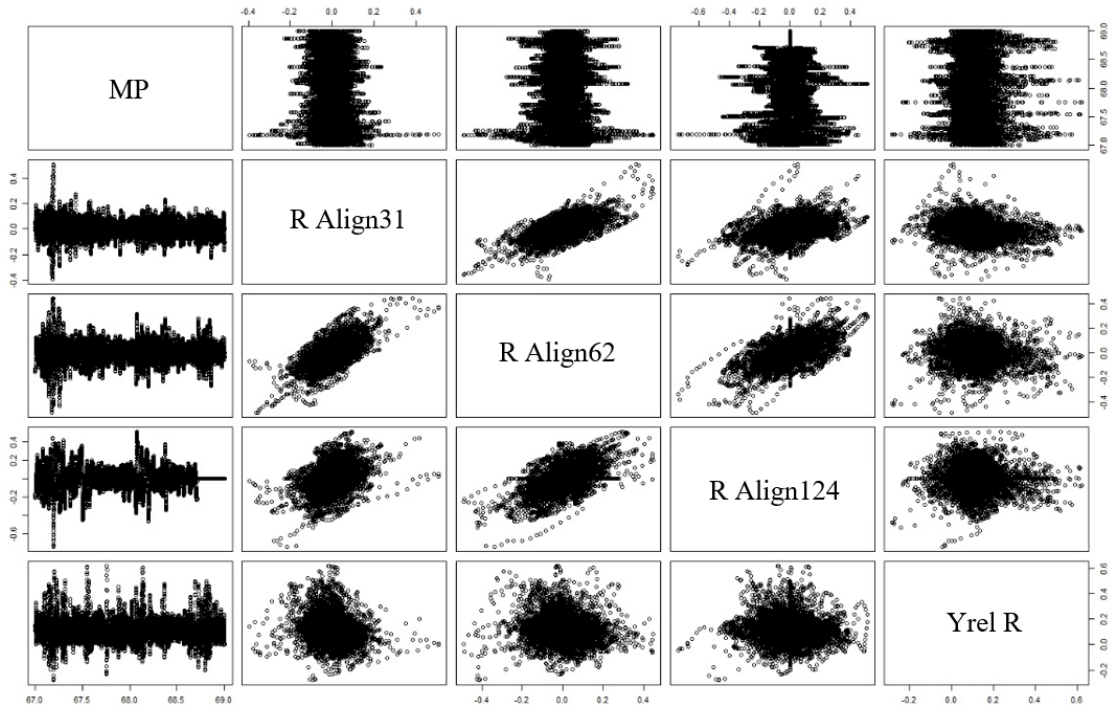


Figure A.3: CSX Peninsula Subdivision MP 67–69 multivariable plot-4

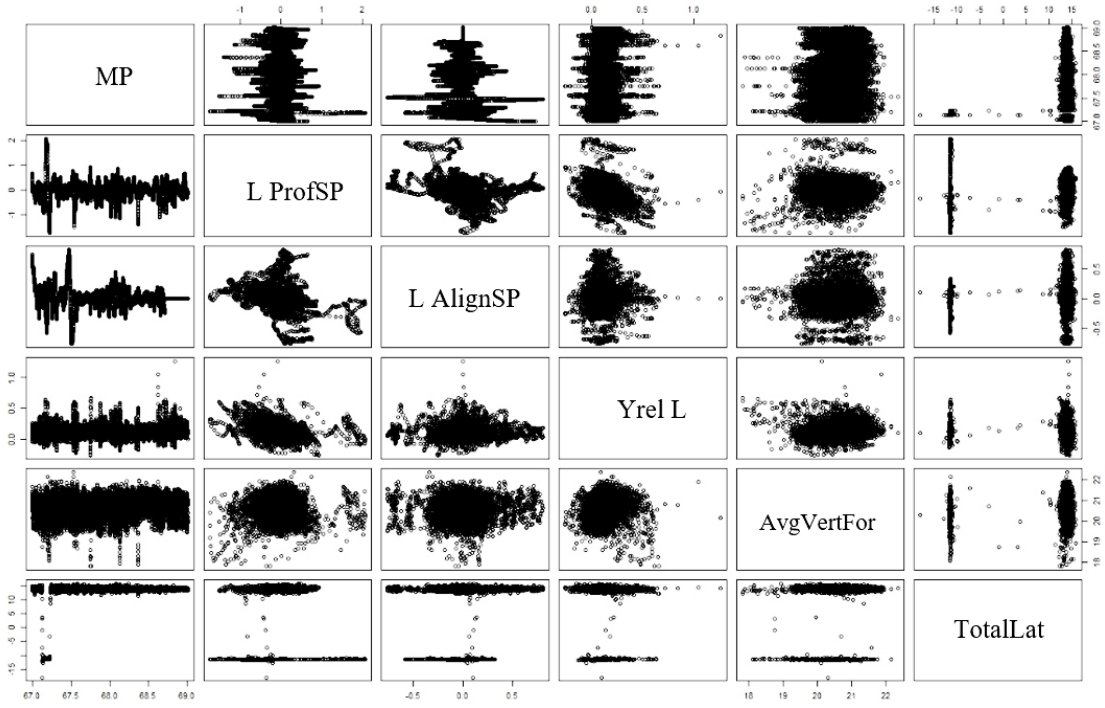


Figure A.4: CSX Peninsula Subdivision MP 67–69 multivariable plot-5

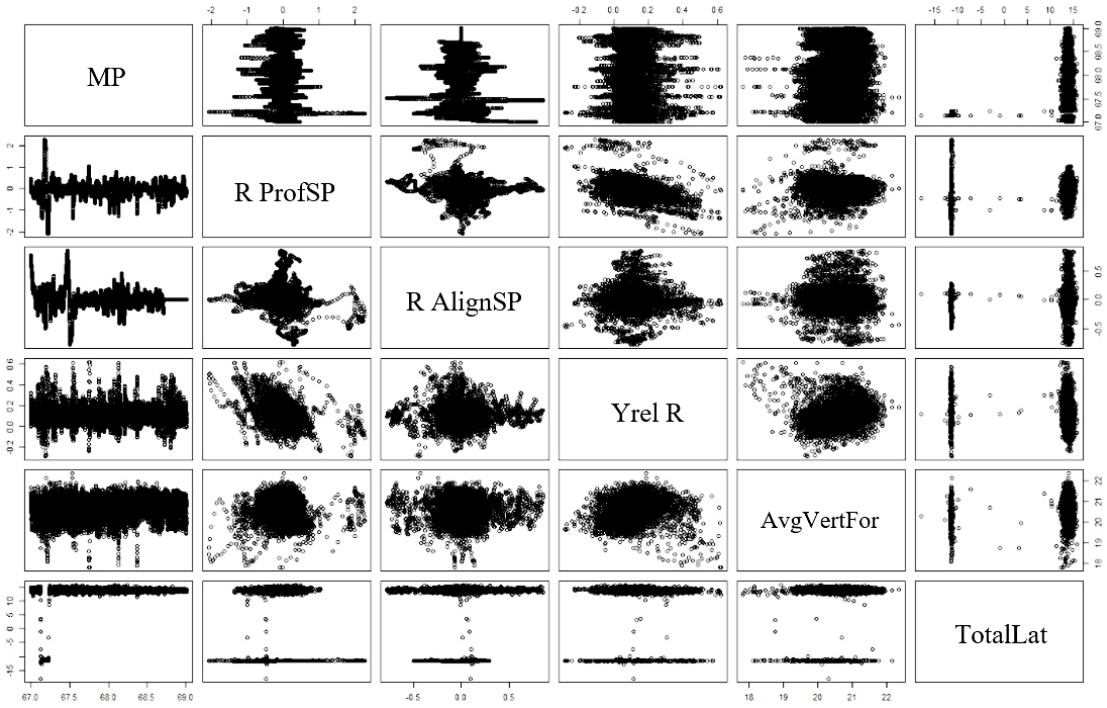


Figure A.5: CSX Peninsula Subdivision MP 67–69 multivariable plot-6

A.1.2 Box Plots

From the CSX Peninsula Subdivision MP 67–69 database box-plot examples.

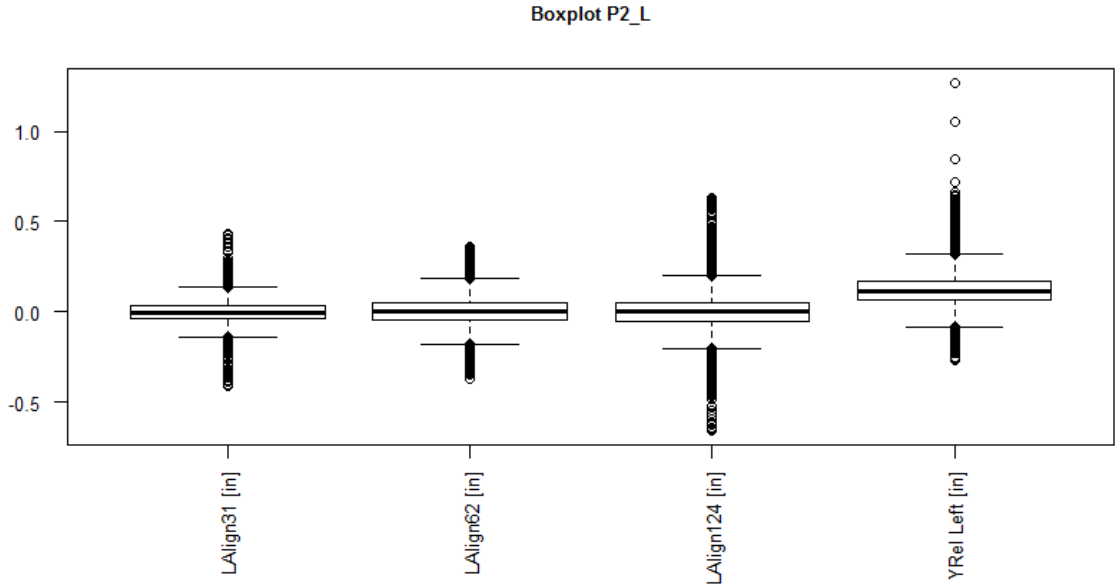


Figure A.6: Box and whisker plot CSX Peninsula Subdivision MP 67–69 data plot-2

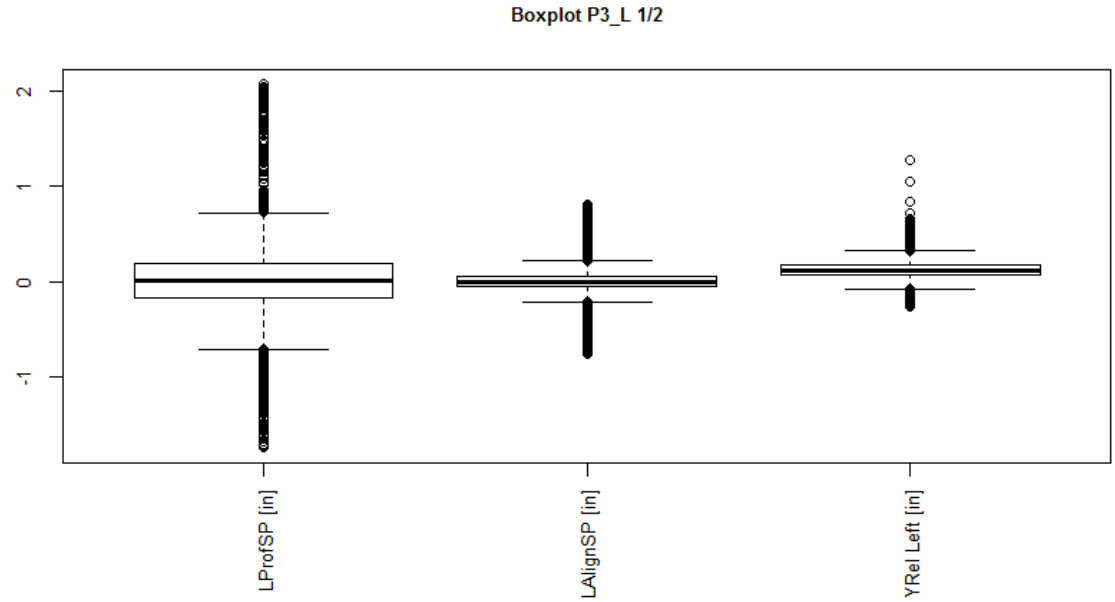


Figure A.7: Box and whisker plot CSX Peninsula Subdivision MP 67–69 data plot-3

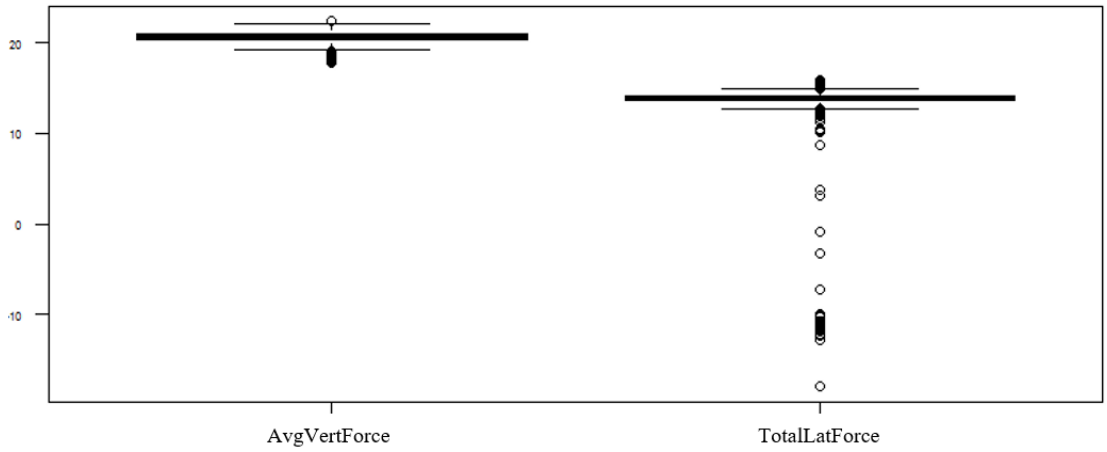


Figure A.8: Box and whisker plot CSX Peninsula Subdivision MP 67–69 data plot-4

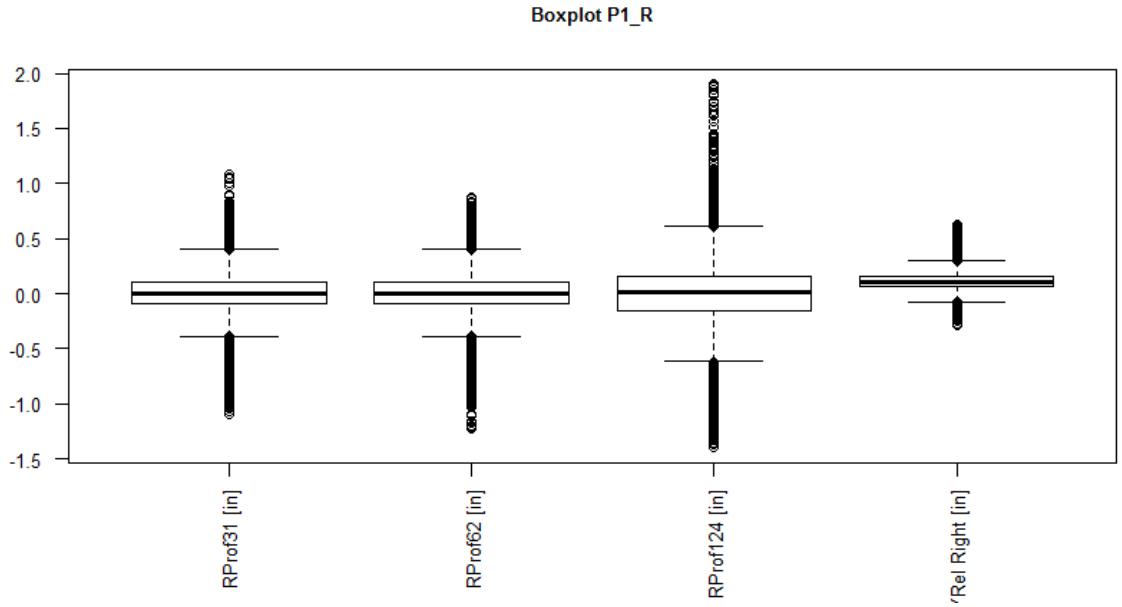


Figure A.9: Box and whisker plot CSX Peninsula Subdivision MP 67–69 data plot-5

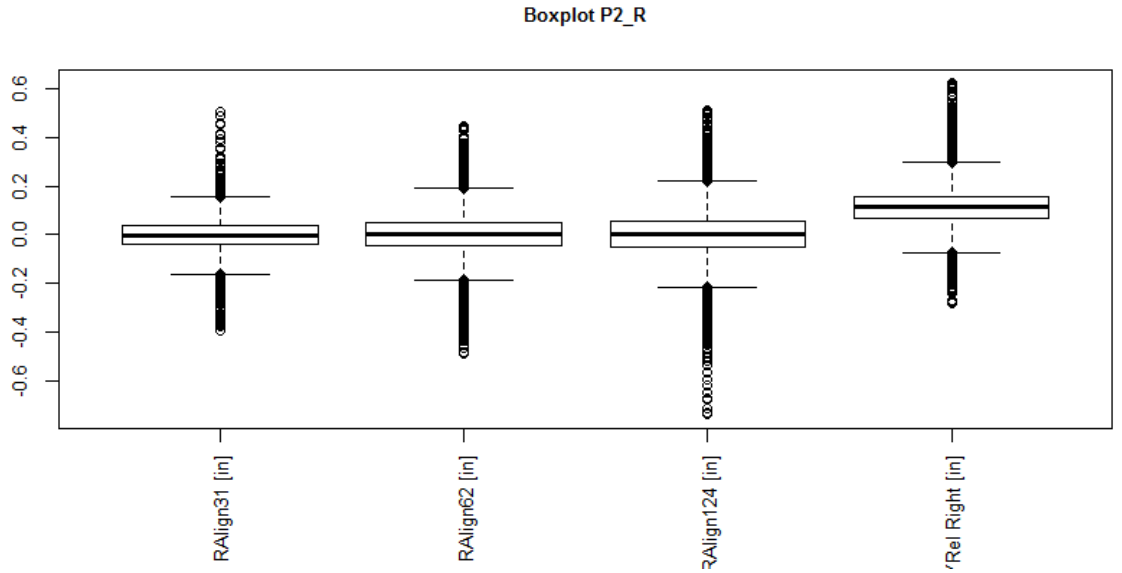


Figure A.10: Box and whisker plot CSX Peninsula Subdivision MP 67–69 data plot-6

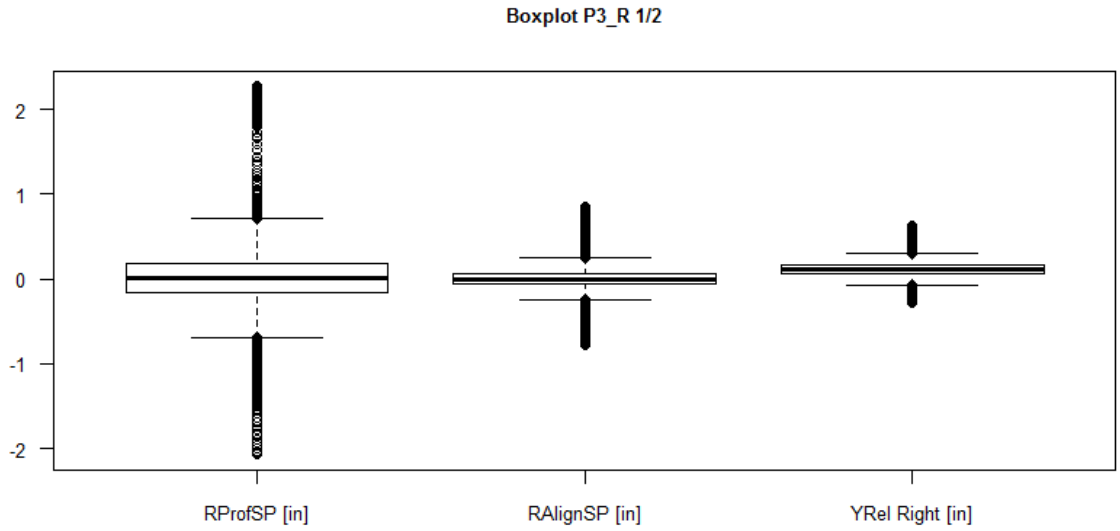


Figure A.11: Box and whisker plot CSX Peninsula Subdivision MP 67–69 data plot-7

A.1.3 Histogram and KDE

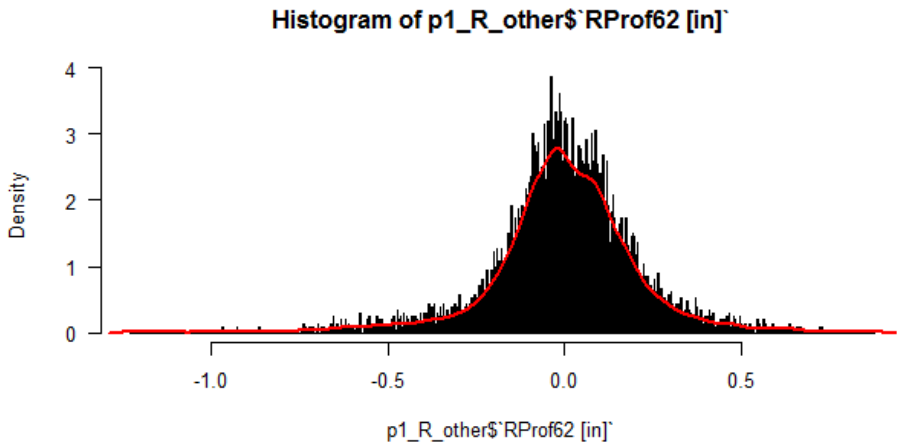


Figure A.12: Histogram and KDE of CSX Peninsula Subdivision MP 67–69 inspection RProf62

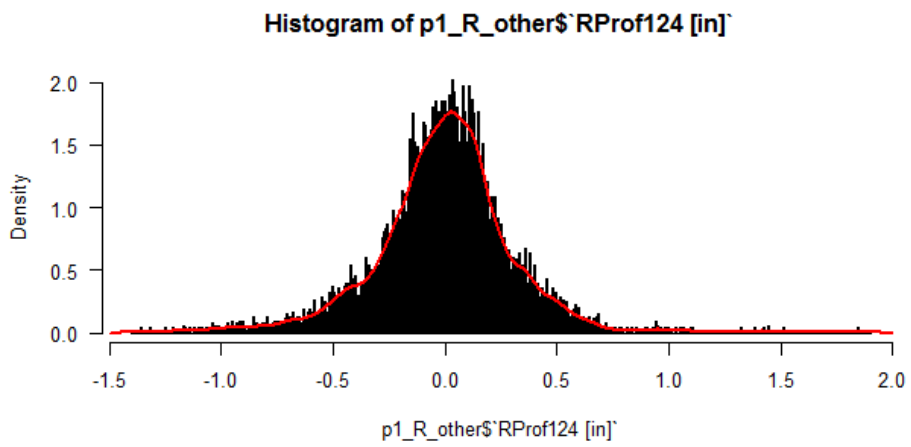


Figure A.13: Histogram and KDE of CSX Peninsula Subdivision MP 67–69 inspection Rprof124

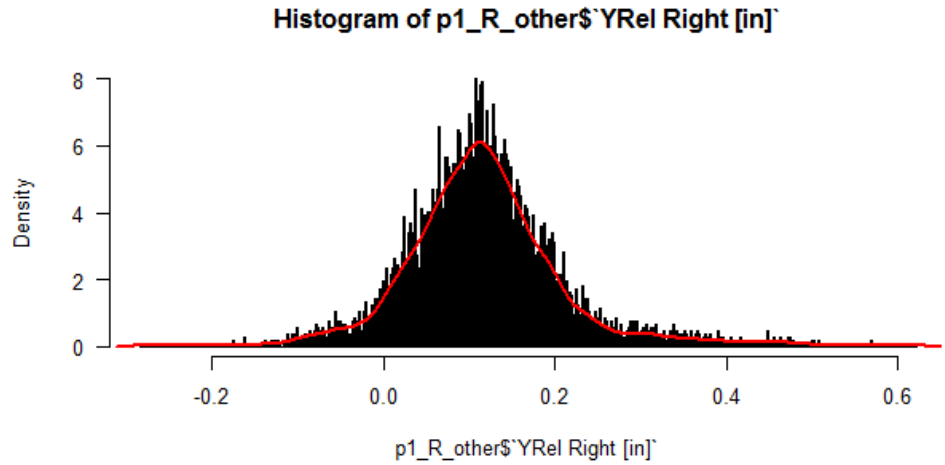


Figure A.14: Histogram and KDE of CSX Peninsula Subdivision MP 67–69 inspection YRel R

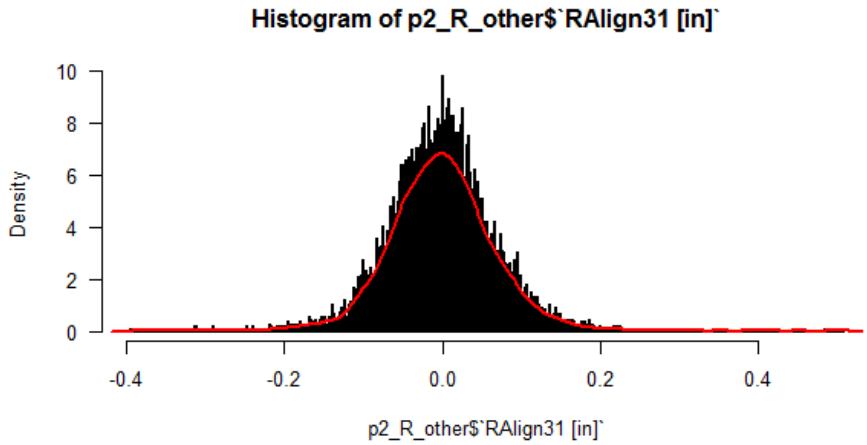


Figure A.15: Histogram and KDE of CSX Peninsula Subdivision MP 67–69 inspection RAlign31

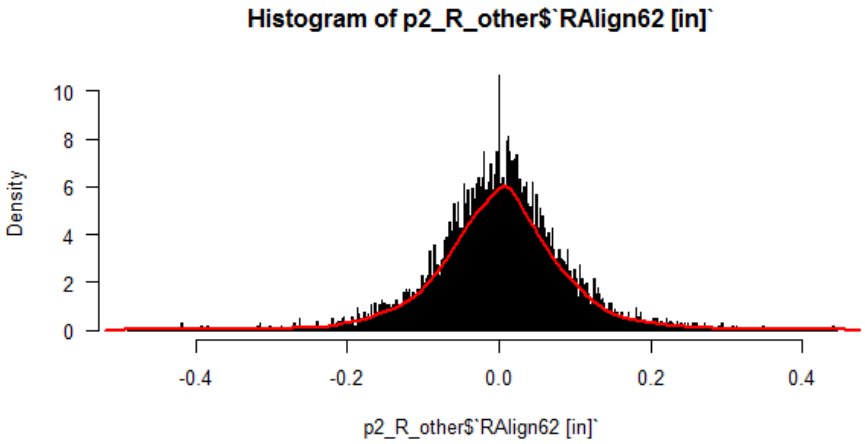


Figure A.16: Histogram and KDE of CSX Peninsula Subdivision MP 67–69 inspection RAlign62

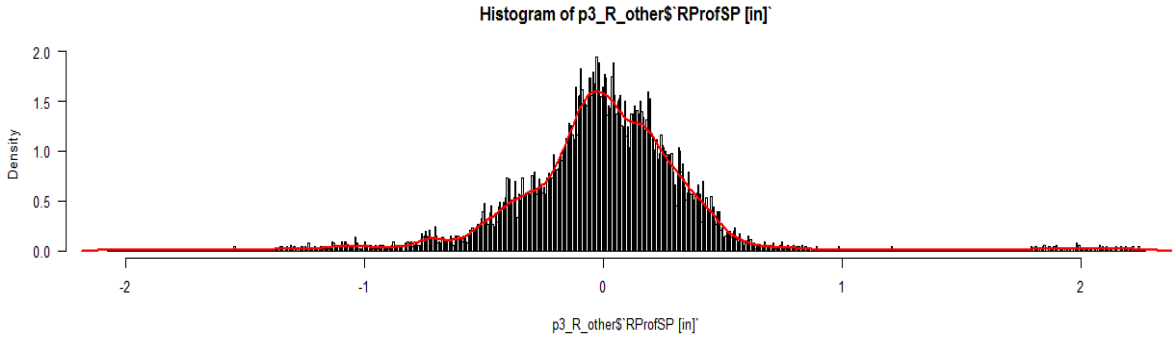


Figure A.17: Histogram and KDE of CSX Peninsula Subdivision MP 67–69 inspection RProfSP

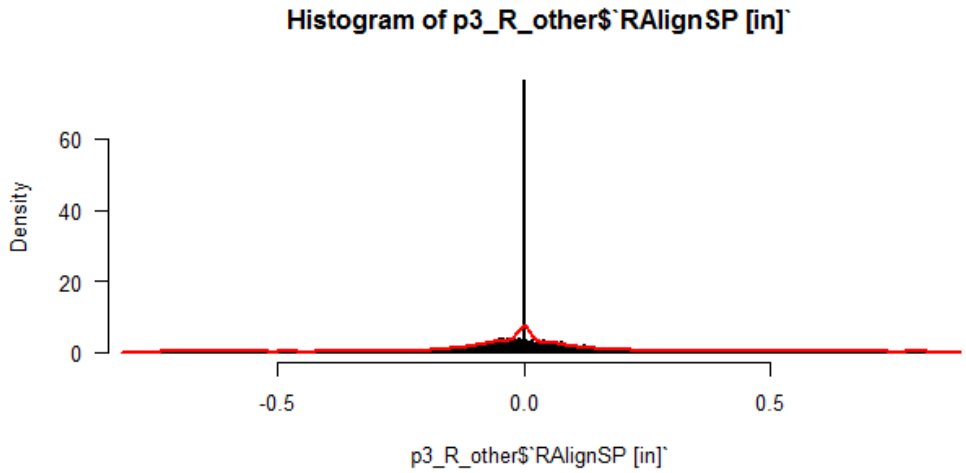


Figure A.18: Histogram and KDE of CSX Peninsula Subdivision MP 67–69 inspection RAlignSP

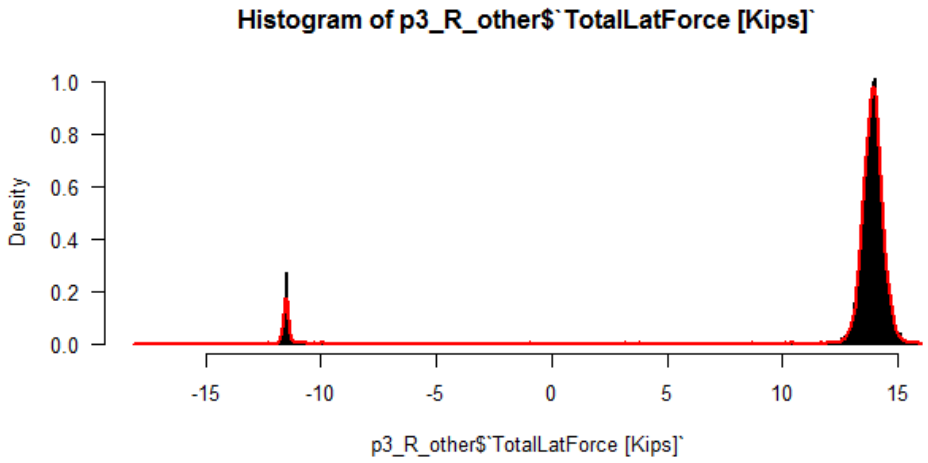


Figure A.19: Histogram and KDE of CSX Peninsula Subdivision MP 67–69 inspection TotalLatForce

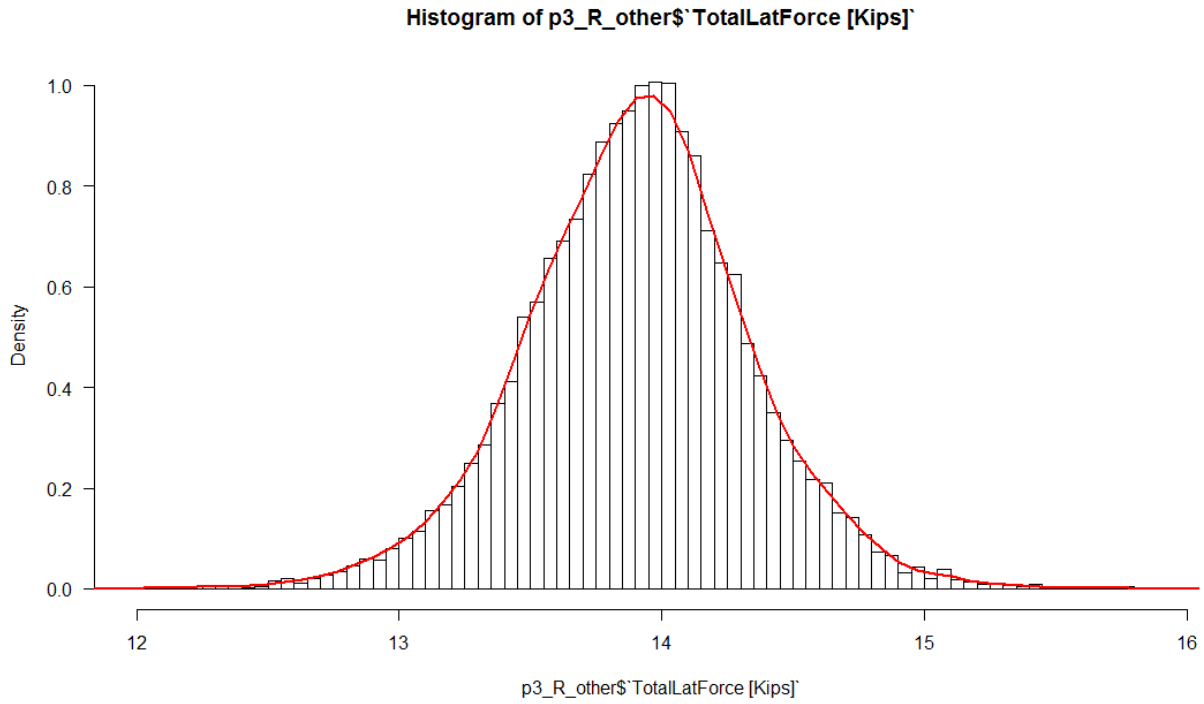


Figure A.20: Histogram and KDE of CSX Peninsula Subdivision MP 67–69 inspection TotalLatForce concentration on the one peak

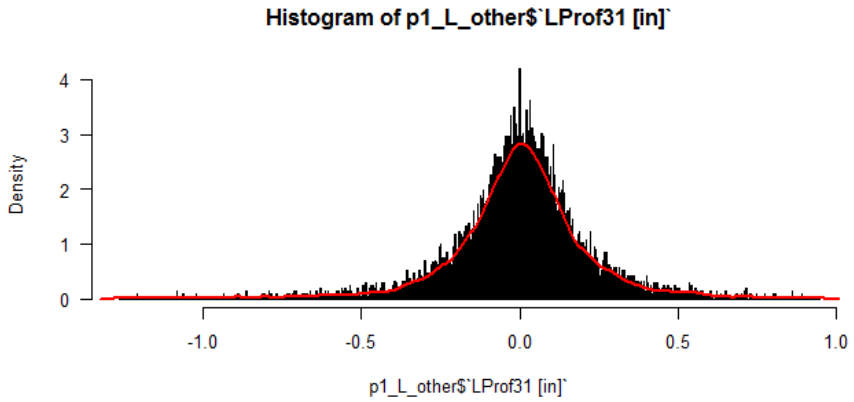


Figure A.21: Histogram and KDE of CSX Peninsula Subdivision MP 67–69 inspection LProf31

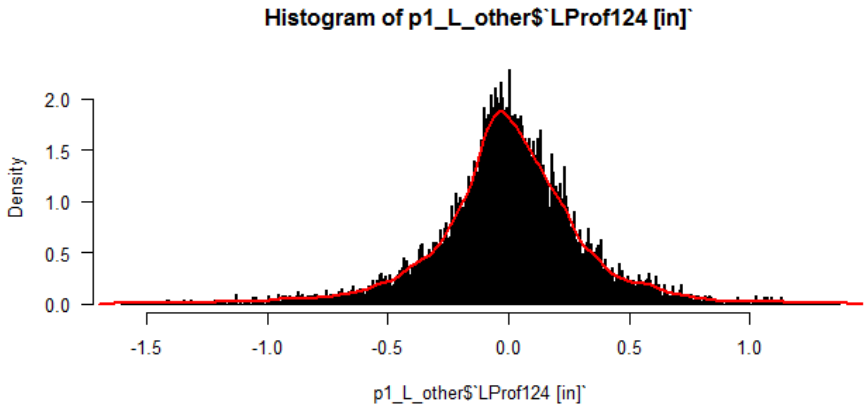


Figure A.22: Histogram and KDE of CSX Peninsula Subdivision MP 67–69 inspection LProf124

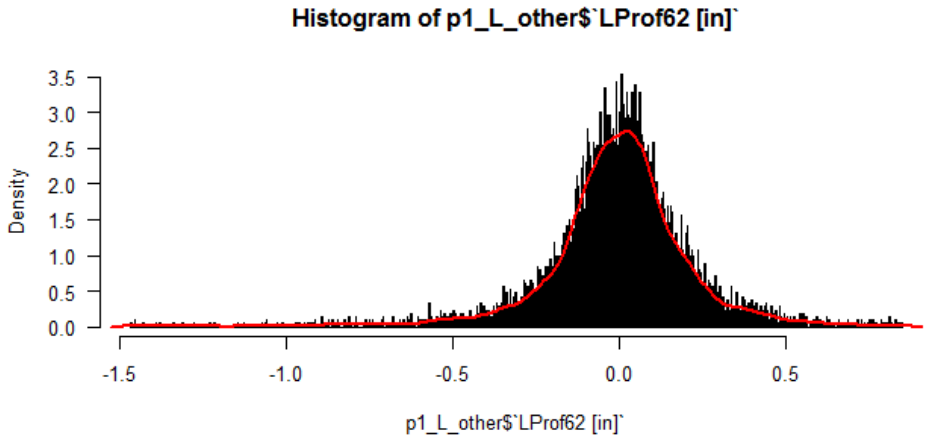


Figure A.23: Histogram and KDE of CSX Peninsula Subdivision MP 67–69 inspection LProf62

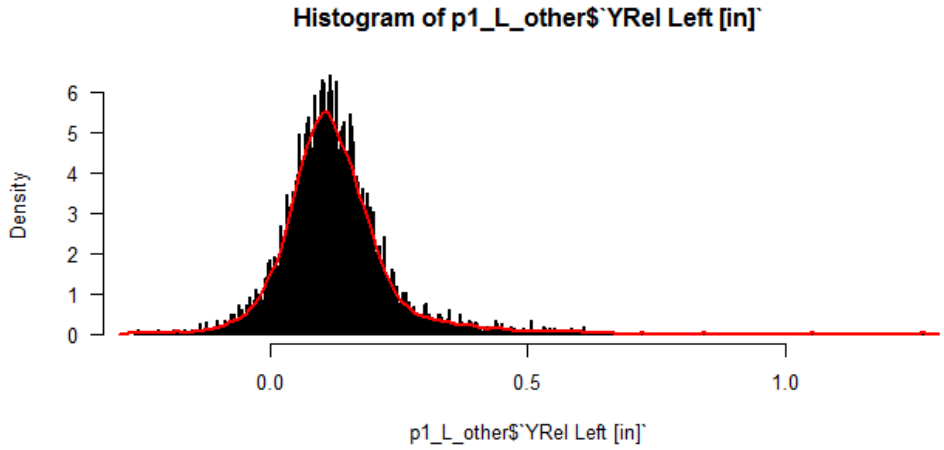


Figure A.24: Histogram and KDE of CSX Peninsula Subdivision MP 67–69 inspection YRel L

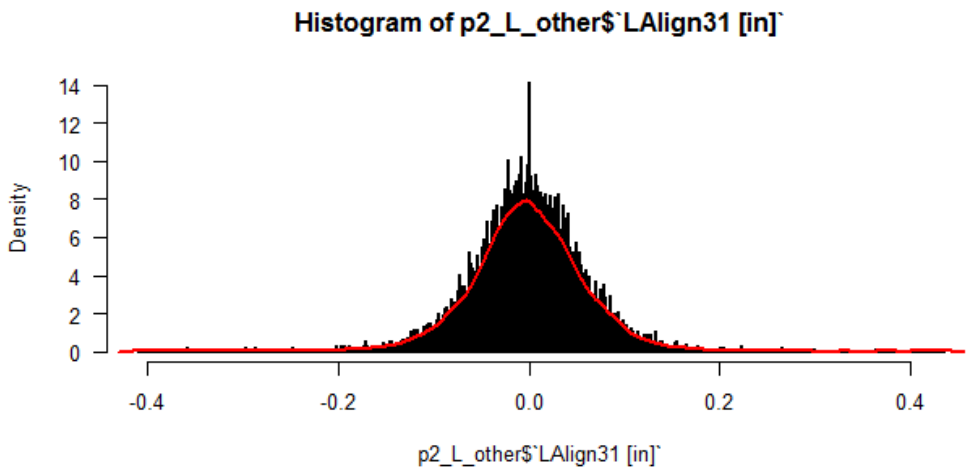


Figure A.25: Histogram and KDE of CSX Peninsula Subdivision MP 67– 69 inspection LAlign31

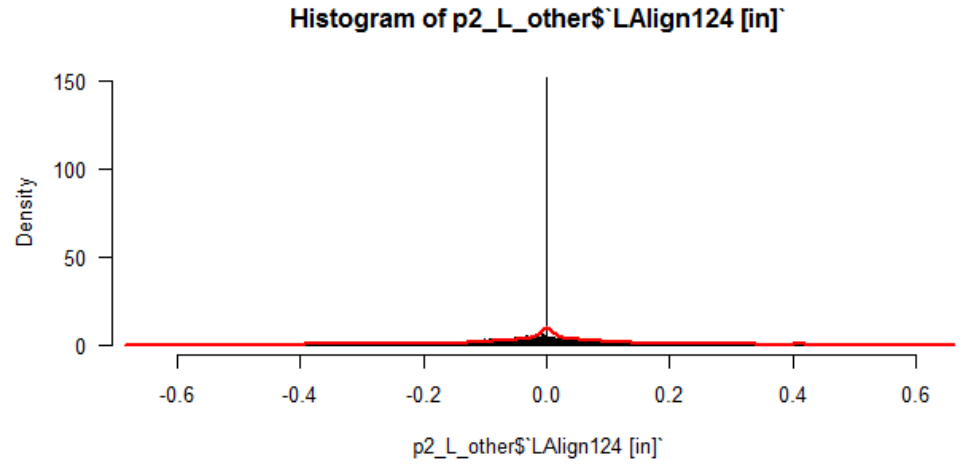


Figure A.26: Histogram and KDE of CSX Peninsula Subdivision MP 67–69 inspection LAlign124. Note, most measurements are at zero

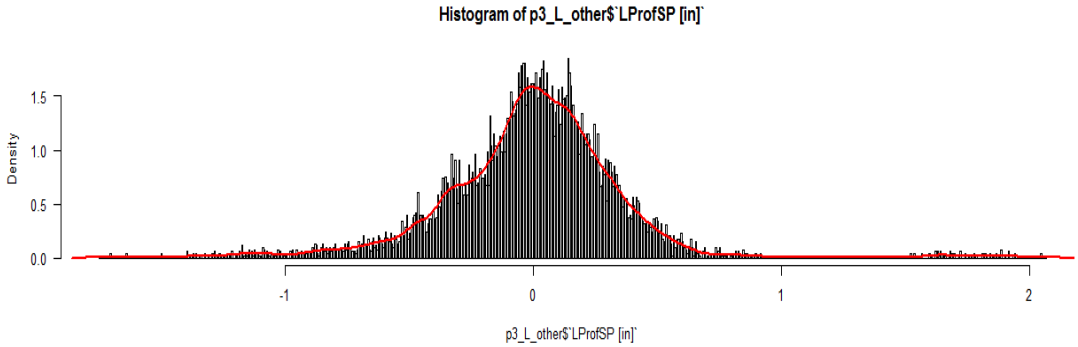


Figure A.27: Histogram and KDE of CSX Peninsula Subdivision MP 67–69 inspection LProfSP

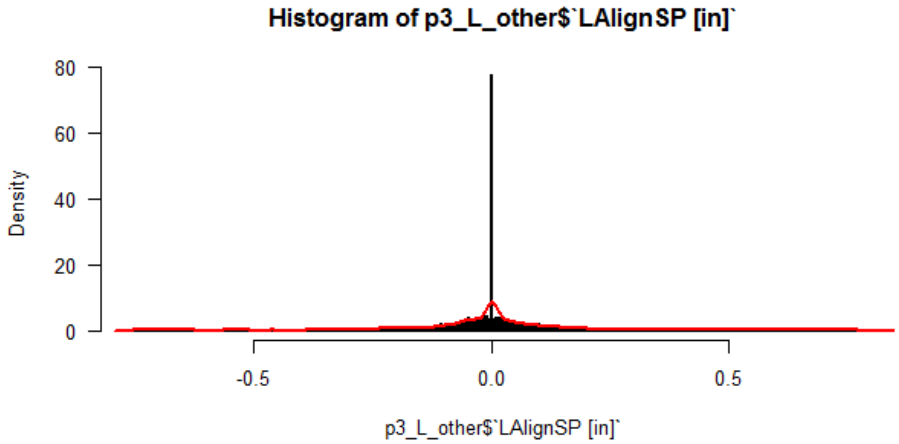


Figure A.28: Histogram and KDE of CSX Peninsula Subdivision MP 67–69 inspection LAlignSP. Note majority of measurements are at zero

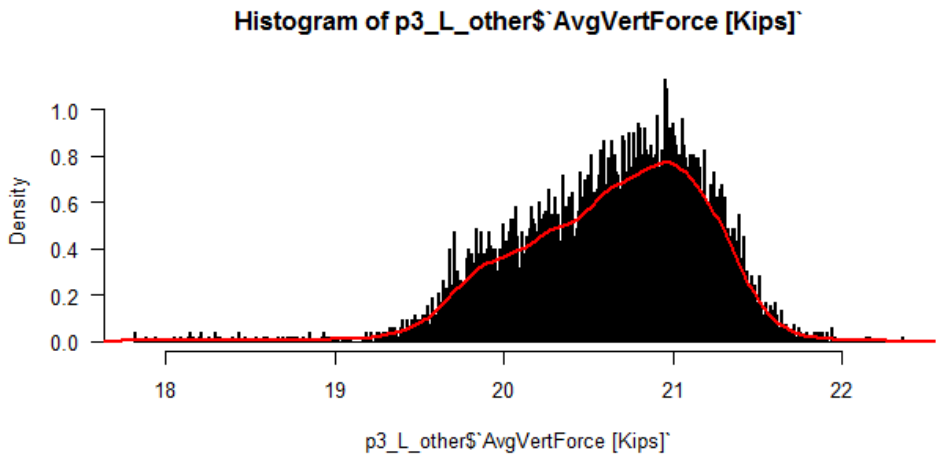


Figure A.29: Histogram and KDE of CSX Peninsula Subdivision MP 67–69 inspection AvgVertForce

A.2 QQ plots

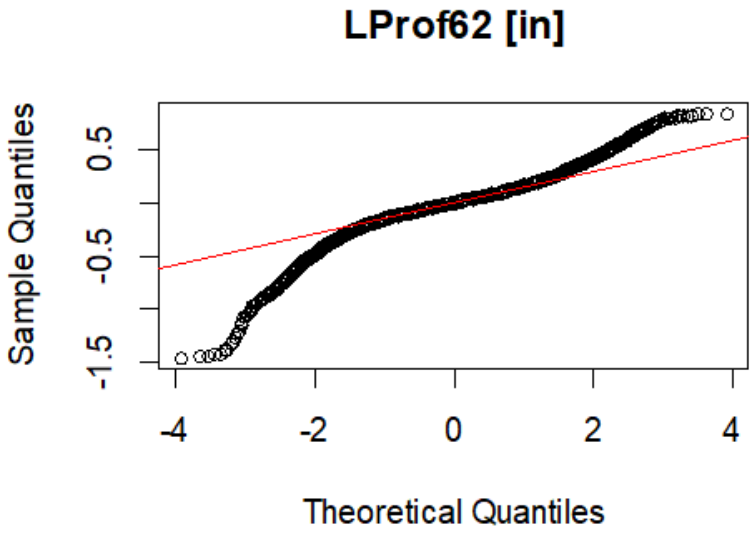


Figure A.30: QQ Plot, CSX Peninsula Subdivision MP 67–69 inspection LProf62

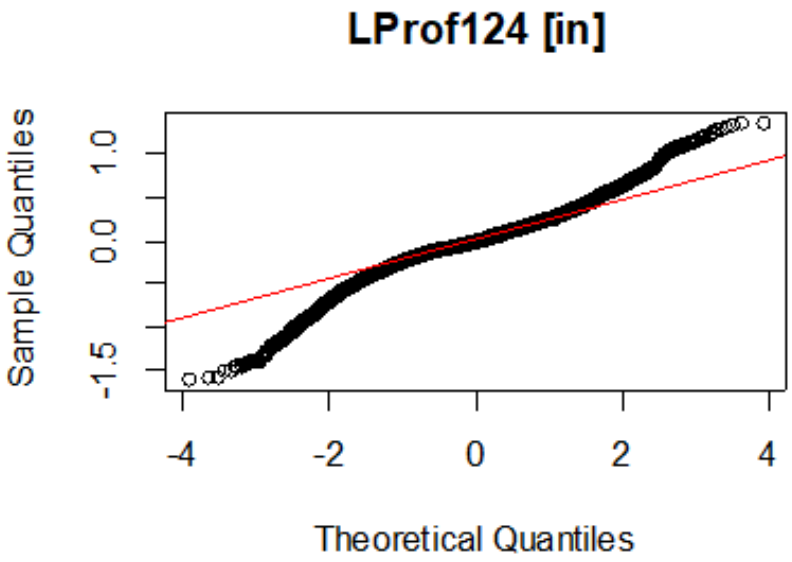


Figure A.31: QQ Plot, CSX Peninsula Subdivision MP 67–69 inspection LProf124

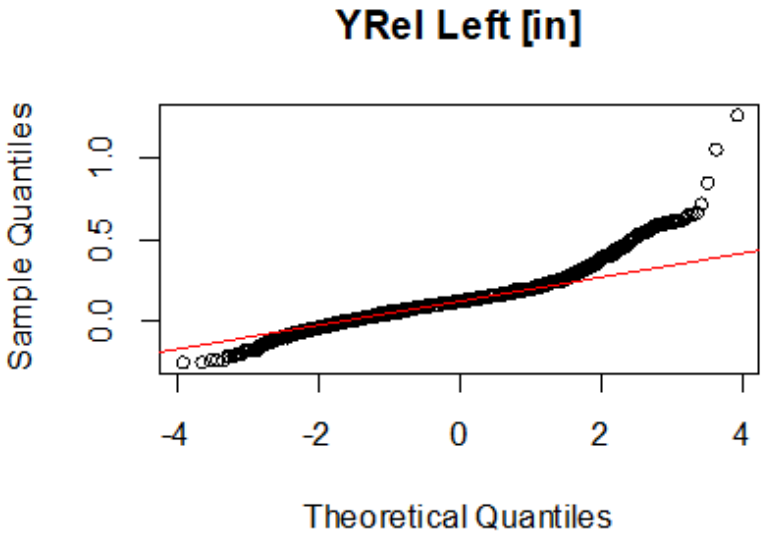


Figure A.32: QQ Plot, CSX Peninsula Subdivision MP 67–69 inspection YRel L

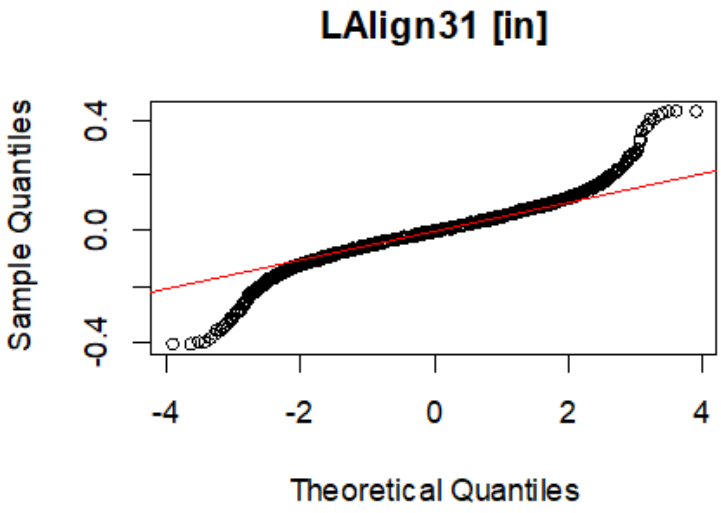


Figure A.33: QQ Plot, CSX Peninsula Subdivision MP 67–69 inspection LAlign31

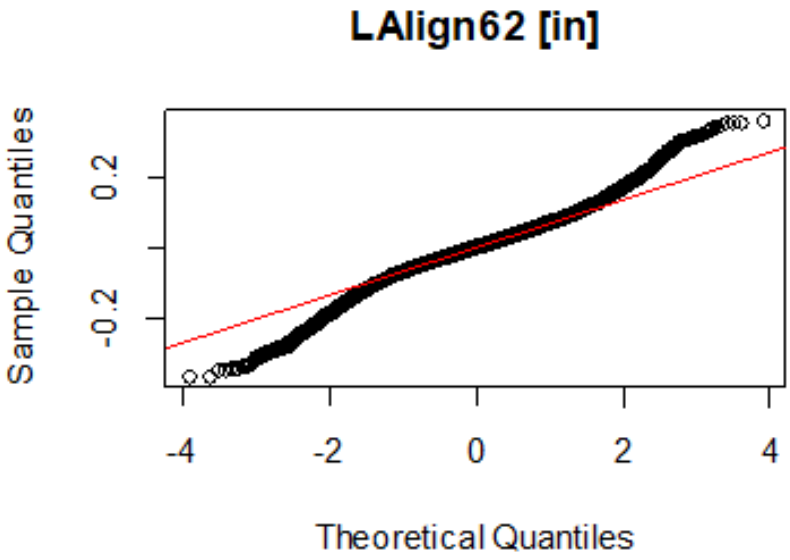


Figure A.34: QQ Plot, CSX Peninsula Subdivision MP 67–69 inspection LAlign62

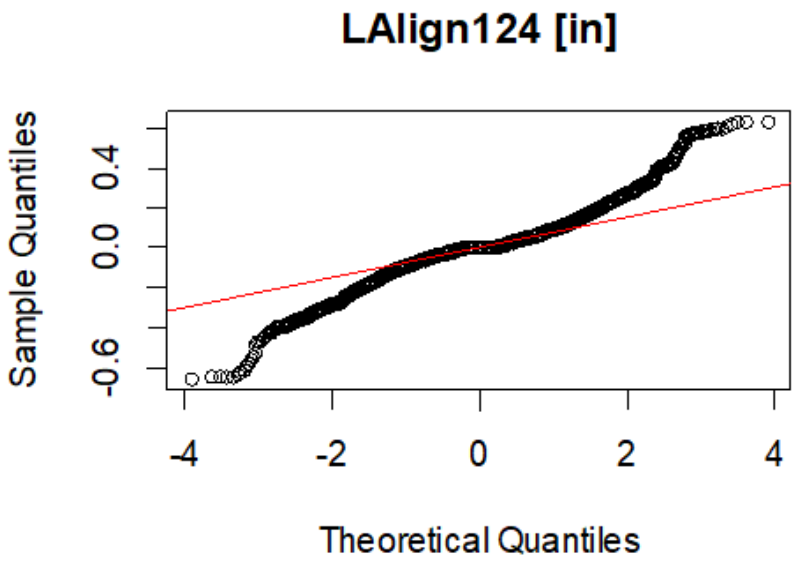


Figure A.35: QQ Plot, CSX Peninsula Subdivision MP 67–69 inspection LAlign124

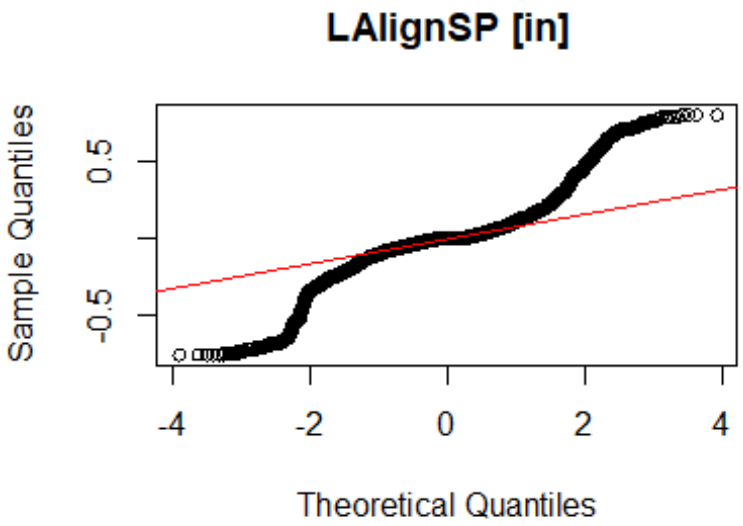


Figure A.36: QQ Plot, CSX Peninsula Subdivision MP 67–69 inspection LAlignSP

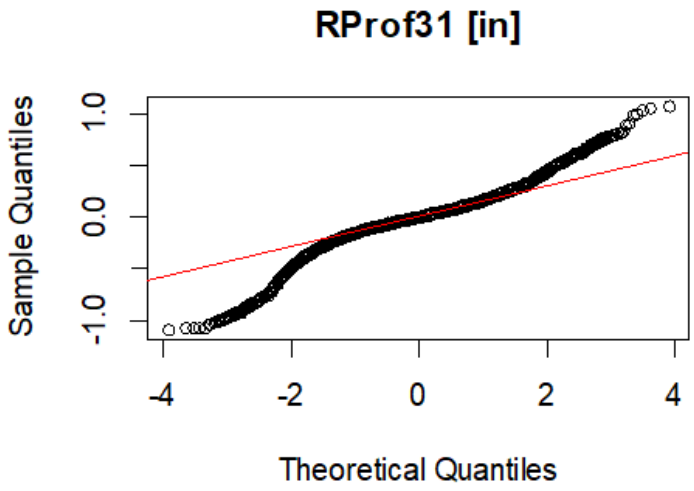


Figure A.37: QQ Plot, CSX Peninsula Subdivision MP 67–69 inspection RProf31

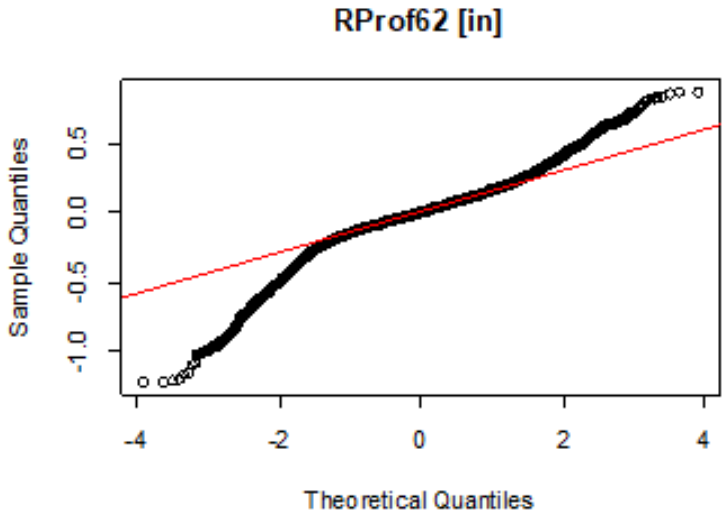


Figure A.38: QQ Plot, CSX Peninsula Subdivision MP 67–69 inspection RProf62

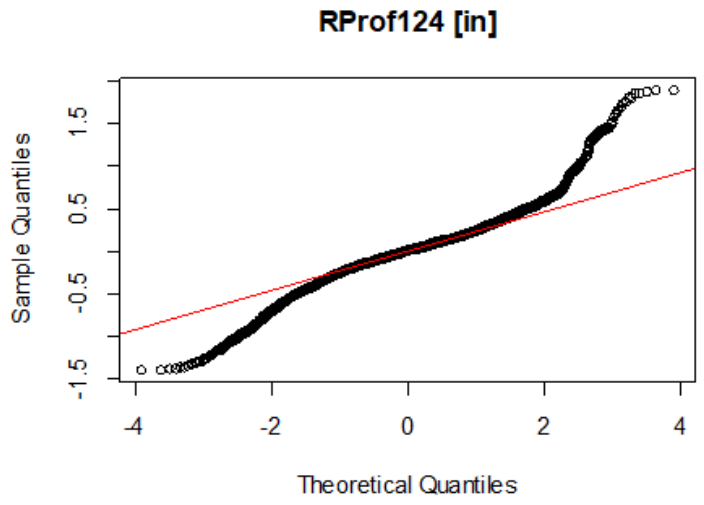


Figure A.39: QQ Plot, CSX Peninsula Subdivision MP 67–69 inspection RProf124

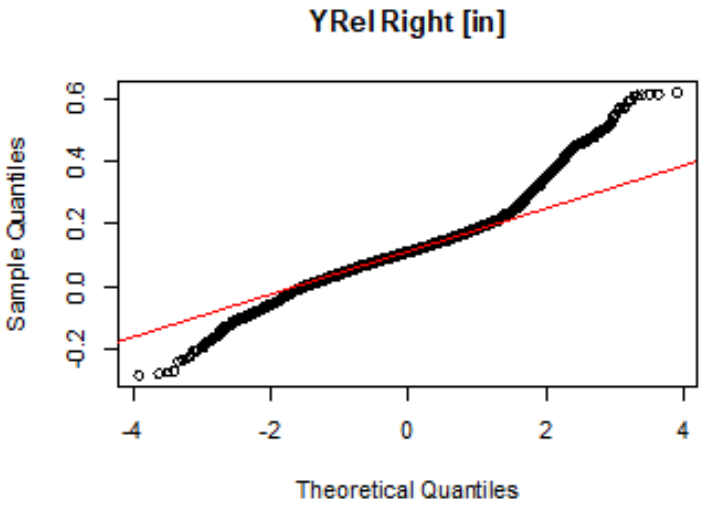


Figure A.40: QQ Plot, CSX Peninsula Subdivision MP 67–69 inspection YRel R

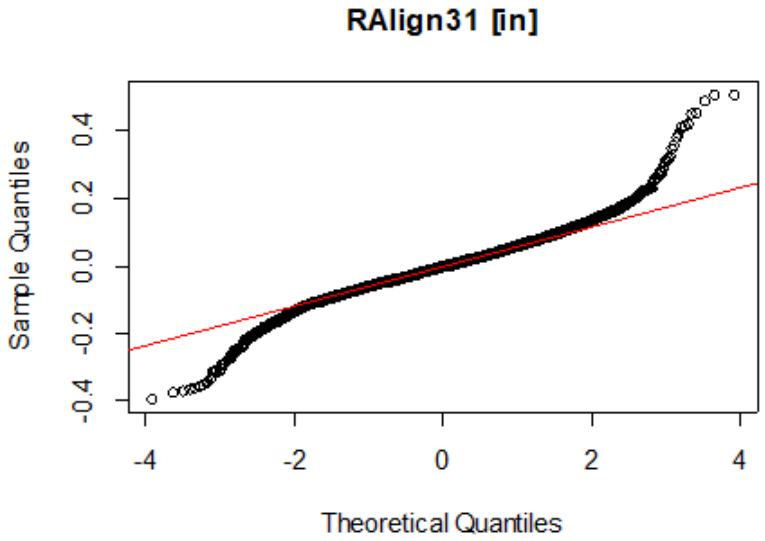


Figure A.41: QQ Plot, CSX Peninsula Subdivision MP 67–69 inspection RAlign31

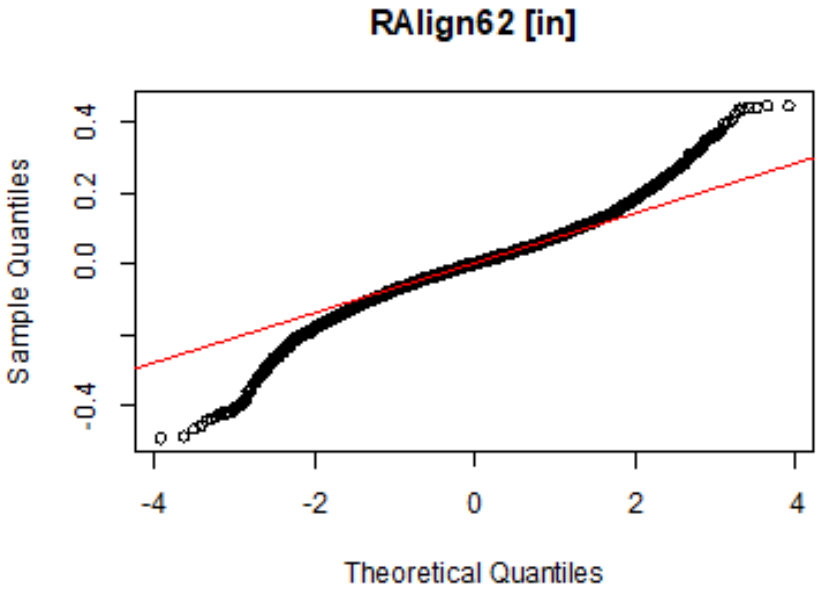


Figure A.42: QQ Plot, CSX Peninsula Subdivision MP 67–69 inspection RAlign62

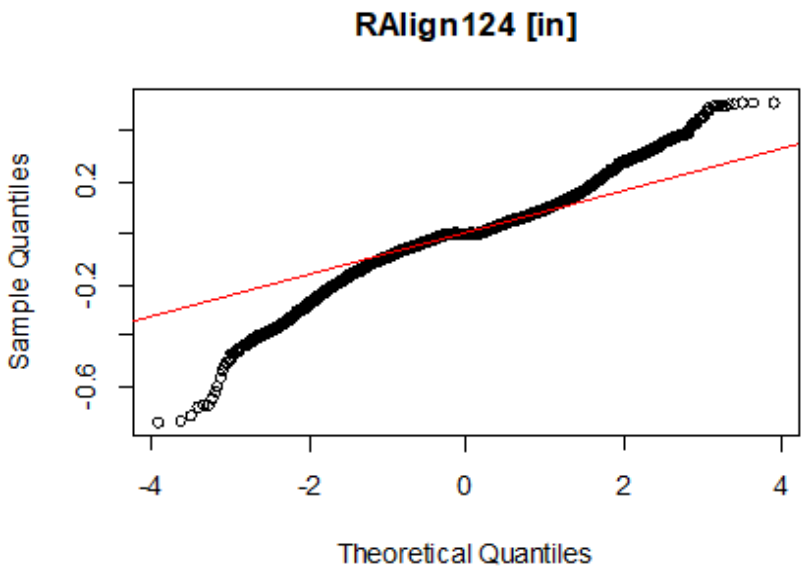


Figure A.43: QQ Plot, CSX Peninsula Subdivision MP 67–69 inspection RAlign124

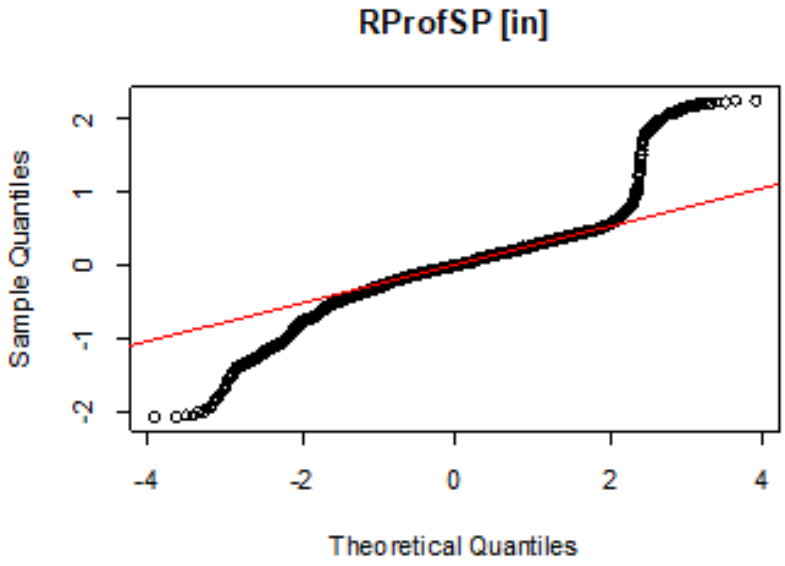


Figure A.44: QQ Plot, CSX Peninsula Subdivision MP 67–69 inspection RProfSP

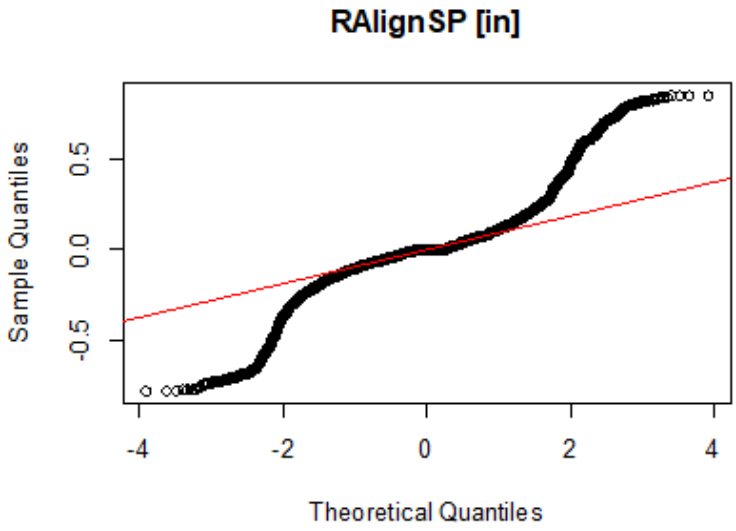


Figure A.45: QQ Plot, CSX Peninsula Subdivision MP 67–69 inspection RAlignSP

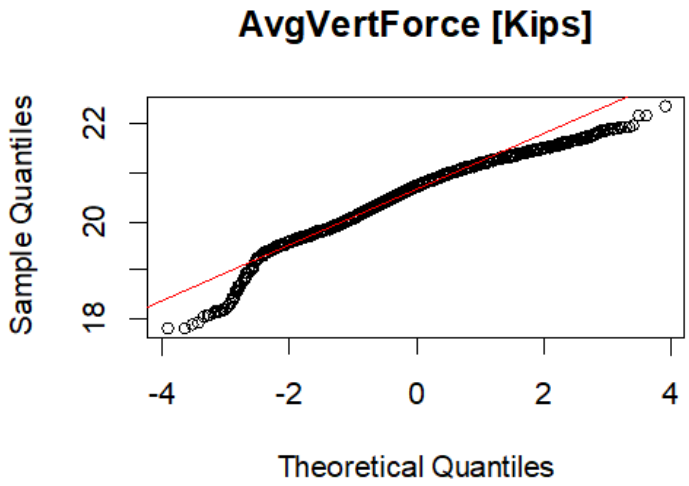


Figure A.46: QQ Plot, CSX Peninsula Subdivision MP 67–69 inspection AvgVertForce

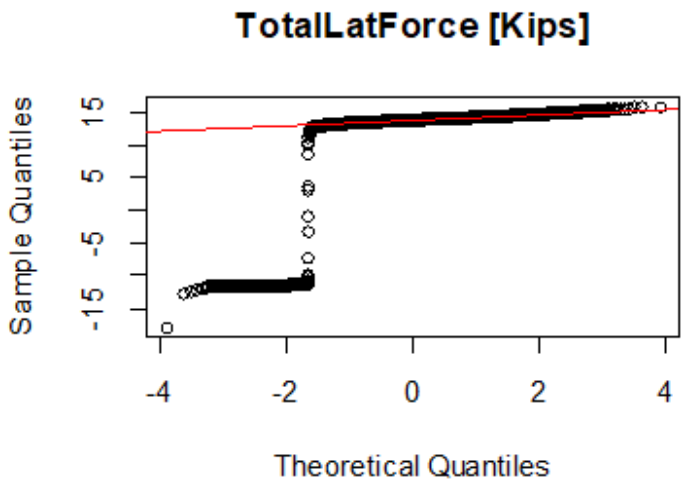


Figure A.47: QQ Plot, CSX Peninsula Subdivision MP 67–69 inspection TotalLatForce

Appendix B:
Logistic Regression Models

B.1 Logistic Regression Model 3.1 P(abs(Rprof62)>0.4) = f(BFIR, BFIC, BLTC)

From R software

```
> summary(logit)
Call:
glm(formula = data1$`abs(Right Prof 62)>0.4` ~ data1$Center +
    data1$Right + data1$`Center: Top of Layer`, family = binomial,
    data = data1)
Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.4819  -0.5349  -0.3641  -0.2875   2.6376
Coefficients:
                              Estimate Std. Error z value Pr(>|z|)
(Intercept)                   -4.98175    1.67692  -2.971  0.00297 **
data1$Center                   0.03969    0.02707   1.466  0.14268
data1$Right                    0.18025    0.04385   4.111 3.94e-05 ***
data1$`Center: Top of Layer` -0.92397    0.56415  -1.638  0.10146
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
(Dispersion parameter for binomial family taken to be 1)
    Null deviance: 207.01  on 252  degrees of freedom
Residual deviance: 169.71  on 249  degrees of freedom
AIC: 177.71
Number of Fisher Scoring iterations: 5
```

B.1.1 Logistic Regression Model 3.1 Sensitivity Analysis

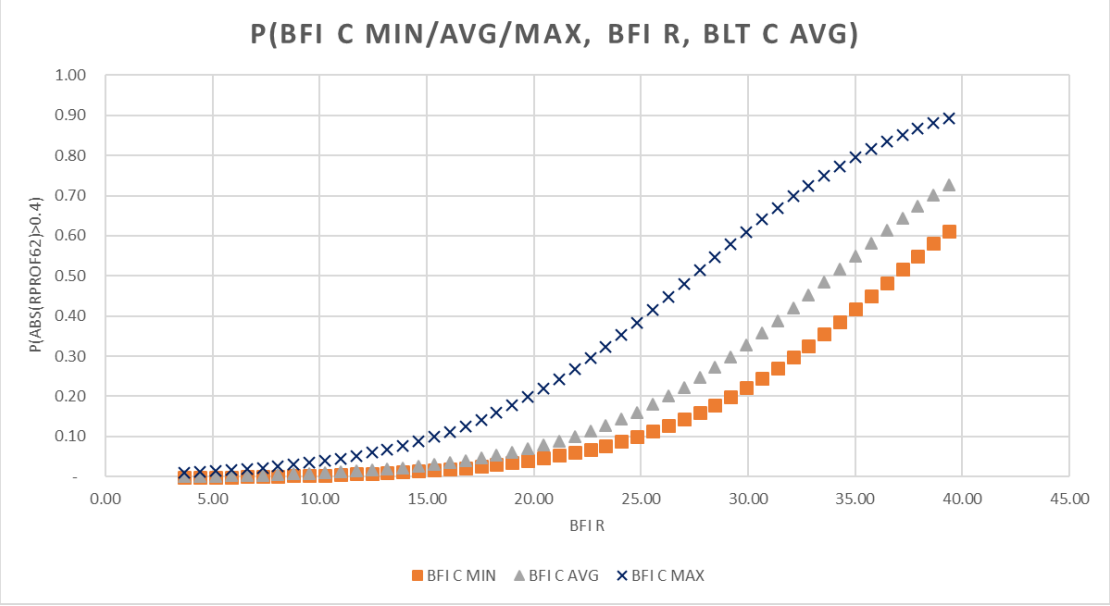


Figure B.1: Sensitivity plot - P(BFI C MIN/AVG/MAX, BFI R, BLT C avg)

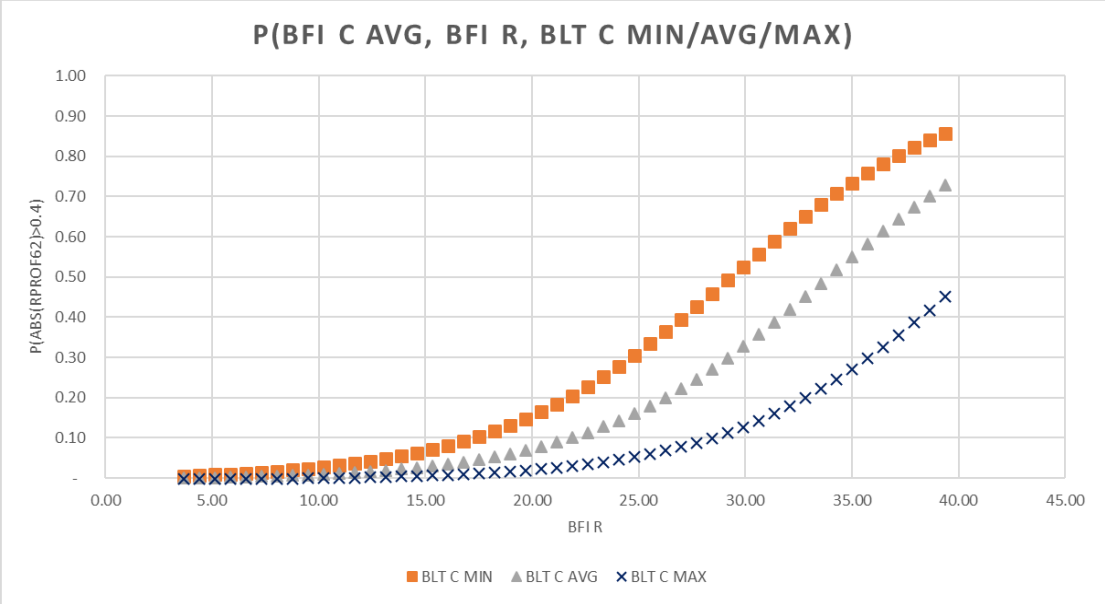


Figure B.2: Sensitivity plot - P(BFI C avg, BFI R, BLT C MIN/AVG/MAX)

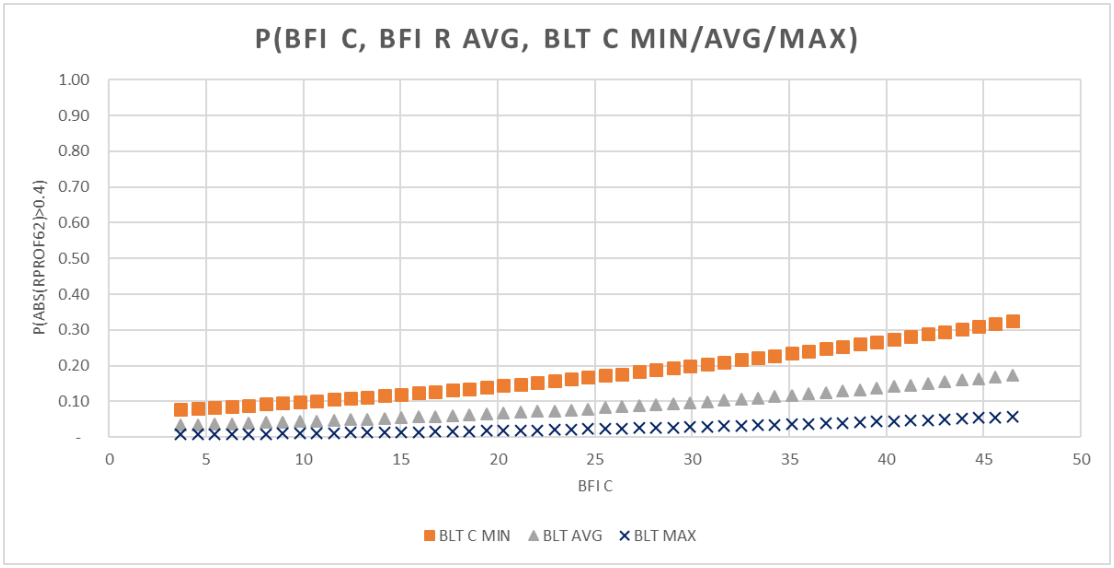


Figure B.3: Sensitivity plot - logistic regression Model 3.1 P(BFI C, BFI R avg, BLT C MIN/AVG/MAX)

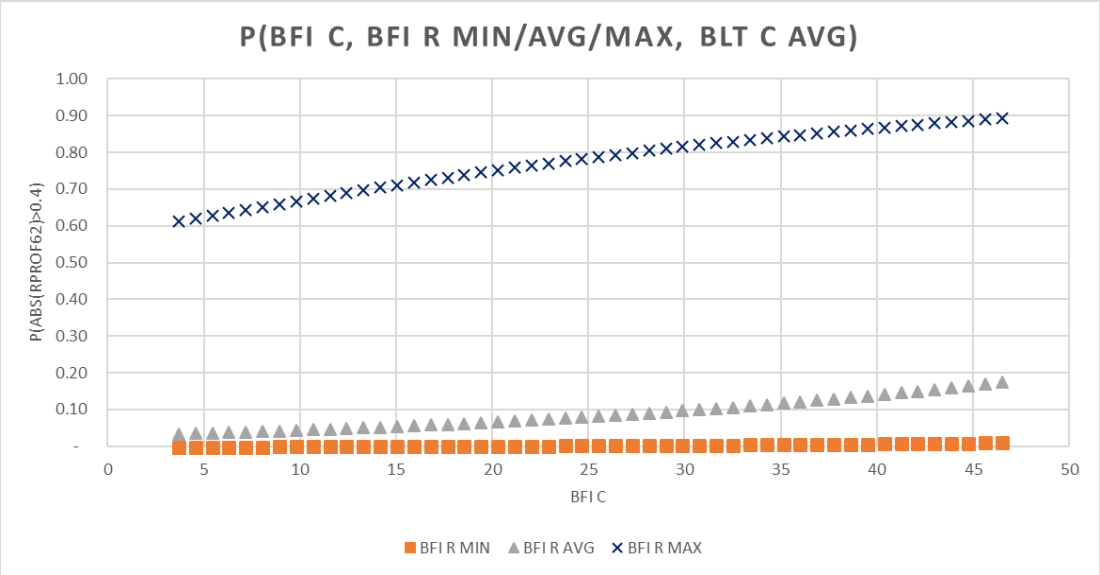


Figure B.4: Sensitivity plot - P(BFI C, BFI R MIN/AVG/MAX, BLT C avg)

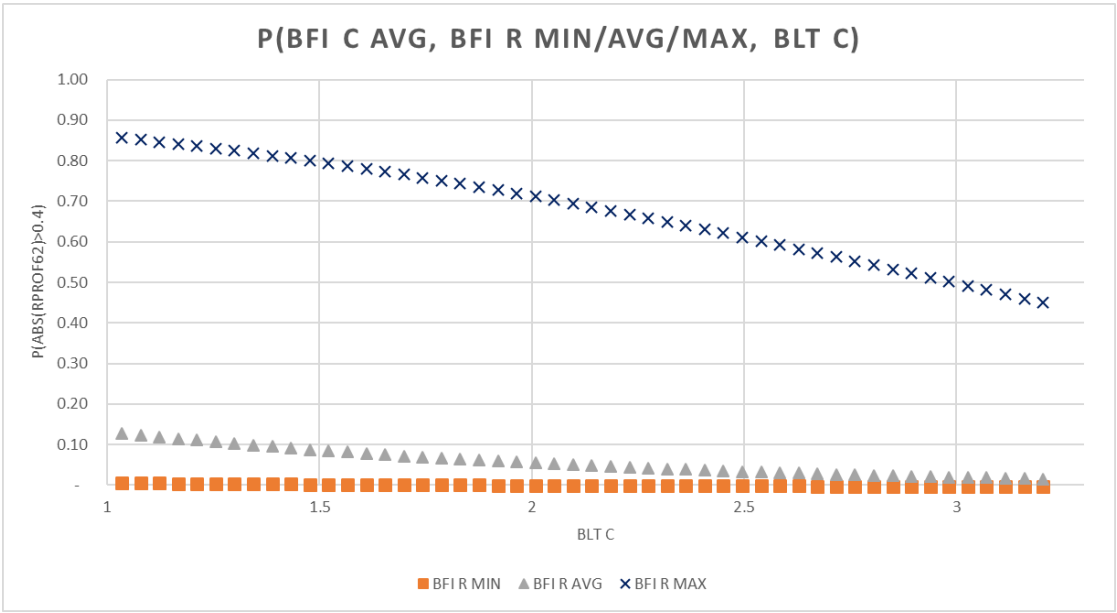


Figure B.5: Sensitivity plot - P(BFI C avg, BFI R MIN/AVG/MAX, BLT C)

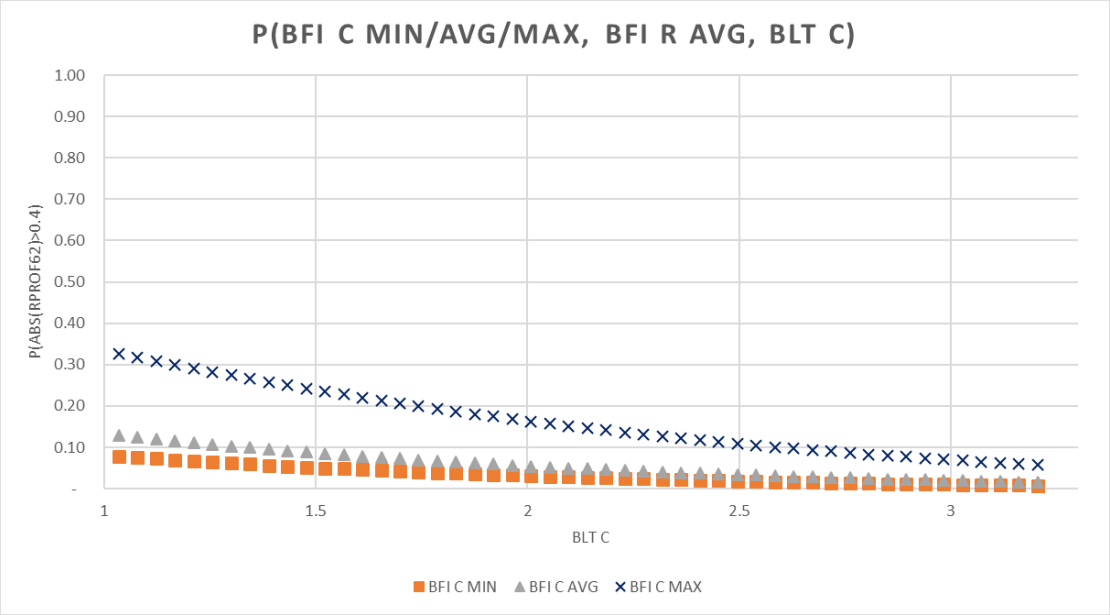


Figure B.6: Sensitivity plot - P(BFI C MIN/AVG/MAX, BFI R avg, BLT C)

B.1.2 Logistic Regression Model 3.1 Statistical Validation

```
> crossVal<- train(as.factor(`abs(Right Prof 62)>0.4`) ~
+                    Center+
+                    Right+
+                    `Center: Top of Layer`
+                    ,data = data1,
+                  family = binomial,  method = "glm", trControl = crossValSettings)
> crossVal
Generalized Linear Model

253 samples
  3 predictor
  2 classes: '0', '1'

No pre-processing
Resampling: Cross-Validated (10 fold, repeated 1 times)
Summary of sample sizes: 228, 227, 228, 229, 229, 227, ...
Resampling results:

  Accuracy   Kappa
  0.8657692  0.2046351
```

B.1.3 Logistic Regression Model 3.1 Confusion Matrix

```
> confusionMatrix(data = pred, data1$`abs(Right Prof 62)>0.4`)
Confusion Matrix and Statistics
          Reference
Prediction   0   1
         0 213  29
         1   4   7

               Accuracy : 0.8696
                 95% CI : (0.8217, 0.9085)
    No Information Rate : 0.8577
```

```
    P-Value [Acc > NIR] : 0.3327

                  Kappa : 0.2478
 Mcnemar's Test P-Value : 2.943e-05

            Sensitivity : 0.9816
            Specificity : 0.1944
         Pos Pred Value : 0.8802
         Neg Pred Value : 0.6364
             Prevalence : 0.8577
         Detection Rate : 0.8419
   Detection Prevalence : 0.9565
      Balanced Accuracy : 0.5880

       'Positive' Class : 0
```

B.1.4 Logistic Regression Model 3.1 ROC and AUC

```
> AUC
An object of class "performance"
Slot "x.name":
[1] "None"

Slot "y.name":
[1] "Area under the ROC curve"

Slot "alpha.name":
[1] "none"

Slot "x.values":
list()

Slot "y.values":
[[1]]
[1] 0.7798899
```

B.2 Logistic Regression Model 3.2 P(abs(Rprof62)>0.4) = f(BFI R, BLT C) and f(BFI R, BFI C, BLT C)

B.2.1 P(abs(Rprof62)>0.4) = f(BFI R, BFI C, BLT C)

```
> logit<-glm(data1$`abs(9/27/2016_Right Prof 62)>0.4` ~
+              data1$Center+
+              data1$Right+
+              data1$`Thickness Center`
+              ,family = binomial, data = data1)
> summary(logit)

Call:
glm(formula = data1$`abs(9/27/2016_Right Prof 62)>0.4` ~ data1$Center +
    data1$Right + data1$`Thickness Center`, family = binomial,
    data = data1)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-0.9677  -0.4063  -0.3487  -0.2559   2.2913

Coefficients:
                          Estimate Std. Error z value Pr(>|z|)
(Intercept)               37.09705 4269.64701   0.009    0.993
data1$Center              -2.01129  238.52749  -0.008    0.993
```

```
data1$Right                 0.12011    0.07979   1.505    0.132
data1$`Thickness Center`   -3.16552    2.70766  -1.169    0.242

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 49.829  on 93  degrees of freedom
Residual deviance: 46.422  on 90  degrees of freedom
AIC: 54.422

Number of Fisher Scoring iterations: 15
```

confidence interval: 2.5 and 95%:

```
> round(exp(cbind(estimate=coef(logit),confint(logit))),2)
Waiting for profiling to be done...
                         estimate 2.5 %   97.5 %
(Intercept)                  1.79  0.00 71794.62
data1$Right                  1.13  0.96     1.33
data1$`Thickness Center`     0.05  0.00     8.18
```

B.2.2 P(abs(Rprof62)>0.4) = f(BFI R, BLT C)

```
> logit<-glm(data1$`abs(9/27/2016_Right Prof 62)>0.4` ~
+              data1$Right+
+              data1$`Thickness Center`
+              ,family = binomial, data = data1)
> summary(logit)

Call:
glm(formula = data1$`abs(9/27/2016_Right Prof 62)>0.4` ~ data1$Right +
    data1$`Thickness Center`, family = binomial, data = data1)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-0.9866  -0.3922  -0.3399  -0.2545   2.2850

Coefficients:
                         Estimate Std. Error z value Pr(>|z|)
(Intercept)               0.58062    5.10543   0.114    0.909
data1$Right               0.12389    0.07964   1.556    0.120
data1$`Thickness Center` -2.95825    2.67466  -1.106    0.269

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 49.829  on 93  degrees of freedom
Residual deviance: 46.794  on 91  degrees of freedom
AIC: 52.794
Number of Fisher Scoring iterations: 5
```

B.2.3 Logistic Regression Model 3.2 Cross Validation and Performance Measurement

B.2.3.1 Logistic Regression Model 3.2 Confusion Matrix

```
> crossVal
Generalized Linear Model

94 samples
 2 predictor
 2 classes: '0', '1'

No pre-processing
Resampling: Cross-Validated (10 fold, repeated 1 times)
Summary of sample sizes: 85, 85, 84, 84, 84, 85, ...
Resampling results:

  Accuracy   Kappa
  0.9166667  0

> confusionMatrix(data = pred, data1$`abs(9/27/2016_Right Prof 62)>0.4`)
Confusion Matrix and Statistics

          Reference
Prediction  0  1
         0 87  7
         1  0  0

               Accuracy : 0.9255
                 95% CI : (0.8526, 0.9695)
    No Information Rate : 0.9255
    P-Value [Acc > NIR] : 0.59879

                  Kappa : 0
 Mcnemar's Test P-Value : 0.02334

            Sensitivity : 1.0000
            Specificity : 0.0000
         Pos Pred Value : 0.9255
         Neg Pred Value :    NaN
             Prevalence : 0.9255
         Detection Rate : 0.9255
   Detection Prevalence : 1.0000
      Balanced Accuracy : 0.5000

       'Positive' Class : 0
```

B.2.3.2 Logistic Regression Model 3.2 ROC and AUC

```
> AUC
An object of class "performance"
Slot "x.name":
[1] "None"

Slot "y.name":
[1] "Area under the ROC curve"

Slot "alpha.name":
[1] "none"

Slot "x.values":
list()

Slot "y.values":
[[1]]
[1] 0.727422

Slot "alpha.values":
```

```
list()
```

B.3 Logistic Regression Model 3.3 P(abs(Rprof62)>0.4) = f(BFIR, BFIC, BLTC)

B.3.1 Logistic Regression Model 3.3

```
> summary(logit)
Call:
glm(formula = data1$`abs(Right Prof 62)>0.4` ~ data1$Center +
    data1$Right + data1$`Center: Top of Layer` + data1$`Right: Top of Layer`,
    family = binomial, data = data1)
Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.4927  -0.4871  -0.3406  -0.2593   2.6621
Coefficients:
                            Estimate Std. Error z value Pr(>|z|)
(Intercept)                 -8.07644    2.06097  -3.919 8.90e-05 ***
data1$Center                 0.02457    0.02831   0.868  0.38541
data1$Right                  0.22252    0.04898   4.543 5.54e-06 ***
data1$`Center: Top of Layer` -3.09444    0.99263  -3.117  0.00182 **
data1$`Right: Top of Layer`   3.43919    1.30317   2.639  0.00831 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
(Dispersion parameter for binomial family taken to be 1)
    Null deviance: 207.01  on 252  degrees of freedom
Residual deviance: 162.38  on 248  degrees of freedom
AIC: 172.38
Number of Fisher Scoring iterations: 5
```

confidence interval: 2.5 and 95%:

```
> round(exp(cbind(estimate=coef(logit),confint(logit))),2)
Waiting for profiling to be done...
                            estimate 2.5 % 97.5 %
(Intercept)                     0.00  0.00   0.02
data1$Center                    1.02  0.97   1.08
data1$Right                     1.25  1.14   1.38
data1$`Center: Top of Layer`    0.05  0.01   0.31
data1$`Right: Top of Layer`    31.16  2.55 434.23
```

B.3.2 Logistic Regression Model 3.3 Cross Validation and Performance Measurement

B.3.2.1 Logistic Regression Model 3.3 Confusion Matrix

Settings for cross validation:

```
> crossVal<- train(as.factor(`abs(Right Prof 62)>0.4`) ~
+                    Center+
+                    Right+
+                    `Center: Top of Layer`+
+                    `Right: Top of Layer`
+                    ,data = data1,
+                  family = binomial,  method = "glm", trControl = crossValSettings)
> crossVal
Generalized Linear Model
```

```
253 samples
  4 predictor
  2 classes: '0', '1'

No pre-processing
Resampling: Cross-Validated (10 fold, repeated 1 times)
Summary of sample sizes: 227, 228, 228, 228, 227, 228, ...
Resampling results:

  Accuracy   Kappa
  0.8733462  0.2688603


> confusionMatrix(data = pred, data1$`abs(Right Prof 62)>0.4`)
Confusion Matrix and Statistics
          Reference
Prediction   0   1
         0 215  24
         1   2  12

               Accuracy : 0.8972
                 95% CI : (0.8531, 0.9318)
    No Information Rate : 0.8577
    P-Value [Acc > NIR] : 0.03938

                  Kappa : 0.435
 Mcnemar's Test P-Value : 3.814e-05

            Sensitivity : 0.9908
            Specificity : 0.3333
         Pos Pred Value : 0.8996
         Neg Pred Value : 0.8571
             Prevalence : 0.8577
         Detection Rate : 0.8498
   Detection Prevalence : 0.9447
      Balanced Accuracy : 0.6621

       'Positive' Class : 0
```

B.3.2.2 Logistic Regression Model 3.3 ROC and AUC

```
> AUC
An object of class "performance"
Slot "x.name":
[1] "None"
Slot "y.name":
[1] "Area under the ROC curve"
Slot "alpha.name":
[1] "none"
Slot "x.values":
list()
Slot "y.values":
[[1]]
[1] 0.7804019
Slot "alpha.values":
list()
```

Appendix C: Hierarchical Clustering Analysis of Histogram-Valued Data

C.1 Summary of All the Variables Normalized and Presented as Histograms

Below is a summary table of all the variables used in this appendix. The variables are normalized and presented as HVD. If the variable is highly distributed, each table shows the first five and the last five bins due to excessive length.

Table C.1: Absolute RProf62 variable as histogram-valued data in a table

	X	p
Bin_1	[-1.0742--0.81397)	0.09881
Bin_2	[-0.81397--0.55374)	0.2253
Bin_3	[-0.55374--0.29351)	0.2569
Bin_4	[-0.29351--0.033283)	0.07905
Bin_5	[-0.033283-0.22695)	0.07115
...	…	…
Bin_14	[2.3088 ; 2.569)	3.95E-06
Bin_15	[2.569 ; 2.8292)	3.95E-06
Bin_16	[2.8292 ; 3.0895)	3.95E-06
Bin_17	[3.0895 ; 3.3497)	0.003953
Bin_18	[3.3497 ; 3.6099)	0.02767

mean = 0.0043 std = 0.997

Table C.2: BFI Center variable as histogram-valued data in a table

	X	p
Bin_1	[-0.54423 ; -0.54423)	0.6996
Bin_2	[-0.54423 ; 1.5393)	0.253
Bin_3	[1.5393 ; 3.6375]	0.04743
mean = -0.1321 std = 0.8241		

Table C.3: BFI Right variable as histogram-valued data in a table

	X	p
Bin_1	[-0.58291 ; -0.58291)	0.7154
Bin_2	[-0.58291 ; 1.1936)	0.2451
Bin_3	[1.1936 ; 4.7717]	0.03953
mean = -0.2243 std = 0.8209		

Table C.4: BLT Center variable as histogram-valued data in a table

	X	p
Bin_1	[-1.7445--1.6405)	0.003953
Bin_2	[-1.6405--1.5364)	0.01976
Bin_3	[-1.5364--1.4323)	0.007905
Bin_4	[-1.4323--1.3283)	0.0751
Bin_5	[-1.3283--1.2242)	0.007905
...	...	...
Bin_34	[1.6897 ; 1.7937)	0.05929
Bin_35	[1.7937 ; 1.8978)	0.007905
Bin_36	[1.8978 ; 2.0019)	0.01186
Bin_37	[2.0019 ; 2.1059)	3.95E-06
Bin_38	[2.1059 ; 2.21)	0.01186
mean = -0.0015 std = 0.9948		

Table C.5: BLT Right variable as histogram-valued data in a table

	X	p
Bin_1	[-2.1518--2.0531)	0.003952
Bin_2	[-2.0531--1.9543)	3.95E-06
Bin_3	[-1.9543--1.8556)	3.95E-06
Bin_4	[-1.8556--1.7568)	3.95E-06
Bin_5	[-1.7568--1.6581)	3.95E-06
...	…	…
Bin_40	[1.6991 ; 1.7978)	3.95E-06
Bin_41	[1.7978 ; 1.8966)	0.003952
Bin_42	[1.8966 ; 1.9953)	0.07114
Bin_43	[1.9953 ; 2.0941)	3.95E-06
Bin_44	[2.0941 ; 2.1928)	0.03952
mean = 0.024 std = 0.999		

C.2 Histogram-Valued Data Plot as Histogram

This subsection, as part of the EDA and after variables normalization, presented for each normalized variable a histogram plot.

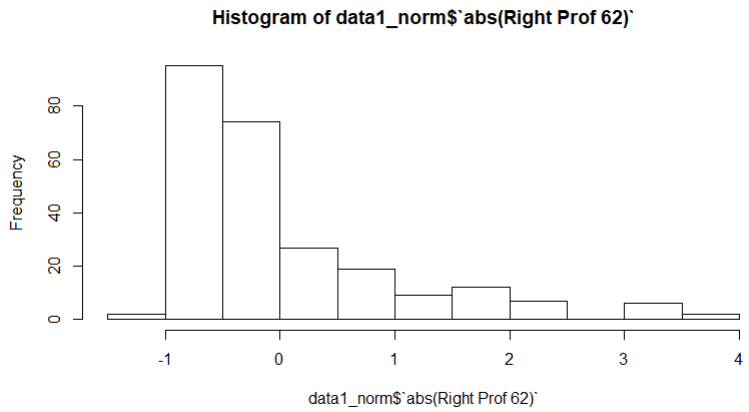


Figure C.1: abs(Rprof62) variable histogram

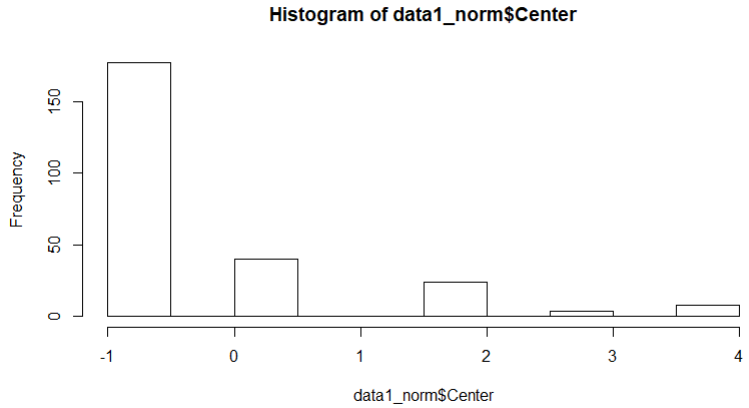


Figure C.2: BFI_C variable histogram

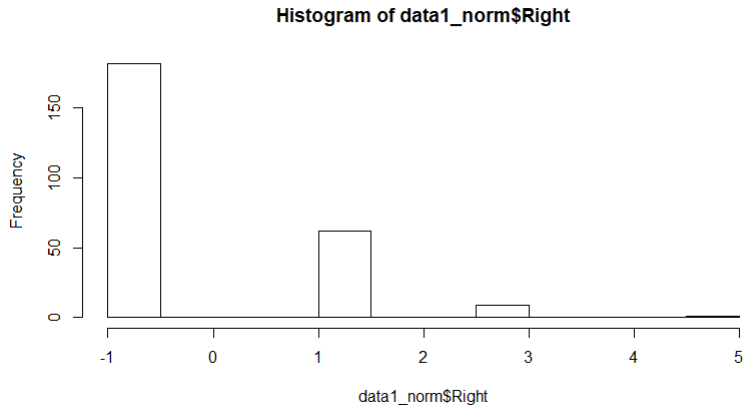


Figure C.3: BFI_R variable histogram

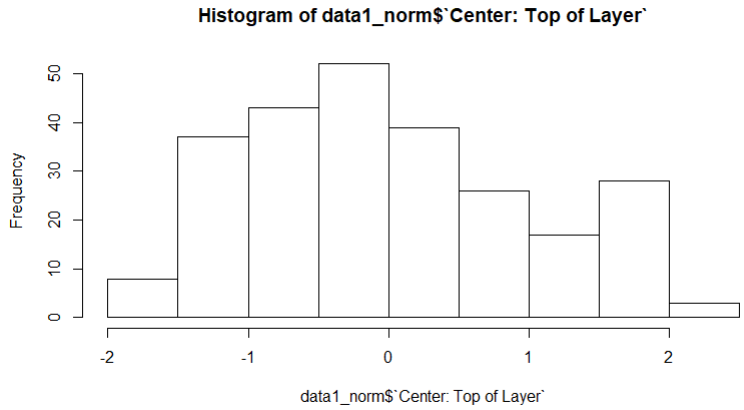


Figure C.4: C_Layer BFI_R variable histogram

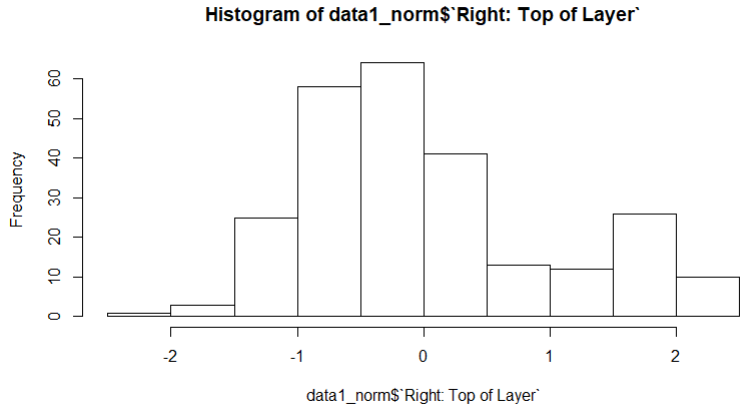


Figure C.5: R_Layer variable histogram

C.3 Histogram-Valued Data Plot as CDF

This subsection, as part of the EDA and after variables normalization, presented for each normalized variable a histogram plot.

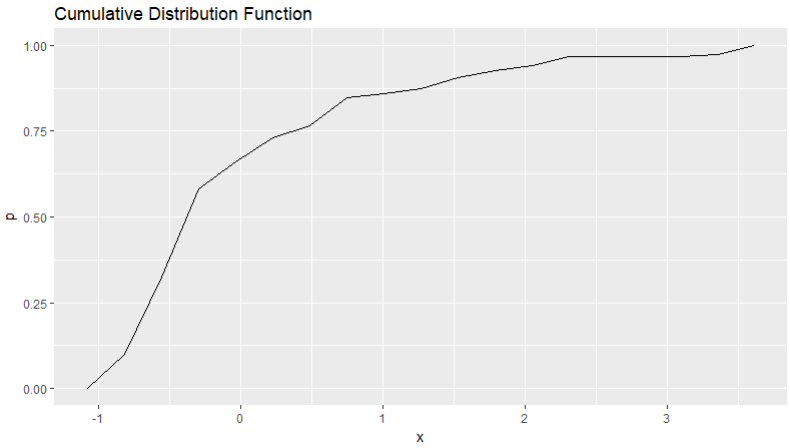


Figure C.6: abs(Rprof62) variable CDF plot

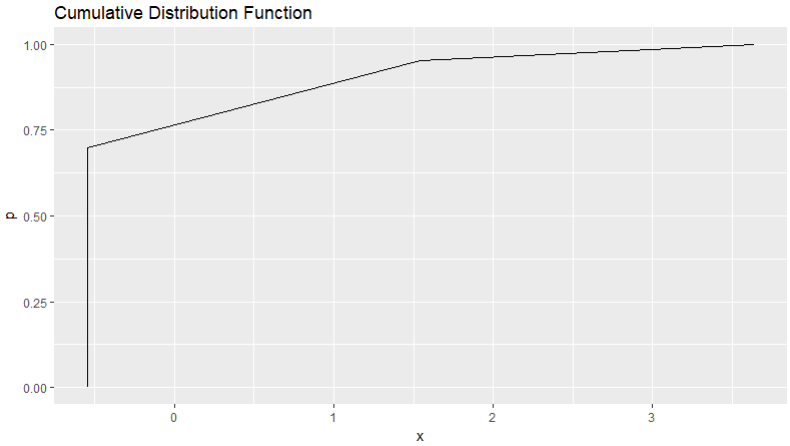


Figure C.7: BFI_C variable CDF plot

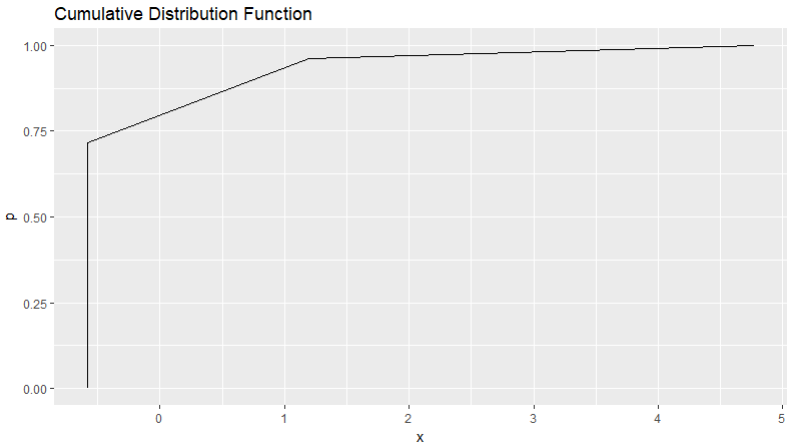


Figure C.8: BFI_R variable CDF plot

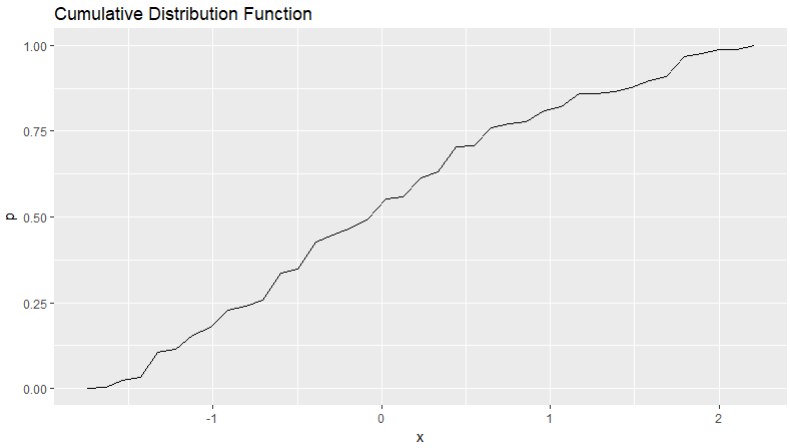


Figure C.9: C_Layer variable CDF plot

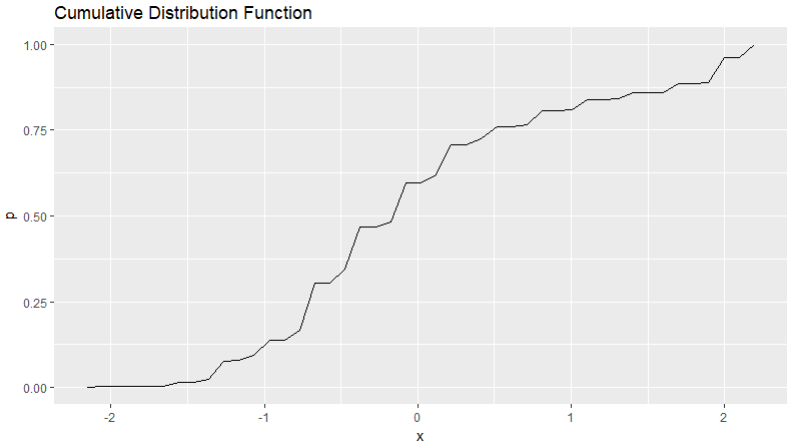


Figure C.10: R_Layer variable CDF plot

C.4 Calculation of Clusters and Dendrogram for Eight Different Linkages Methods

Figures C.11 to C.17 present the resulted dendrograms of the hierarchical structure of the data according to different linkage methods.

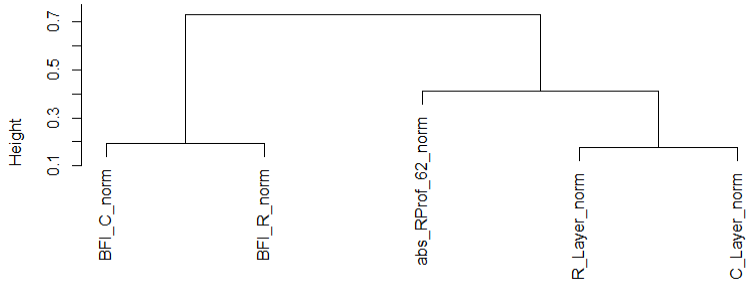


Figure C.11: Cluster dendrogram with maximum dissimilarity (complete linkage)

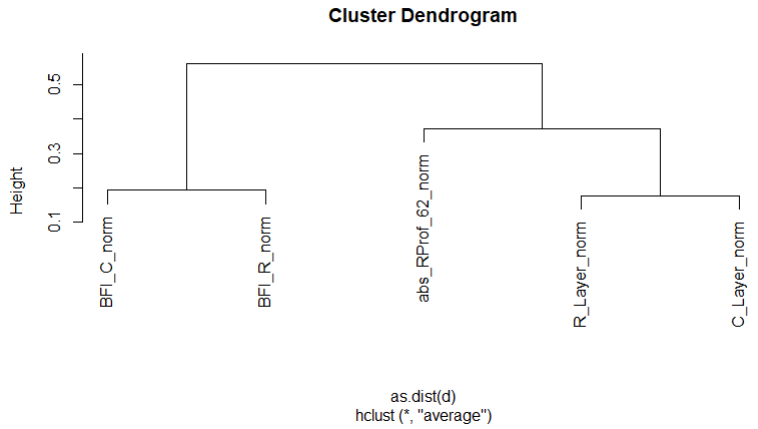


Figure C.12: Cluster dendrogram with average dissimilarity (average linkage)

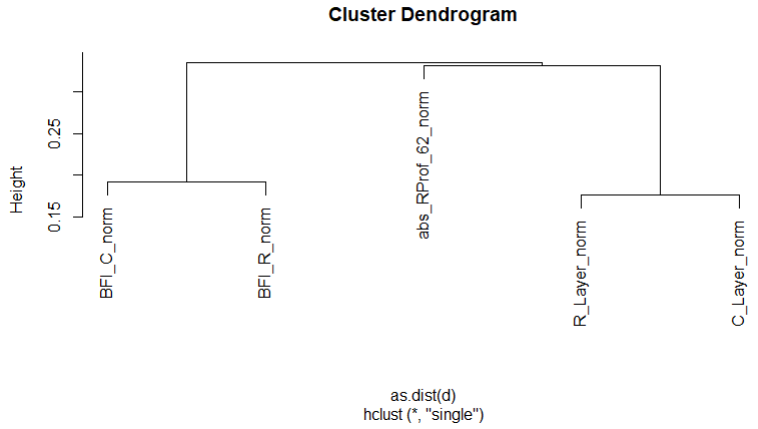


Figure C.13: Cluster dendrogram with minimum dissimilarity (single linkage)

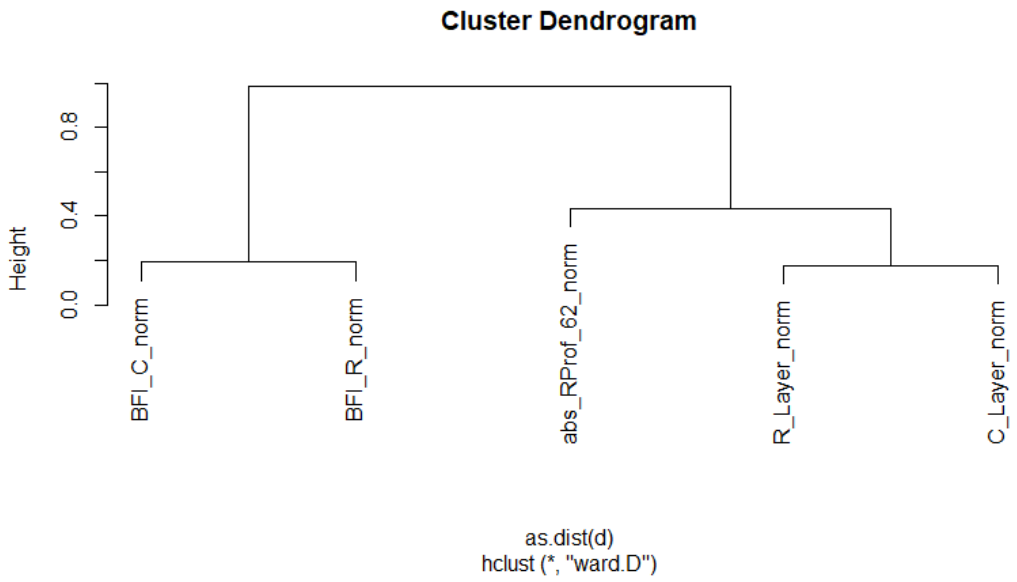


Figure C.14: Cluster dendrogram with ward linkage

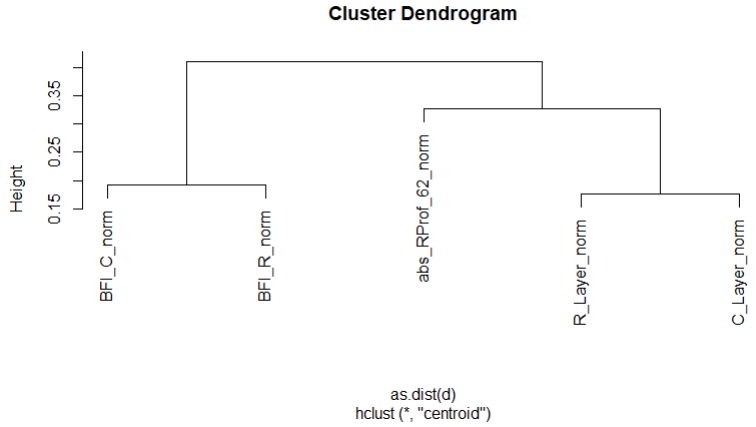


Figure C.15: Cluster dendrogram with centroid linkage

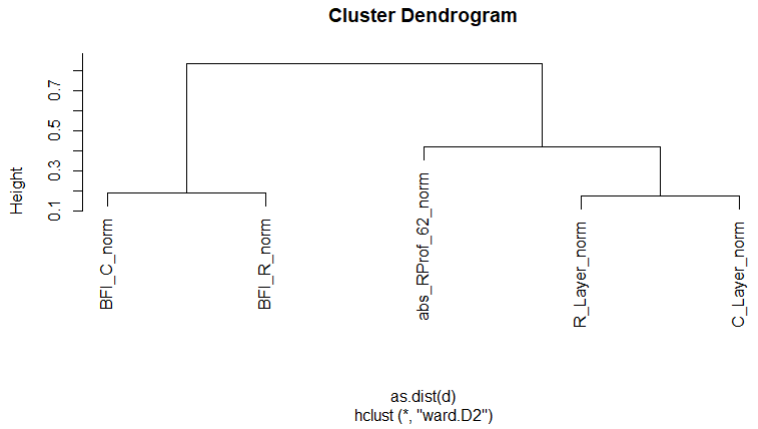


Figure C.16: Cluster dendrogram with Ward.D2 linkage

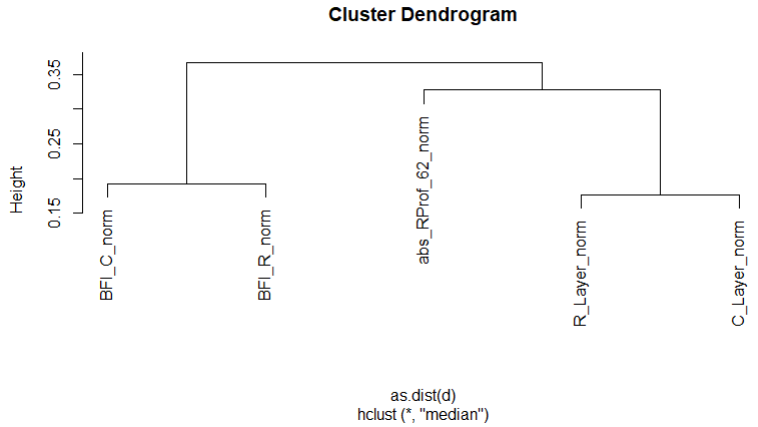


Figure C.17: Cluster dendrogram with median linkage

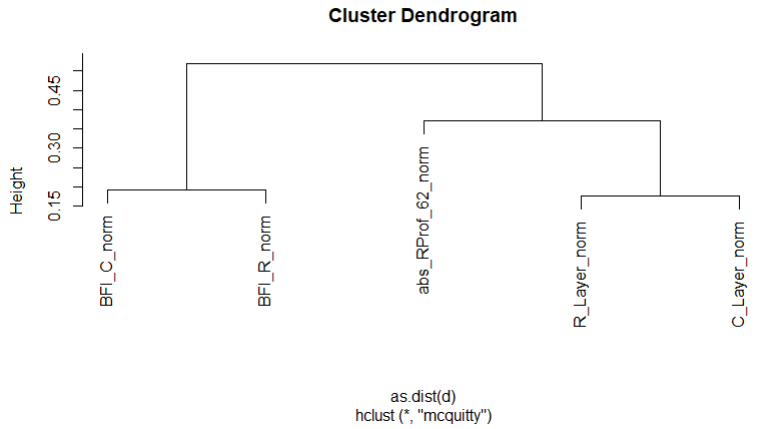


Figure C.18: Cluster dendrogram with McQuitty linkage

C.5 Cut the Tree for Four Groups

The following eight R software codes plots, presented in Table C.6, illustrate variables classification by groups. For example, first three lines of code plot Table C.6 belong complete linkage cut tree results, where that the parameters: abs_RProf_62_norm, BFI_C_norm, and BFI_R_norm create three different groups and the fourth group contains two parameters: R_Layer_norm and C_Layer_norm.

Table C.6: Cut tree for four groups plot summary

> cutree(HCA_HVD_complete,k = 4)				
abs_RProf_62_norm	BFI_C_norm	BFI_R_norm	R_Layer_norm	C_Layer_norm
1	2	3	4	4
> cutree(HCA_HVD_average,k = 4)				
abs_RProf_62_norm	BFI_C_norm	BFI_R_norm	R_Layer_norm	C_Layer_norm
1	2	3	4	4
> cutree(HCA_HVD_centroid,k = 4)				
abs_RProf_62_norm	BFI_C_norm	BFI_R_norm	R_Layer_norm	C_Layer_norm
1	2	3	4	4
> cutree(HCA_HVD_mcquitty,k = 4)				
abs_RProf_62_norm	BFI_C_norm	BFI_R_norm	R_Layer_norm	C_Layer_norm
1	2	3	4	4
> cutree(HCA_HVD_median,k = 4)				
abs_RProf_62_norm	BFI_C_norm	BFI_R_norm	R_Layer_norm	C_Layer_norm
1	2	3	4	4
> cutree(HCA_HVD_single,k = 4)				
abs_RProf_62_norm	BFI_C_norm	BFI_R_norm	R_Layer_norm	C_Layer_norm
1	2	3	4	4
> cutree(HCA_HVD_ward.D,k = 4)				
abs_RProf_62_norm	BFI_C_norm	BFI_R_norm	R_Layer_norm	C_Layer_norm
1	2	3	4	4

C.6 Logistic Regression models following HCA of HD

C.6.1 Logistic Regression Model 5.1 Plot from R Software

```
> logit<-glm(data1$`abs(Right Prof 62)>0.4` ~
+               data1$C+
+               data1$R+
+               data1$LC+
+               data1$C2+
+               data1$R2+
+               data1$LC2+
+               data1$CLC+
+               data1$RLC
+             ,family = binomial, data = data1)
> summary(logit)

Call:
glm(formula = data1$`abs(Right Prof 62)>0.4` ~ data1$C + data1$R +
```

```
    data1$LC + data1$C2 + data1$R2 + data1$LC2 + data1$CLC +
    data1$RLC, family = binomial, data = data1)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.6908  -0.5126  -0.3503  -0.2217   2.8842

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept)  1.7150559 11.9074993   0.144    0.885
data1$C      0.0174463  0.2760985   0.063    0.950
data1$R      0.1058282  0.5553687   0.191    0.849
data1$LC    -8.1665796  7.4685535  -1.093    0.274
data1$C2    -0.0046809  0.0035432  -1.321    0.186
data1$R2     0.0008867  0.0092530   0.096    0.924
data1$LC2    0.6866286  1.4040470   0.489    0.625
data1$CLC    0.1963029  0.0920449   2.133    0.033 *
data1$RLC    0.0130212  0.1517269   0.086    0.932
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 207.01  on 252  degrees of freedom
Residual deviance: 158.73  on 244  degrees of freedom
AIC: 176.73

Number of Fisher Scoring iterations: 6

LR model 1 confidence interval:

> round(exp(cbind(estimate=coef(logit),confint(logit))),2)
Waiting for profiling to be done...
            estimate 2.5 %       97.5 %
(Intercept)     5.56  0.00 8.478787e+10
data1$C         1.02  0.57 1.710000e+00
data1$R         1.11  0.35 3.190000e+00
data1$LC        0.00  0.00 9.010300e+02
data1$C2        1.00  0.99 1.000000e+00
data1$R2        1.00  0.98 1.020000e+00
data1$LC2       1.99  0.11 2.886000e+01
data1$CLC       1.22  1.04 1.510000e+00
data1$RLC       1.01  0.74 1.350000e+00
```

C.6.2 LR Model 5.2 Plot from R Software

```
> logit2<-glm(data1$`abs(Right Prof 62)>0.4` ~
+              data1$C+
+              data1$R+
+              data1$LR+
+              data1$C2+
+              data1$R2+
+              data1$LR2+
+              data1$CLR+
+              data1$RLR
+            ,family = binomial, data = data1)
> summary(logit2)

Call:
glm(formula = data1$`abs(Right Prof 62)>0.4` ~ data1$C + data1$R +
    data1$LR + data1$C2 + data1$R2 + data1$LR2 + data1$CLR +
    data1$RLR, family = binomial, data = data1)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.9116  -0.5180  -0.3347  -0.2245   3.2440
Coefficients:
```

```
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -4.003450  17.477024  -0.229  0.81882
data1$C      0.333616   0.295725   1.128  0.25926
data1$R     -0.570349   0.726061  -0.786  0.43214
data1$LR    -0.002112  11.677144   0.000  0.99986
data1$C2    -0.010419   0.003803  -2.739  0.00616 **
data1$R2     0.006826   0.010126   0.674  0.50023
data1$LR2   -2.528446   2.188817  -1.155  0.24802
data1$CLR    0.201844   0.128047   1.576  0.11495
data1$RLR    0.254390   0.218514   1.164  0.24435
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 207.01  on 252  degrees of freedom
Residual deviance: 155.88  on 244  degrees of freedom
AIC: 173.88

Number of Fisher Scoring iterations: 6
```

C.6.3 Logistic Regression Model 5.3 Plot from R Software

```
> logit3<-glm(data1$`abs(Right Prof 62)>0.4` ~
+               data1$C+
+               data1$R+
+               data1$LR
+               ,family = binomial, data = data1)
> summary(logit3)

Call:
glm(formula = data1$`abs(Right Prof 62)>0.4` ~ data1$C + data1$R +
    data1$LR, family = binomial, data = data1)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.5237  -0.5262  -0.3490  -0.3464   2.3864

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -7.36131    1.93037  -3.813 0.000137 ***
data1$C      0.06265    0.02466   2.540 0.011083 *
data1$R      0.18886    0.04503   4.194 2.74e-05 ***
data1$LR     0.04563    0.65117   0.070 0.944132
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 207.01  on 252  degrees of freedom
Residual deviance: 172.61  on 249  degrees of freedom
AIC: 180.61

Number of Fisher Scoring iterations: 5

LR Model 5.4 plot from R software.
> logit4<-glm(data1$`abs(Right Prof 62)>0.4` ~
+               data1$C+
+               data1$R+
+               data1$LC
+               ,family = binomial, data = data1)
> summary(logit4)

Call:
glm(formula = data1$`abs(Right Prof 62)>0.4` ~ data1$C + data1$R +
    data1$LC, family = binomial, data = data1)
Deviance Residuals:
```

```
    Min       1Q   Median       3Q      Max
-1.4819  -0.5349  -0.3641  -0.2875   2.6376

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -4.98175    1.67692  -2.971  0.00297 **
data1$C      0.03969    0.02707   1.466  0.14268
data1$R      0.18025    0.04385   4.111 3.94e-05 ***
data1$LC    -0.92397    0.56415  -1.638  0.10146
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 207.01  on 252  degrees of freedom
Residual deviance: 169.71  on 249  degrees of freedom
AIC: 177.71

Number of Fisher Scoring iterations: 5
```

C.7 Logistic Regression Following HCA of HD Cross Validation and Confusion Matrix

```
> crossVal<- train(as.factor(`abs(Right Prof 62)>0.4`) ~
+                    C+
+                    R+
+                    LC+
+                    C2+
+                    R2+
+                    LC2+
+                    CLC+
+                    RLC
+                  ,family = binomial, data = data1[,c(1:4,5,7,9,11,13)], method =
"glm", trControl = crossValSettings)
> crossVal
Generalized Linear Model

253 samples
  8 predictor
  2 classes: '0', '1'

No pre-processing
Resampling: Cross-Validated (10 fold, repeated 1 times)
Summary of sample sizes: 227, 228, 228, 227, 227, 228, ...
Resampling results:

  Accuracy   Kappa
  0.8728718  0.2398409
```

C.8 Logistic Regression Model 5.1 Confusion Matrix and Statistics

```
> confusionMatrix(data = pred, as.factor(data1$`abs(Right Prof 62)>0.4`))
Confusion Matrix and Statistics

          Reference
Prediction   0   1
         0 215  27
         1   2   9
               Accuracy : 0.8854
                 95% CI : (0.8395, 0.9219)
    No Information Rate : 0.8577
    P-Value [Acc > NIR] : 0.1192

                  Kappa : 0.3389
 Mcnemar's Test P-Value : 8.324e-06
```

```
            Sensitivity : 0.9908
            Specificity : 0.2500
         Pos Pred Value : 0.8884
         Neg Pred Value : 0.8182
             Prevalence : 0.8577
         Detection Rate : 0.8498
   Detection Prevalence : 0.9565
      Balanced Accuracy : 0.6204

       'Positive' Class : 0
```

C.9 Logistic Regression Model 5.1 Performance ROC Curve

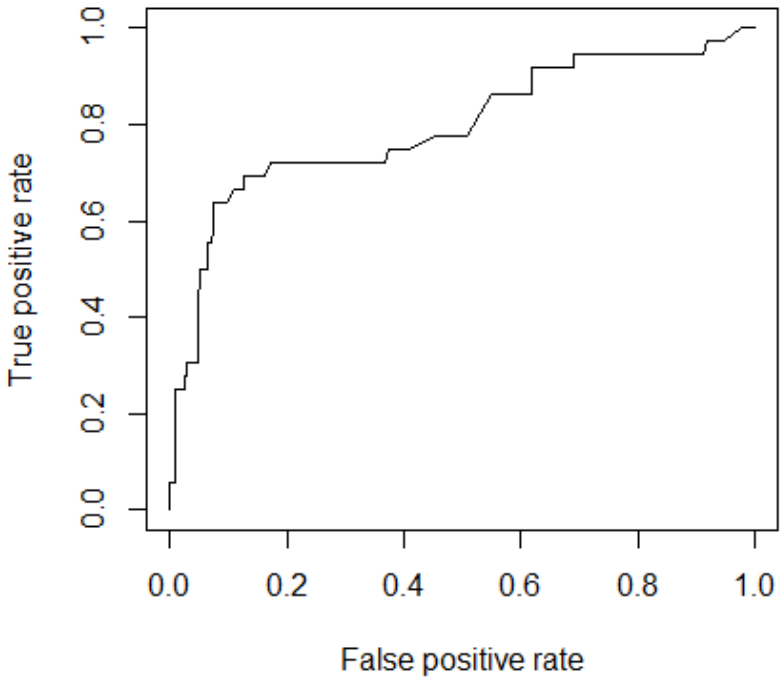


Figure C.19: Model 5.1 ROC curve

```
> AUC
An object of class "performance"
Slot "x.name":
[1] "None"

Slot "y.name":
[1] "Area under the ROC curve"

Slot "alpha.name":
[1] "none"

Slot "x.values":
list()

Slot "y.values":
[[1]]
[1] 0.7951229
```

```
Slot "alpha.values":
list()
```

C.10 Model 5.1 Sensitivity Analysis Two-Dimensional Plots

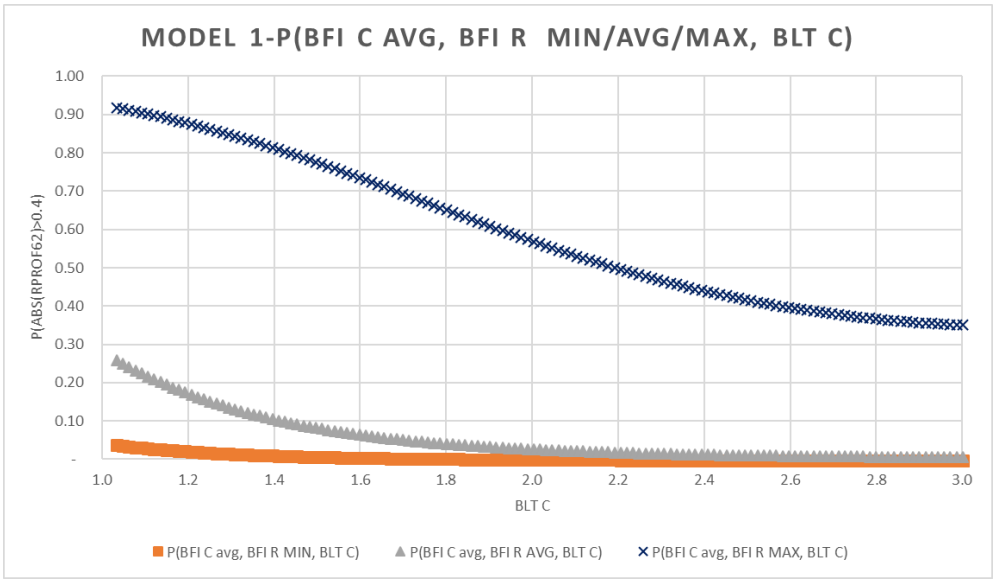


Figure C.20: Model 5.1 - P(BFI C avg, BFI R MIN/AVG/MAX, BLT C)

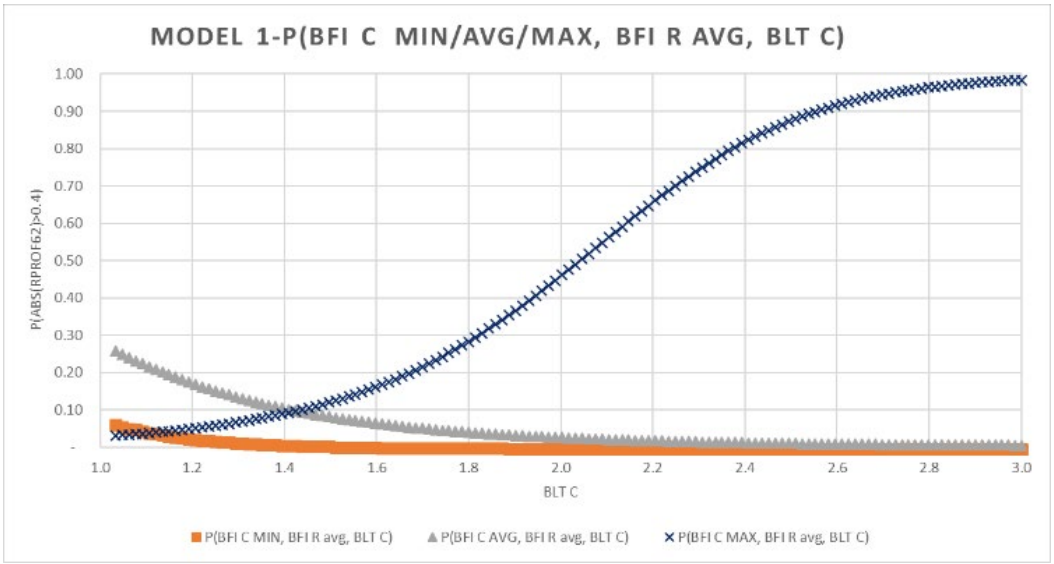


Figure C.21: Model 5.1 - P(BFI C MIN/AVG/MAX, BFI R avg, BLT C)

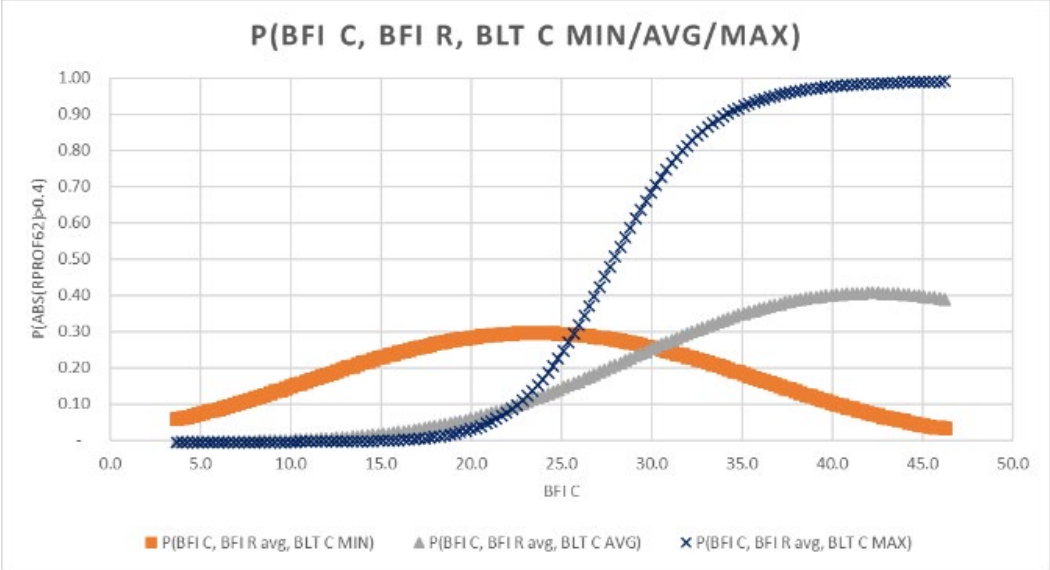


Figure C.22: Model 5.1 - P(BFI C, BFI R avg, BLT C MIN/AVG/MAX)

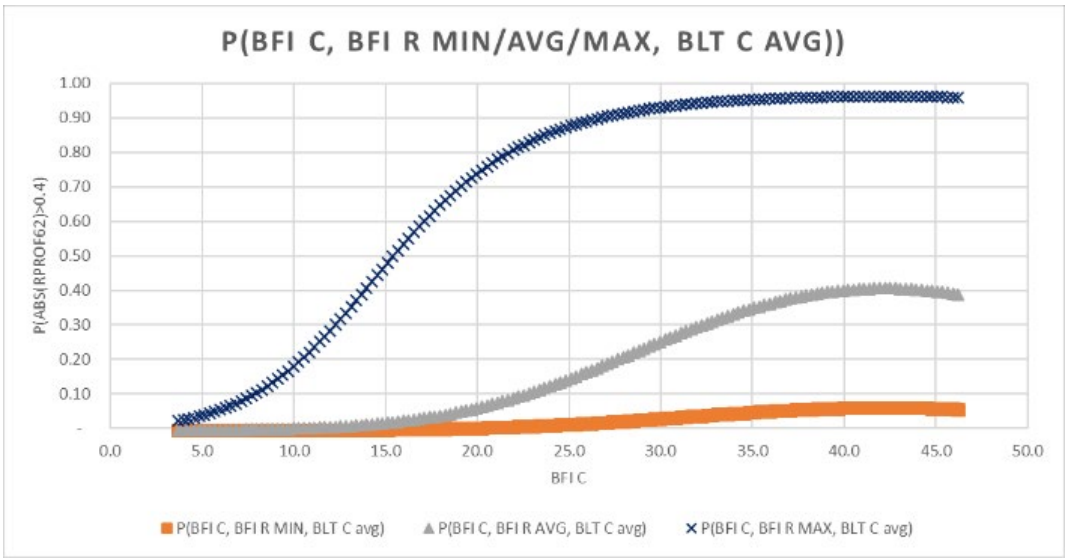


Figure C.22: Model 5.1 - P(BFI C, BFI R MIN/AVG/MAX, BLT C avg)

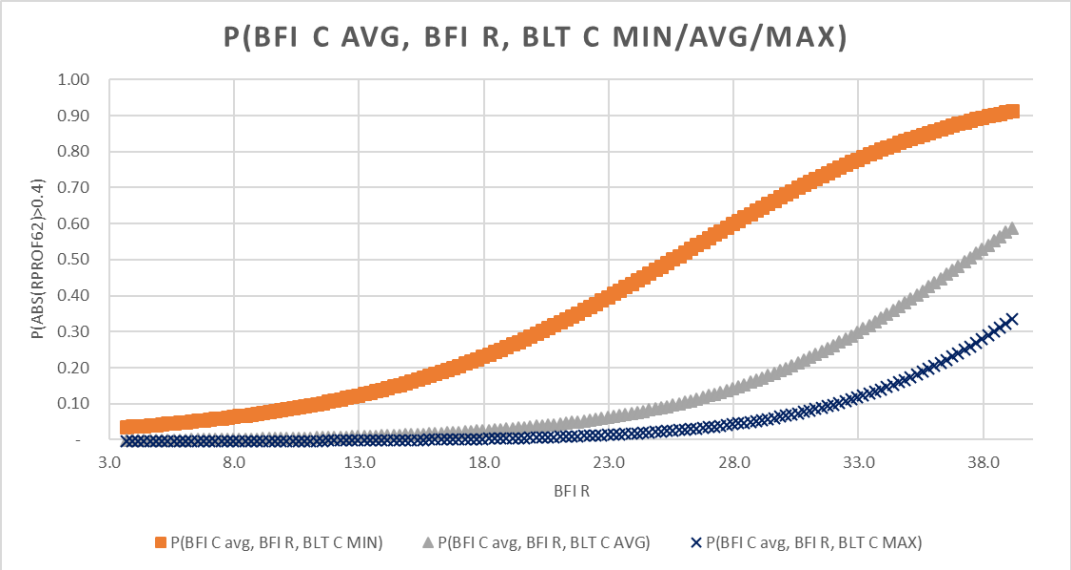


Figure C.23: Model 5.1 - P(BFI C avg, BFI R, BLT C MIN/AVG/MAX)

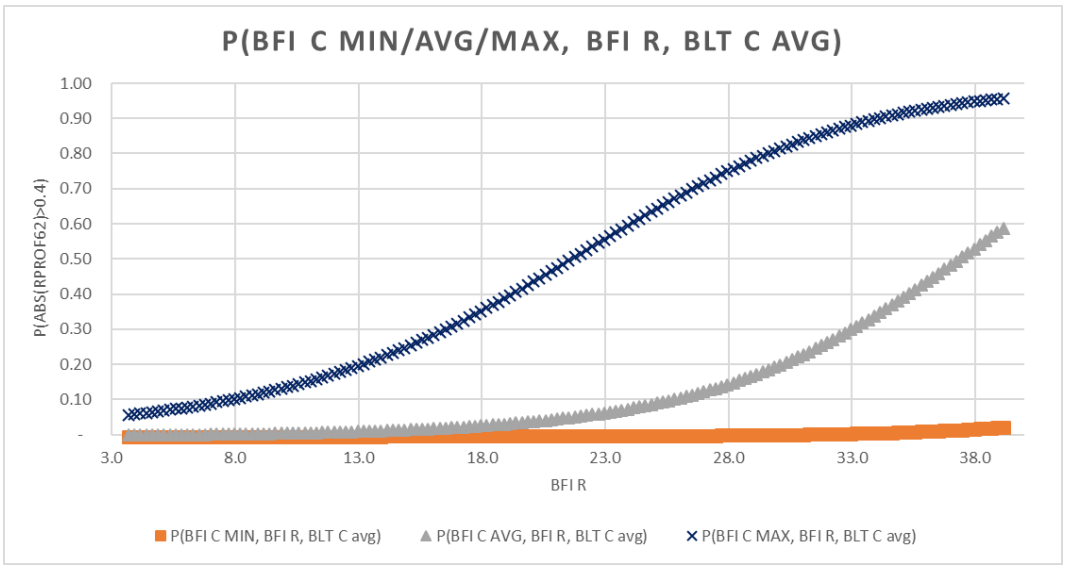


Figure C.24: Model 5.1 - P(BFI C MIN/AVG/MAX, BFI R, BLT C avg)

Abbreviations and Acronyms

ACRONYMS	EXPLANATION
AIC	Akaike Information Criterion
AUC	Area Under Curve
BFI	Ballast Fouling Index
BLT	Ballast Layer Thickness
BTI	Ballast Thickness Index
BNSF	Burlington Norther Santa Fe Railway
CP	Canadian Pacific Railway
CDF	Cumulative Distribution Function
EDA	Exploratory Data Analysis
FPR	False Positive Rate
FRA	Federal Railroad Administration
FDL	Free Draining Layer
GLM	Generalized Linear Model
GPR	Ground Penetrating Radar
HCA	Hierarchical Clustering Analysis
HVD	Histogram Valued Data
KDE	Kernel Density Estimation
LRI	Layer Roughness Index
LIDAR	Light Detection and Ranging
MLE	Likelihood Estimation
LR	Logistic Regression
MP	Milepost
MISSQ	Minimal Increase of Sum-of-Squares
NS	Norfolk Southern Corporation
QQ	Quantile-Quantile
ROC	Receiver Operating Characteristic
SD	Standard Deviation
SDA	Symbolic Data Analysis
MRail	Track Deflection Measurements
TQI	Track Quality Indices

ACRONYMS	EXPLANATION
TPR	True Positive Rate
UP	Union Pacific Railroad