Derailment Prediction Models for Canada's Rail Network

TAVIA WING TUNG CHOW

A THESIS SUBMITTED TO
THE FACULTY OF GRADUATE STUDIES
IN PARTIAL FULFILLMENT OF THE REQUIREMENTS
FOR THE DEGREE OF
MASTER OF APPLIED SCIENCE IN CIVIL ENGINEERING

GRADUATE PROGRAM IN CIVIL ENGINEERING
YORK UNIVERSITY
TORONTO, ONTARIO

March 2020

Abstract

Train related accidents, particularly derailments, can lead to severe consequences especially when they involve injuries or fatalities or involve dangerous goods that might result in environmental impacts. A literature review found that rail safety assessment and derailment prediction models have often been constrained by aggregated data which may yield inaccurate assessments of the safety performance of a rail network by, for example, failing to consider segment-level characteristics.

This study focused on the development of segment-level derailment prediction models for Canada's rail network using negative binomial and logistic regression modelling methods. The study used a network screening process to identify key segments of derailment concern. A thorough quantitative review of the models' results and performance was conducted to understand the predictive capabilities and applications of the models in derailment prediction. The analytical approach and findings in this study have strong implications for advancing research on rail safety in North America.

Keywords: rail safety, track segment, network screening, safety performance function, derailments, risk prediction, negative binomial regression, logistic regression.

Dedication

I want to dedicate my thesis work to my husband and new baby girl for giving me the motivation to pursue and finish my research. Also, to my family who provided tremendous support in every way possible in this journey. I would never be able to make this accomplishment without your love and understanding.

Acknowledgements

I would like to thank my supervisor, Dr. Peter Park for his continuous guidance, understanding and support. His professional knowledge and experience in industry were extremely helpful for my research. I would also like to express my appreciation to Dr. Kevin Gingerich and Dr. Mehdi Nourinejad for their time reviewing this thesis. I am thankful for Tayyab Shah and Seun Oluwajana's contribution towards the model development. Additionally, I would like to thank Transport Canada for partially funding my research as part of the Collaborative Research and Development for Improved Rail Safety in Canada project.

Table of Contents

Abstract ..... **ii**

Dedication ..... **iii**

Acknowledgements ..... **iv**

List of Tables ..... **viii**

List of Figures ..... **x**

CHAPTER 1. INTRODUCTION **1**

1.1. Research Overview 1

1.2. Safety Network Screening 6

1.3. Research Goal and Objectives 6

1.4. Scope and Structure of Thesis 7

1.5. Chapter Summary 7

CHAPTER 2. LITERATURE REVIEW **9**

2.1. Railway Safety 9

2.2. Derailment Prediction Models and Factors 14

2.3. Modelling Methods 21

2.4. Chapter Summary 25

CHAPTER 3. STUDY DATA **27**

3.1. Data Sources 27

3.2. Data Preparation and Validation 28

3.3. Descriptive Data Analysis 40

3.4. Chapter Summary 48

CHAPTER 4. MODELLING METHODOLOGY **49**

4.1. Input Data Preparation 49

4.2. Model Descriptions 51

4.3. Model Form 54

4.4. Empirical Bayes (EB) Approach 55

4.5. Model Calibration and Validation 56
4.6. Goodness-of-fit Tests 56
4.7. Variable Selection 61
4.8. Chapter Summary 62
CHAPTER 5. DEVELOPMENT AND ANALYSIS OF PREDICTION MODELS 64
5.1. Derailment Prediction Model for Canada 65
5.2. Derailment Prediction Model for Eastern Canada 68
5.3. Derailment Prediction Model for Western Canada 71
5.4. Derailment Prediction Model for the Canadian National Railway 75
5.5. Derailment Prediction Model for the Canadian Pacific Railway 78
5.6. Chapter Summary 81
CHAPTER 6. SAFETY NETWORK SCREENING 82
6.1. Hotspot Analysis for Canada 82
6.2. Hotspot Analysis for Eastern Canada 84
6.3. Hotspot Analysis for Western Canada 86
6.4. Hotspot Analysis for Canadian National Railway 88
6.5. Hotspot Analysis for Canadian Pacific Railway 89
6.6. Chapter Summary 91
CHAPTER 7. DEVELOPMENT OF BINOMIAL LOGISTIC MODELS 92
7.1. Logit Model Development 92
7.2. Model Calibration and Validation 93
7.3. Goodness-of-fit Tests for Logit Model 93
7.4. Logit Model for Canada 97
7.5. Logit Model for Eastern Canada 100
7.6. Logit Model for Western Canada 104
7.7. Logit Model for Canadian National Railway 107
7.8. Logit Model for Canadian Pacific Railway 111

7.9. Chapter Summary 114
CHAPTER 8. COMPARISON OF PREDICTION RESULTS 116
8.1. Tetrachoric Correlation Analysis 116
8.2. Comparison of Network Screening Results 117
8.3. Segment Identification Methods 118
8.4. Discussion of Model Performances 119
8.5. Chapter Summary 121
CHAPTER 9. CONCLUSIONS AND RECOMMENDATION 122
9.1. Research Contribution 123
9.2. Research Limitations 124
9.3. Future Study 125
REFERENCES 127
APPENDICES 133
Appendix A: Negative Binomial Models – Candidate Models 134
Appendix B: Model Forms and Cure Plots for Shortlisted Models 154
Appendix C: Variance Inflation Factors (VIFs) for Selected Model 167
Appendix D: Logit Models – Candidate Models 168
Appendix E: Converted Prediction Outcomes for Negative Binomial Models 188
Appendix F: Key Segments of Concerns from Both Negative Binomial and Logit Models 193

List of Tables

Table 1: Top 10 Common Causes for Main-track Derailments in US, 2001-2010 (Liu et al., 2012) 4
Table 2: Top 10 Common Causes for Main-track Derailments in Canada, 2001-2014 (Leishman, 2017) .. 4
Table 3: Estimated Economic Costs for Lac-Mégantic Derailment (Goodman and Rowan, 2013) 5
Table 4: Summary of Analysis Periods of Rail Safety Research 19
Table 5: Description of Study Data 27
Table 6: Track Classification and Number of Features 30
Table 7: Annual Main-Track Derailment Occurrence, 1999 to 2018 32
Table 8: A Screenshot of Grade Crossing Inventory Database 36
Table 9: Description of Rail Network Features (NRC, 2012) 37
Table 10: Segmentation Method Analysis Results and Comparison 38
Table 11: Number of Railcars Involved in Derailments, 1999-2018 43
Table 12: Derailments by Track Owner, 1999-2018 45
Table 13: Model Variable Names and Descriptions 49
Table 14: Summary Statistics for Independent Variables 50
Table 15: Rail System Characteristics of Eastern and Western Canada 52
Table 16: Correlation Matrix of Model Variables 61
Table 17: Summary of GOF Measures and Preferred Values 63
Table 18: Summary Statistics for Key Independent Variables – Canada Model 65
Table 19: Summary of Goodness-of-fit Test Results for Shortlisted Canada Models 66
Table 20: Negative Binomial Model Results for Canada 67
Table 21: Summary Statistics for Key Independent Variables – Eastern Canada 69
Table 22: Summary of Goodness-of-fit Tests for Shortlisted Eastern Canada Models 70
Table 23: Negative Binomial Model Results for Eastern Canada 71
Table 24: Summary Statistics for Key Independent Variables – Western Canada 72
Table 25: Summary of Goodness-of-fit Tests for Shortlisted Western Canada Models 73
Table 26: Negative Binomial Model Results for Western Canada 74
Table 27: Summary Statistics for Key Independent Variables in CN Model 75
Table 28: Summary of Goodness-of-fit Tests for Shortlisted CN Models 76
Table 29: Negative Binomial Model Results for CN 77
Table 30: Summary Statistics for Key Explanatory Variables in CN Model 78
Table 31: Summary of Goodness-of-fit Tests for Shortlisted CP Models 79
Table 32: Negative Binomial Model Results for CP 80

Table 33: Safety Network Screening Results of Canada 82
Table 34: Safety Network Screening Results for Eastern Canada 85
Table 35: Safety Network Screening Results for Western Canada 87
Table 36: Safety Network Screening Results for CN 88
Table 37: Safety Network Screening Results for CP 90
Table 38: Segments Identified as Hotspots in Multiple Models 91
Table 39: A Sample Classification Table 95
Table 40: Comparison of Candidate Logit Mode ls for Canada 97
Table 41: Logit Model Results for Canada 98
Table 42: Classification Table for Canada Model 98
Table 43: Comparison of Shortlisted Logit Models for Eastern Canada 100
Table 44: Logit Model Results for Eastern Canada 101
Table 45: Classification Table for Eastern Canada Model 102
Table 46: Comparison of Candiate Logit Models for Western Canada 104
Table 47: Logit Model Results for Western Canada 105
Table 48: Classification Table for Western Canada Model 106
Table 49: Comparison of Candidate Logit Models for CN 108
Table 50: Logit Model Results for CN 108
Table 51: Classification Table for CN model 109
Table 52: Comparison of Candidate Logit Models for CP 111
Table 53: Logit Model Results for CP 112
Table 54: Classification Table for CP model 113
Table 55: Tetrachoric Correlation Analysis Results 116
Table 56: Selected Statistics of Track Segments with Predicted Derailments Risks 117
Table 57: Selected Statistics of Segments with Derailments Risks from NB and Logit Models 118
Table 58: Summary of Selected Statistics for Difference Segment Identification Methods 119
Table 59: Comparison of Model Statistics Based on Cut-off Values 120

List of Figures

Figure 1: Derailment Rate per Million Train Miles in Canada and the United States 2
Figure 2: Number of Derailments and Total Train Miles in the US (2009-2018) 9
Figure 3: Iso-line Graphs for Derailments (Wang et al., 2017, p. 822) 11
Figure 4: Number of Main-Track Derailments and Million Train Miles in Canada (2009-2018) 11
Figure 5: Main-track and Non-main-track Derailments in Canada (2009-2018) 12
Figure 6: Train Accidents by Type (2009-2018) 12
Figure 7: Distribution of Railcars Involved in Derailments (2009-2018) 13
Figure 8: Expected Accidents Relative to Train Length 15
Figure 9: Sample of RTM effect (Liu and Rodriguez, 2016, p. 4) 17
Figure 10: Rail Operating Revenues and Million Main-Track Train-Miles, 1999-2017 20
Figure 11: Empirical Bayes Method 23
Figure 12: Dataset Development Process 29
Figure 13: Rail Network in Canada 30
Figure 14: Example of Multi-Lane Tracks into a Single/Centreline Track Conversion 31
Figure 15: Derailments by Province, 1999 - 2018 33
Figure 16: Excerpt from Derailments Table 34
Figure 17: Derailments Geocoded in Reference to the Rail Track Centreline 35
Figure 18: Sample Grade Crossing Locations with Total Daily Train Traffic 36
Figure 19: Distribution of Segment Length by Segment Method 37
Figure 20: Annual Number of Main-Track Derailments, 1999-2018 40
Figure 21: Number of Main-Track Derailments by Month, 1999-2018 41
Figure 22: Number of Main-Track Derailments by Season, 1999-2018 41
Figure 23: Derailments by Day of the Week 42
Figure 24: Hourly and Daily Trends of Derailments, 1999-2018 42
Figure 25: Distribution of Number of Railcars Involved in Derailments, 1999-2018 43
Figure 26: Number of Derailments by Province and Territory, 1999-2018 44
Figure 27: Derailment Rate per Kilometre Track by Province, 1999-2018 44
Figure 28: Derailment and Train Volumes on Segments (1999-2018) 46
Figure 29: Maximum Train Speeds (mph) Recorded at At-Grade Crossings in Canada, 2019 47
Figure 30: Derailment and Segment Train Speeds 47
Figure 31: Box Plot for each Independent Variable 50
Figure 32: Number of Derailments by Number of Track Segments, 2009-2018 51

Figure 33: System Characteristics for Eastern and Western Canada 52
Figure 34: Rail Track by Ownership 53
Figure 35: Derailments per Thousand Kms of Track Owned (2009-2018) 53
Figure 36: Sample Cumulative Residual (CURE) Plot 60
Figure 37: Sample CURE Plot Showing a Poor Fitting Model 60
Figure 38: Organization of Model Analysis Discussion 64
Figure 39: Track Segments in Canada 65
Figure 40: Track Segments in Canada Model 66
Figure 41: Cumulative Plot of Negative Binomial Model for Canada 67
Figure 42: Track Segments in Eastern Canada Model 68
Figure 43: Track Segments in Eastern Canada Model 69
Figure 44: Cumulative Plot of Negative Binomial Model for Eastern Canada 70
Figure 45: Track Segments in Western Canada 72
Figure 46: Track Segments in Western Canada Model 73
Figure 47: Cumulative Plot of Negative Binomial Model for Western Canada 74
Figure 48: Track Segments Owned by CN 75
Figure 49: Track Segments in CN Model 76
Figure 50: Cumulative Plot of Negative Binomial Model for CN 77
Figure 51: Track Segments Owned by CP 78
Figure 52: Track Segments in the CP Model 79
Figure 53: Cumulative Plot of Negative Binomial Model for CP 80
Figure 54: Top 10 Segments with Highest Derailment Risks in Canada 83
Figure 55: Track Owners of Top 10 Segments with Highest Derailment Risks in Canada 84
Figure 56: Top 10 Segments with Highest Derailment Risks of Eastern Canada Model 85
Figure 57: Track Owners of Top 10 Segments with Highest Derailment Risks in Eastern Canada 86
Figure 58: Top 10 Segments with Highest Derailment Risks in Western Canada 87
Figure 59: Track Owner of Top 10 Segments with Highest Derailment Risks in Western Canada 88
Figure 60: Top 10 Segments with Highest Derailment Risks for CN 89
Figure 61: Top 10 Segments with Highest Derailment Risks for CP 90
Figure 62: Sample ROC Curve 96
Figure 63: ROC Curve for Canada Model 99
Figure 64: Track Segments with Observed Derailment(s) 99
Figure 65: Track Segments Classified with Derailment Risk for Canada 100
Figure 66: ROC Curve for Eastern Canada Model 103

Figure 67: Track Segments with Observed Derailment(s) in Eastern Canada 103
Figure 68: Track Segments Classified with Derailment Risk in Eastern Canada 104
Figure 69: ROC Curve for Western Canada Model 106
Figure 70: Track Segments with Observed Derailment(s) in Western Canada 107
Figure 71: Track Segments Classified with Derailment Risk in Western Canada 107
Figure 72: ROC Curve for Eastern Canada Model 110
Figure 73: Track Segments with Observed Derailment(s) on CN Railway Network 110
Figure 74: Track Segments Classified with Derailment Risk on CN Railway Network 111
Figure 75: ROC Curve for CP Model 113
Figure 76: Track Segments with Observed Derailment(s) on CP Railway Network 114
Figure 77: Track Segments Classified with Derailment Risk for CP Model 114
Figure 78: Utilization of Research Findings 124

CHAPTER 1. INTRODUCTION

This chapter provides an overview of the research topic and a description of the safety network screening process. The research scope, research objectives and structure of the thesis are also presented.

1.1. Research Overview

Rail transportation is pivotal to North America's economy as it facilitates efficient movement of both commodities and passengers. Canada has over 46,000 kilometres of rail infrastructure supporting the transportation of goods and passengers (Transport Canada, 2019). A majority of these rail lines are classified as Class 1 such as the transcontinental freight railway systems owned by Canadian National and Canadian Pacific and the intercity passenger rail lines owned by VIA Rail (Transport Canada, 2011). The remaining regional and shorter railways are classified as Class II. The freight rail sector is responsible for transporting commodities over long distances while the passenger rail sector provides commuter, intercity and tourist transportation services. In 2017, the traffic levels, measured in train-miles for freight and passenger trains were 65 and 12 million respectively.

In contrast, the United States (hereafter "US") has approximately 140,000 miles (or 225,308 km) of rail infrastructure operated by over 650 rail companies supporting the delivery of goods and passenger transportation services (Association of American Railroads, 2019). As freight rail infrastructure is privately owned, railroad companies are responsible for their own maintenance and improvement. Rail tracks in the US are defined from Class 1 to Class 9. The Class is based on construction details and geometric variables and determines the maximum allowable speed limits. Seven[1] Class I railroads account for nearly 68% (477 million train-miles) of freight rail mileage connecting multiple states. Smaller, non-Class I rail tracks operate rail lines with shorter mileage and lower frequencies. Passenger rail services in the US are managed by the National Railroad Passenger Corporation, also known as Amtrak. Passenger rail services recorded a train mileage of 38 million in 2017 through the provision of medium- and long-distances intercity services (Bureau of Transportation Statistics, 2017).

1.1.1. Derailment Trends

Figure 1 shows the 10-year trend for main-track derailments for Canada and the US between 2009 and 2018. The Canadian data was retrieved from the Transportation Safety Board of Canada[2] (TSB) whilst the

[1] The seven Class I freight railroads are: BNSF Railway, CSX Transportation, Grand Trunk Corporation, Kansas City Southern Railway, Norfolk Southern Combined Railroad Subsidiaries, Soo Line Railroad, and Union Pacific Railroad (FRA, 2019)
[2] TSB (2019). Table 1. Railway occurrences and casualties, 2008-2018. Retrieved from https://www.tsb.gc.ca/eng/stats/rail/2018/sser-ssro-2018.html#3.0

US data was published by from the Federal Railroad Administration Office[3] (FRA). For comparison purposes, the numbers of derailments were normalized by millions-train-miles-traveled, a common measure for rail activity. The solid lines represent the number of main-track derailments per million train miles and the dashed lines represent the linear trend associated with the data points.

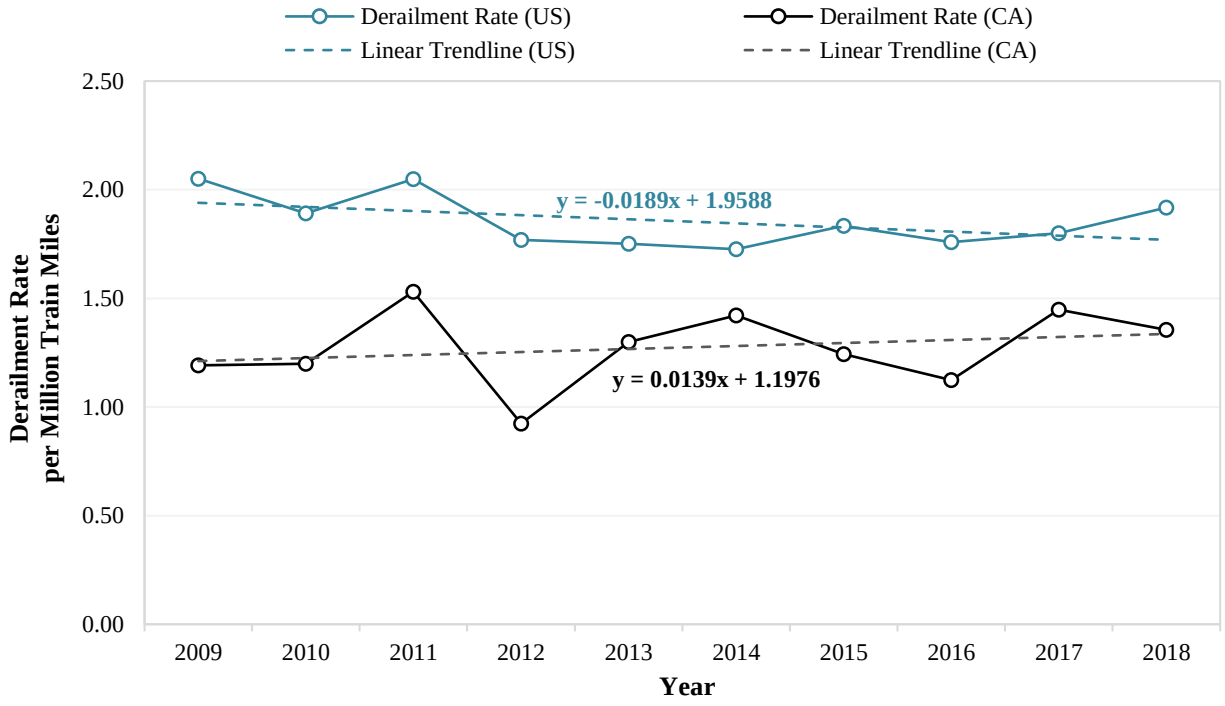


Figure 1: Derailment Rate per Million Train Miles in Canada and the United States

The US had a consistently higher derailment rate than did Canada, 1.85 compared to Canada's 1.3 over the ten-year period. The number of derailments in both countries fluctuated each year. The US shows an overall decreasing trend despite the increase of 7.8% from 2009 to 2010 followed by an 8.3% increase the year after. Canada shows an overall increasing trend with higher fluctuations than in the US. Derailment rates in Canada ranged from 1.19 to 1.53. For example, derailment rates compared to the previous year increased by 28% in 2011, declined by 40% in 2012, and decreased by 10% in 2016.

Although the derailment rates may appear to be very low for both countries, derailments are often associated with severe consequences. In recognizing this, governing agencies and rail companies put significant effort into ensuring the safety of rail transportation. The current rail safety improvement procedure involves a report-and-respond system where train-related accidents are self-reported by rail

[3] FHWA (2019). Ten Year Accident/Incident Overview. Retrieved from https://safetydata.fra.dot.gov/OfficeofSafety/publicsite/Query/TenYearAccidentIncidentOverview.aspx

operators. Transport agencies then respond through investigations, maintenance and/or implementation of mitigation measures.

In Canada, a rail operator files a report to the Transportation Safety Board of Canada (TSB) for review after an accident has occurred. The TSB then reviews the report, transcribes information to their central rail occurrence database and initiates appropriate actions according to the gravity of the situation. In certain situations, accidents may warrant an investigation in which the TSB is deployed to the occurrence site to conduct a detailed assessment (TSB, 2019). In such a case, the field investigators and/or engineers examine the site, identify safety deficiencies and provide recommendations which may include remedial actions and safety countermeasures. Thus, the current safety network screening approach is reactive rather than proactive since it lacks an ability to identify and predict areas of safety concern *prior* to accidents.

This study recognizes the limitations of the current system and applies a network screening method which involves the development of derailment prediction models to identify specific rail segments (hereafter "segments") of safety concern in Canada's rail network. A network screening tool may help the government, transport authorities and rail companies to identify, evaluate, and implement cost-effective risk mitigation strategies.

Train related accidents have major implications for the safety of the public, the environment and service interruptions. As such, commitment to safety is crucial to both Canada and the US's rail industries. Derailments and other train accidents often lead to severe consequences such as injuries and casualties when people are involved, and damage to the environment when dangerous goods are involved. Derailments can occur on main-tracks or non main-tracks. Main-tracks are the continuous main-line tracks used for through-trains while non-main tracks are the discontinuous sections of tracks that require trains to travel at reduced speed (Transport Canada, 2018a). Non-main tracks include spurs, rail yards, crossovers, etc. (see **Chapter 3.2.1** for more information on track classification). This thesis focuses on main-track derailments following the rationale discussed in **Chapter 2.1.2**.

1.1.2. Derailment Causes

Derailments are very rare events and the factors that contribute to this type of accident can be complex. Some derailments might involve a combination of factors and causes. In general, the cause of a derailment can be broadly categorized as track, equipment, human factors, signal, and miscellaneous. Each broad group can be divided into more specific causes. **Table 1** presents the top 10 common causes for main-track derailments in the US. Broken rails and welds are noted as the most frequent derailment causes and attributed to 15.3% of main-track derailments between 2001 and 2010 (Liu et al., 2012).

Table 1: Top 10 Common Causes for Main-track Derailments in US, 2001-2010 (Liu et al., 2012)

Rank	Common Causes	Percentage
1	Broken Rails or Welds	15.3%
2	Track Geometry (excl. Wide Gauge)	7.3%
3	Bearing Failure (Car)	5.9%
4	Broken Wheels (Car)	5.2%
5	Train Handling (excl. Brakes)	4.6%
6	Wide Gauge	3.9%
7	Obstructions	3.5%
8	Buckled Track	3.4%
9	Track-Train Interaction	3.4%
10	Other Axle/Journal Defects (Car)	3.3%

In Canada, the predominate causes for main-track derailments are rail, joint bar and rail anchoring (10.8%) closely followed by track geometry (9.7%). **Table 2** presents the top 10 common causes for main-track derailments in Canada (Leishman, 2017).

Table 2: Top 10 Common Causes for Main-track Derailments in Canada, 2001-2014 (Leishman, 2017)

Rank	Common Causes	Percentage
1	Rail, joint bar and rail anchoring	10.8%
2	Track Geometry (excl. Wide Gauge)	9.7%
3	Environmental conditions	6.9%
4	Wheels	6.8%
5	Train handling	6.6%
6	Miscellaneous	6.2%
7	Axles and journal bearings	5.3%
8	General switching rules	5.0%
9	Switches	4.8%
10	Brakes	3.9%

Research has found that the cause of a derailment may be affected by the seasonal. For example, wheel and rail failures are more prevalent in the winter months between December and March (English and Moynihan, 2007). Broken rail is also a seasonal variation as rails are more prone to thermal expansion and stresses due to overheating in the summer months (Liu et al, 2012; Leishman, 2017). Deterioration of track conditions can be related to train traffic since track degradation may be more severe along segments with high train volumes, heavy train loads, and/or high operating speeds. Causes related to track geometry and components can often be mitigated by inspection and maintenance.

1.1.3. Derailment Impacts

As previously mentioned, derailments can interrupt services, damage infrastructure, impact the environment, and lead to injuries and even casualties. The financial consequences of derailments depend on the severity (number of railcars involved) and location. Derailment severity is measured by the number of railcars involved. More severe derailments, i.e. those with more railcars involved, can lead to more damage and higher costs (e.g. locomotive repair, track maintenance, etc.) and greater economic loss through

loss in productivity. Derailments that occur in proximity to built-up areas can lead to more severe consequences (e.g. injuries and casualties) and higher recovery costs than derailments in a rural environment.

The economic costs incurred by derailments can be significant, particularly if dangerous goods are involved. The impact on society and the environment can be exacerbated by fires or explosions that may followed a derailment. In July 2013, a tragic train accident occurred in Lac-Mégantic, Quebec where a train carrying flammable petroleum crude oil was derailed due to operational failure of the braking system (TSB, 2013). This accident has led to 63 derailed cars and six million litres of oil leakage which turned into an explosion resulting in 47 deaths.

Table 3 provides a high-level cost breakdown for the Lac-Mégantic accident using data retrieved from an economic assessment report on accidents related to transporting crude by rail. The estimate for the environmental cleanup was $200 million in addition to other recovery costs of $500 million. These costs could be substantially higher if a derailment of similar scale occurs in a more populated area. In some cases, the environmental cleanup costs can be partially funded by the federal and/or provincial governments. Even with insurance coverage and financial assistance from governments, the financial liability associated with severe derailments may be too substantial for responsible parties to bear.

Table 3: Estimated Economic Costs for Lac-Mégantic Derailment (Goodman and Rowan, 2013)

Item	**Cost Estimates ($)**
Decontamination	$200 to $500 million
Town reconstruction, economic recovery and compensation for victims' families	$500 million
Total	$1 billion or more

Other indirect costs of derailments include rail service disruptions which can be difficult to quantify and not part of the scope of this thesis. At a high-level, such costs could be associated with loss in freight activities and the increase in travel times for passenger transportation services. For example, Northeast Corridor Commission's (NEC) annual report has estimated that one day of disrupted train service could translate to $100 million for the economy due to loss in productivity (NEC, 2017).

The safety impact and cost implications of derailments emphasize the needs for a more proactive method for mitigating derailment risks. Network screening is a systematic approach to identifying key areas of concern that may warrant consideration for safety improvements.

1.2. Safety Network Screening

Network screening is a common technique used for analyzing and comparing transportation facilities in accordance with safety performance. The Highway Safety Manual (HSM) (AASHTO, 2010) defines the safety network screening as a systematic process for analyzing a transportation network to identify and prioritize sites based on the potential for safety improvement.

This process is widely used in the transportation engineering industry to evaluate the safety performance of transportation facilities. As part of the process, prediction models are developed to predict the average number of derailments based on a set of characteristics (independent variables) associated with a segment. These prediction models are also referred to as Safety Performance Functions (SPFs) which are further explained in **Chapter 2.3.1**.

A network screening process typically consists of five major steps, as outlined below.

1. *Establish focus*: Describe the goal for network screening.
2. *Identify network and establish reference populations*: Identify the area of focus for screening, for example, derailments along main tracks.
3. *Select performance measures:* Select appropriate performance measures for reducing the frequency and severity of accidents and evaluating safety risks. For example, a performance measure might be the frequency of derailments.
4. *Select screening method:* Select a screening method. The screening method may, for example, involve the ranking of segments based on the expected frequency of derailments.
5. *Screen and evaluate results*: Carry out the network screening process and evaluate the results and ranking produced.

The networking screening results help transportation authorities to predict potential safety risks, allocate resources and implement appropriate mitigation measures more effectively and strategically.

1.3. Research Goal and Objectives

The primary goal of this research is to develop a network screening tool to identify segments with derailment risk (also referred as ‘hotspots’) in the rail network.

The study has the following key objectives:

1. Provide a literature review on state-of-the-art methodologies, tools and practices for rail risk management associated with derailment;
2. Conduct a descriptive analysis to gain an understanding of historical trends and patterns of derailments in Canada;
3. Develop segment-level derailment prediction models using negative binomial and logistic regression modelling methods;
4. Carry out network screening and identify derailment-prone segments; and
5. Compare and discuss the model performance of the negative binomial and logistic regression modelling techniques.

1.4. Scope and Structure of Thesis

The scope of this thesis consists of conducting safety network screening on Canada's rail network with a focus on main-track derailments involving freight and passenger trains. Main-track derailments are any occurrence where the wheels of the trains come off the main-tracks, excluding collisions. A crucial component of the network screening process includes the development of segment-level derailment prediction models by employing negative binomial and logistic regression modelling techniques.

At the outset of the study, it is important to gain an in-depth understanding of the research topic and data. This primarily involves a literature review as documented in **Chapter 2**. It also involves data collection, data validation and a trend analysis as documented in **Chapter 3**. The subsequent tasks pertain to the development of prediction models and the identification of segments with derailment risk. These issues are discussed in **Chapters 4** through **7**. **Chapter 8** then provides a comprehensive review of the results of the analysis and the model performances. **Chapter 9** discusses the study's research contributions and limitations and makes suggestions for future research.

1.5. Chapter Summary

Rail transportation is a crucial component of the overall transportation system for commodities and passengers in both the US and Canada. It is important to ensure the safety of the rail infrastructure to minimize safety risks and loss in productivity. Train accidents, derailments in particular, have strong implications for the safety of the public and may be associated with very substantial costs in terms of

environmental impact and disruption to services. Canada's current rail safety management system is reactive rather than proactive. Safety measures are often considered and implemented only after an accident has occurred. Thus, the current approach lacks a predictive ability for identifying future hotspot locations *prior* to accidents.

To overcome this limitation, this study applies a safety network screening process to evaluate the safety performance of Canada's rail network. Such a process involves the development of segment-level derailment prediction models and identification of segments with derailment risk. An ability to predict derailment risks would allow governing agencies and rail companies to systematically evaluate the safety performance of the network, perform maintenance and take remedial actions in a proactive manner.

The research methodology and findings in this thesis have benefits for transportation agencies, rail companies and safety practitioners. The study involves the development of segment-level prediction models for estimating derailment risks which help to identify the key segments that warrant special attention for further investigation and/or implementation of preventive measures.

CHAPTER 2. LITERATURE REVIEW

This chapter presents a literature review of state-of-the-art methodologies and emerging practices for rail safety management, particularly for prediction of derailment risks. It provides insights into the use of scientific methods to evaluate the safety performance of rail infrastructure. Note that a fair amount of research has been done on derailment prediction in the United States, but the relevant topics have not been extensively explored in Canada.

2.1. Railway Safety

The following sections provide an overview of historical trends and patterns of derailment occurrences in the US and Canada.

2.1.1. Derailments in the United States

Figure 2 shows the number of derailments and millions-train-miles-traveled in the US using data retrieved from FRA (2019) for 2009 to 2018. The number of train miles is a common measure of rail activity. It is calculated by multiplying the number of trains operated in a given year by the distance travelled. The number of derailments remained fairly consistent for the 10-year period. The highest number of derailments was observed in 2011**.** The number of train miles was also fairly consistent.

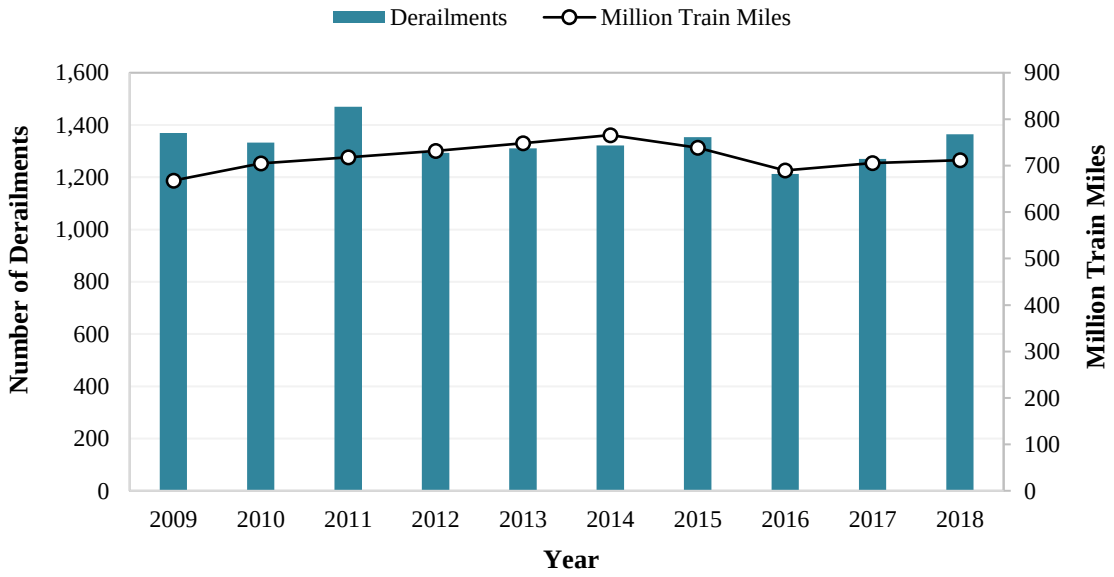


Figure 2: Number of Derailments and Total Train Miles in the US (2009-2018)

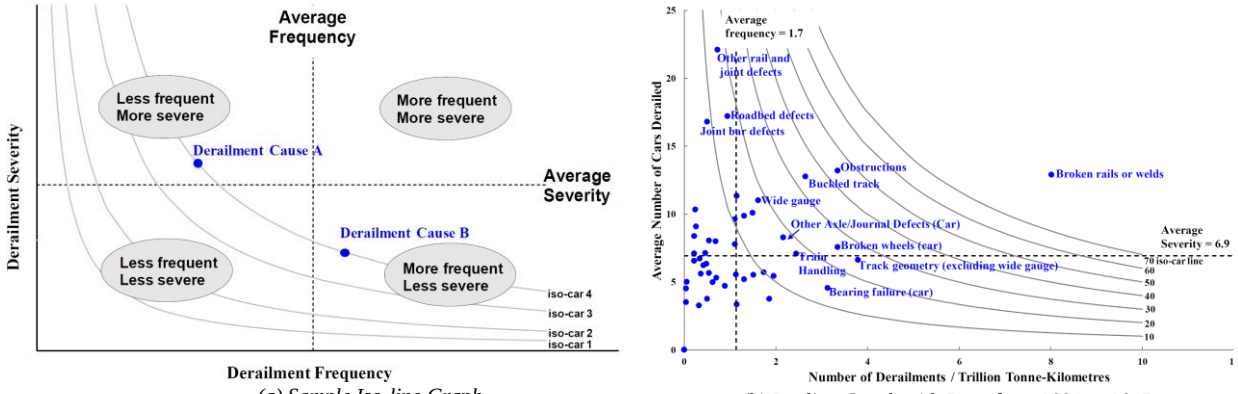
In the US, extensive research has focused on derailments risk analysis. The analyses and findings have been used to address a variety of rail risk analysis and management questions. Many of the studies

analyzed derailment potential based on system characteristics such as track classes. Rail tracks in the US are classified from 1 to 9 according to the geometrical requirements and speed limits. The allowable speed increases as the track class number increases; the speed limits for Class 1 and Class 9 are 10 mph and 110 mph respectively (FRA, 2019). Nayak et al. (1983) noted a strong statistical correlation between track class and the number of derailments with higher FRA track classes statistically associated with fewer derailments. Other studies found the same relationship (Treichel and Barkan, 1993; Liu et al., 2012). The finding is not surprising as higher FRA track classes imply higher train speeds and higher train which in turn imply stringent engineering and maintenance standards to ensure safe operations.

Derailment causes have also been analyzed to quantify derailment risks and determine areas for further safety improvements. Other than using tabular data, another tool that was introduced by researchers to assess the risk level of derailments is called an iso-car line graph. **Figure 3** presents two sample iso-car line graphs. This graphical technique can be used to compare and assess the risk levels of different derailment causes based on severity (y-axis) and frequency (x-axis). The measure of severity for derailments is the number of railcars involved. The iso-car line represents the equal level of risk based on the number of derailed cars, an inverse function of the average number of derailed cars, and the number of derailments (Wang et al, 2017). The iso-lines and iso-car numbers represent risk levels based on the distance from the graph's origin.

In **Figure 3(a)**, for example, the bottom left quadrant represents derailments that occur less frequently and that are less severe (lower number of derailed cars). The top right quadrant represents derailments that occur more frequently and that are more severe (higher number of derailed cars). This is the quadrant with the greatest risks.

Figure 3(b) is based on the same principles and shows real data from 2006 to 2015. The Figure shows that broken rails or welds were the most frequent and severe causes of derailment. This pattern is consistent with another study's findings (Liu, 2015). Broken rails or welds is also the cause associated with the highest iso-car level of 85. Other studies have shown that track defects and mechanical failures are the most prevalent casual factors associated with derailments (Liu et al., 2012; Liu, 2015; Leishman, 2017). Nonetheless, the iso-line technique can be useful for gaining an understanding on historical derailments and potential causes.



(a) Sample Iso-line Graph *(b) Iso-line Graph with Data from 2006 to 2015*

Figure 3: Iso-line Graphs for Derailments (Wang et al., 2017, p. 822)

2.1.2. Derailments in Canada

In Canada, the TSB maintains accident records for all federally-regulated railways. The records are published on TSB's online data portal. Railway companies also manage and maintain their own accident databases which may or may not contain more information than the TSB records (access to this data was not available for this study). **Figure 4** presents the annual statistics for derailments in Canada by millions-train-miles-traveled[4] between 2009 and 2018. The Figure shows a positive relationship between the number of derailments and train activity levels except for 2012.

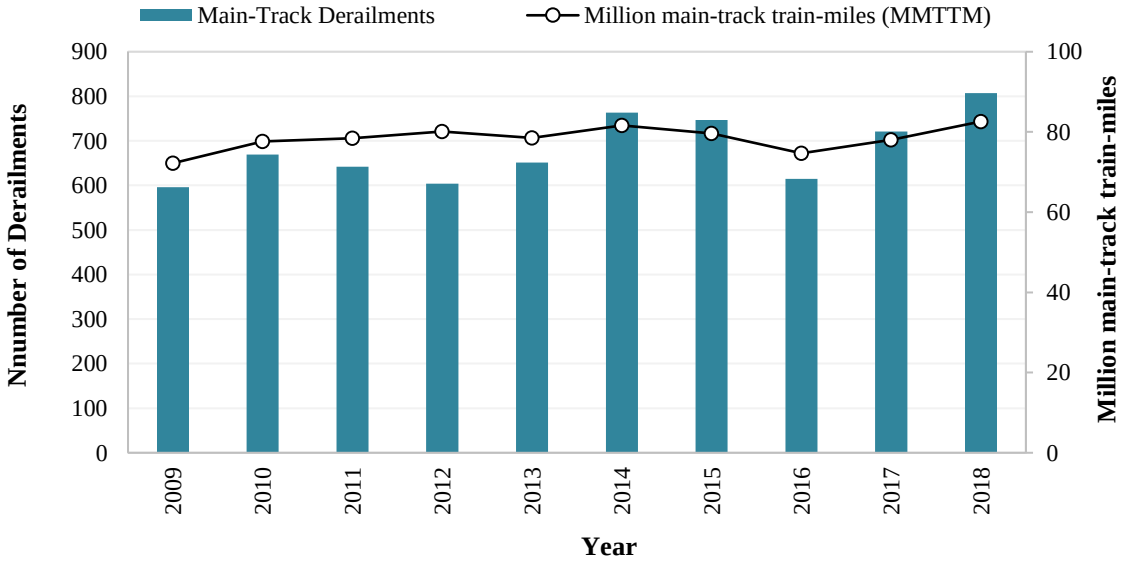


Figure 4: Number of Main-Track Derailments and Million Train Miles in Canada (2009-2018)

[4] Retrieved from Rail transportation occurrences in 2018, Table 1. Railway occurrences and casualties, 2008-2018 (TSB, 2018)

A comparison of main-track and non main-track derailments in Canada for the same time period is presented in **Figure 5**. The number of main-track and non main-track derailments fluctuated from year to year with a slight upward trend over the 10-year period. There are consistently over 500 derailments of non-main-track derailments recorded over the analysis period. In general, there is a steady trend over the 10-year period with an average of 100 main-track derailments annually. Freight trains accounted for 99.3% and 99.8% of derailments on main-tracks and non main-tracks respectively.

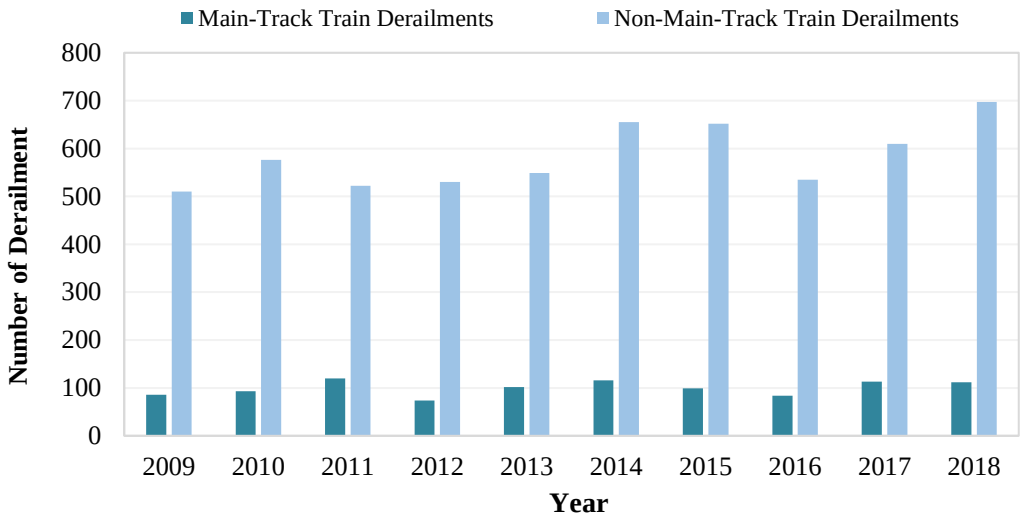


Figure 5: Main-track and Non-main-track Derailments in Canada (2009-2018)

The TSB publishes rail safety data for reportable accidents. These include derailments, collisions[5], crossing accidents and other (e.g., accidents involving track units, employees/passengers, fires, and explosions). **Figure 6** shows 10 years (2009-2018) of reportable accidents of which 55% were derailments. Non-main-track derailment was the most common type of accident and accounted for 47% of all accidents.

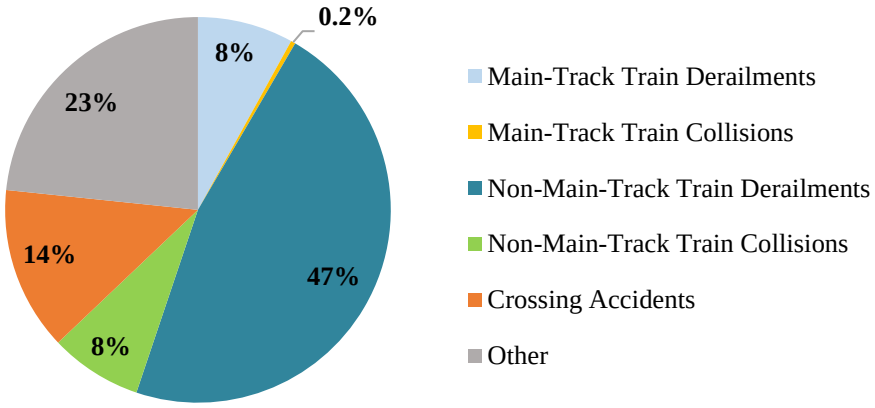
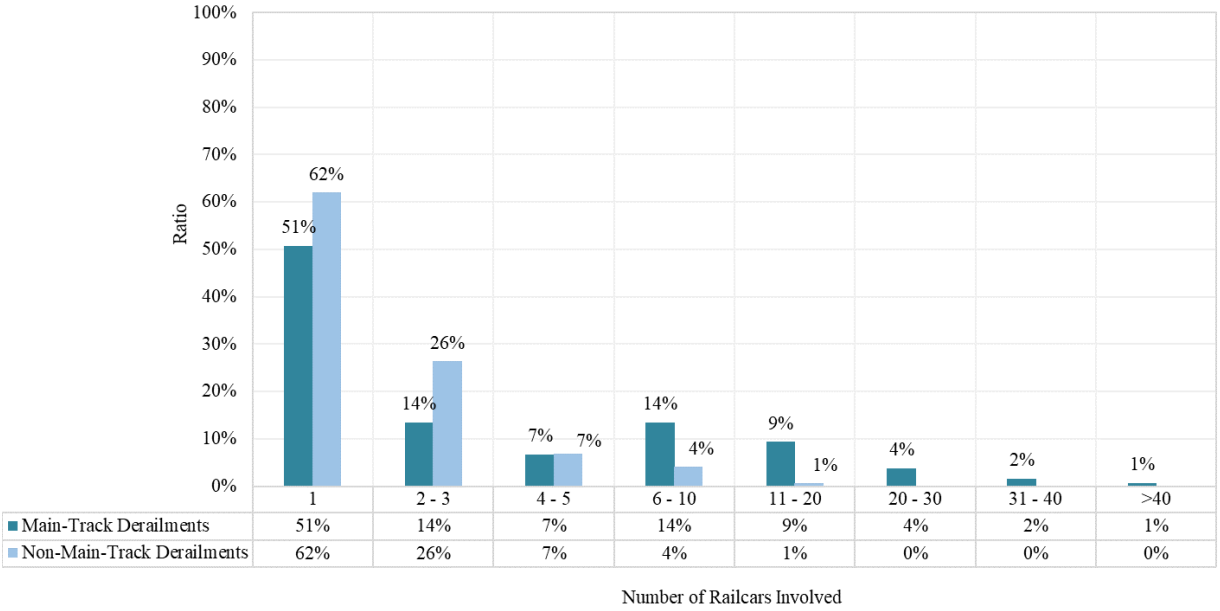


Figure 6: Train Accidents by Type (2009-2018)

[5] "Collision" means an impact, other than an impact associated with normal operating circumstances, between (a) rolling stock; (b) rolling stock and a person or vehicle; or (c) rolling stock and an object or animal, if the rolling stock is damaged or derailed. (Transportation Safety Board Regulations SOR/2018-258, s. 4 (6), 2014)

In terms of severity, main-track derailments typically involved a higher number of railcars compares to non-main-track derailments. **Figure 7** compares the number of cars involved in main-track and non-main-track derailments in Canada (TSB, 2019). Derailments along main-tracks involved more railcars than those occurring on non-main-tracks. Severe main-track derailments could involve more than 40 railcars. The number of railcars involved in non-main-track derailments was generally less than 20.



	1	2 - 3	4 - 5	6 - 10	11 - 20	20 - 30	31 - 40	>40
Main-Track Derailments	51%	14%	7%	14%	9%	4%	2%	1%
Non-Main-Track Derailments	62%	26%	7%	4%	1%	0%	0%	0%

Figure 7: Distribution of Railcars Involved in Derailments (2009-2018)

As previously mentioned, derailments can have a very significant safety and environmental impact in addition to leading to service disruptions. Despite the higher proportion of non-main track derailments (e.g., a derailment in a rail yard), this type of accident often involves less than two railcars, low-speeds and relatively minor consequences whereas main-track derailments are often associated with much more serious consequences due to higher train speeds and train traffic. Transport Canada (2018) noted that main-track collisions and derailments are the most serious train accidents in terms of potential risk to the public and economic loss.

The number of fatalities or injuries is relatively low in train related accidents. None of the derailments involving passenger trains resulted in injuries between 2009 and 2018. However, the potential release of dangerous goods or hazardous materials is a major concern. In 2018, 135 train accidents involved dangerous good and 65% of the 135 were due to derailments (TSB, 2019). The release of dangerous goods or hazardous materials can result in toxic chemical exposure that is harmful to train crews and people in close proximity to the accident. When explosions occur, there may be fatalities, as occurred in Lac-

Mégantic. Due to the scale and nature of train accidents, it is important to consider the application of prediction models to estimating the number of derailments.

2.2. Derailment Prediction Models and Factors

Two types of prediction model are relevant to estimating and predicting derailments: exposure-only models, and multivariate models. Exposure models consist of one input variable whereas multivariate model include more than one independent variable. Other factors crucial to derailment prediction include regression-to-the-mean bias, the track segmentation method, and the length of the periods of analysis.

2.2.1. Exposure-Based Models

Accident prediction has strong implications for rail safety policies, operating practices, risk management and resource allocation. It was also recognized that mitigation strategies can be improved by having the ability to identify safety deficiencies before a concern or issue arises (English and Moynihan, 2007).

In general, safety estimation can be expressed as the number of accidents (frequency) or as accident rate. Accident rate is the number of accidents normalized by a metric of traffic exposure such as train-miles, railcar-miles or gross ton-miles. These metrics represent total distance traveled multiplied by number of trains, number of railcars, and weight (e.g., tonnage) respectively.

Schafer and Barkan (2008) investigated train accidents (including derailments, collisions and highway-rail crossing accidents) by considering train-miles and railcar-miles as measures of exposure. The study used US train accident data from 1990 to 2005. Using exposure as the only input, Schafer and Barkan developed the following model **(Equation 1)** to predict the number of accidents for Class 1 freight rail lines. **Equation 1** shows a positive relationship between the likelihood of train accident and traffic exposure.

$$A_{EXP} = 1.05 \times 10^{-8} M_C + 8.62 \times 10^{-7} M_T \quad \text{(Eq.1)}$$

where:

A_{EXP} = expected number of accidents,

M_C = number of railcar miles, and

M_T = number of train miles

In this model, number of accidents is estimated by taking the sum of the railcar mile accident rate multiplied by the number of railcar miles and the sum of train mile accident rate multiplied by the number of train miles. The model suggests that a single train will experience a higher number of accidents with an increase in the number of railcar miles. See **Figure 8a**.

Conversely, **Figure 8b** shows that longer trains would result in fewer accidents assuming that the number of railcars remains constant. This finding is based on the assumption that more trains are required to transport the same number of railcars and this increases the possibility of accidents due to the higher level of traffic exposure (i.e., an increase in train-miles).

However, exposure-based models are of limited value as they lack explanatory factors which may influence the likelihood of accidents. For example, **Equation 1** does not account for the increased risk that are associated with longer trains (e.g., train handling issues and braking).

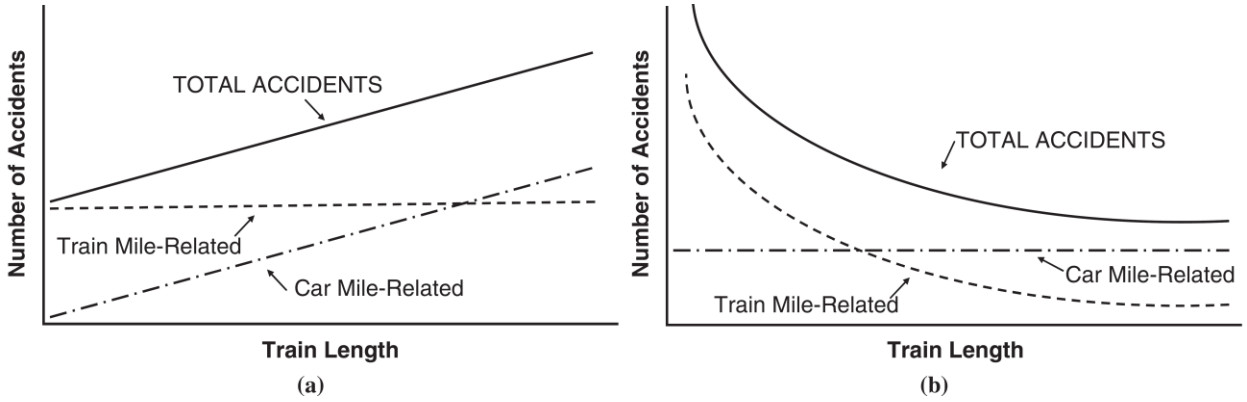


Figure 8: Expected Accidents Relative to Train Length
(a) for a single train and (b) for a fixed amount of traffic. (Scharfer and Barkan, 2008, p. 74)

In Canada, English and Moynihan (2007) indicated that derailment rate can be used as a measure for rail network screening. They suggested that, tonnage-miles might be more suitable than distance traveled for calculating derailment rate as tonnage-miles captures the train activity level (e.g., how much cargo being is by the train) (English and Moynihan, 2007). If tonnage-miles is used as an exposure measure, an increase in axle loads would reduce the derailment rate given the increase in productivity level. English and Moynihan also noted that it is appropriate to evaluate rail operators that carry similar commodities. Since heavy train loads are likely to cause more track damage and increase derailment potential, tonnage-miles appeared to be an appropriate exposure variable.

The approach could, however, be biased when different railways carry a wide range of products or when the product types change over time. Since train load data was not available for our study, we used train traffic as the exposure variable for derailment prediction.

Exposure-based models may provide an estimation of average risk, but most risk management decisions require greater precision. Unfortunately, lack of data often constrains analysis. If more detailed data is available (e.g., finer differentiation of the data categories especially for traffic density) a more robust statistical estimation can be carried out to provide greater resolution in the results (Liu, et al., 2016).

2.2.2. Multivariate-Based Models

As more complex questions have arisen, researchers have developed alternative approaches in prediction modelling, for example, approaches that consider multiple independent variables of interest (Liu et al., 2016; Fienberg, 1980; Agresti, 2007). In our context, this type of approach considers the total number of derailments and combinations of different independent variables. The following variables have been commonly used in derailment prediction modelling: FRA track class, operation (signal, non-signaled), and traffic density.

In derailment prediction models, a common statistical method used to develop multivariate models is negative binomial (NB) distributions.

The NB distribution estimates the number of derailments using historical data such as traffic exposure, method of train operation and track classification (Liu et al., 2012, 2015, 2016). **Equations 2** through **4** show the model form of negative binomial regression to estimate the number of derailments on US Class 1 railroads.

$$Y = \text{Poisson}\ (\lambda) \tag{Eq.2}$$

$$\lambda = \text{Gamma}\ \left(f, \frac{f}{m}\right) \tag{Eq.3}$$

$$m = exp\left(\sum_{p=0}^{k} b_p\ X_p\right) M \tag{Eq.4}$$

where:

Y= observed number of derailments,

m = estimated number of derailments,

b_p= p^{th} parameter coefficient,

X_p= p^{th} independent variable,

M = traffic exposure (gross ton-miles), and

f = inverse dispersion parameter.

The estimated derailment rates by FRA track classes are given in **Equation 5**, as shown below.

$$z = exp(b_0 + b_{trk}X_{trk} + b_{moo}X_{moo} + b_{den}X_{den}) \tag{Eq.5}$$

where:

Z = estimated derailment rate per gross ton-miles,

X_{trk} = FRA track class (1 to 5),

X_{moo} = method of operation (1 for signaled, 0 for non-signaled),

X_{den} = annual traffic density level (1 for $\geq$ 20 MGT, 0 for $<$ 20MGT), and

b = parameter coefficients.

Agresti (2007) used a maximum likelihood method in a NB model to estimate parameter coefficients where all three variables were statistically significantly related to the number of derailments. Individual railcar derailment was also calculated. It represented the likelihood that an individual railcar is involved in a derailment and was measured by calculating the number of railcars derailed per unit of traffic exposure. Agresti noted a correlation between higher track classes and fewer derailments. This is consistent with previous studies.

In a different study, another influential factor for derailments was the type of rail track operation method (signal or non-signaled track). Using likelihood ratio tests, non-signaled tracks were found to have a greater risk that than signaled track of derailments caused by broken rails (Liu et al, 2013).

2.2.3. Regression-to-the-Mean Bias

Liu and Rodriguez (2016) have shown that derailment estimates may be vulnerable to the statistical error known as regression-to-the-mean (RTM) bias. RTM is a statistical phenomenon. In the case of safety research, where accidents occur as random events that fluctuate over time around a long-term average, selecting an individual year is unlikely to provide an appropriate representative picture of actual safety and does not provide enough information to draw conclusions about changes in the number of accidents. When evaluating derailment risks, RTM bias may occur when a study fails to consider the long term mean value for the number of derailments.

Figure 9 illustrates the effects of RTM**.** Accident rates in this figure represent the number of accidents normalized by train traffic (million car-miles). The solid and dotted lines represent the average accident rates for railway lines A and B respectively, and the dots show the observed accident rates.

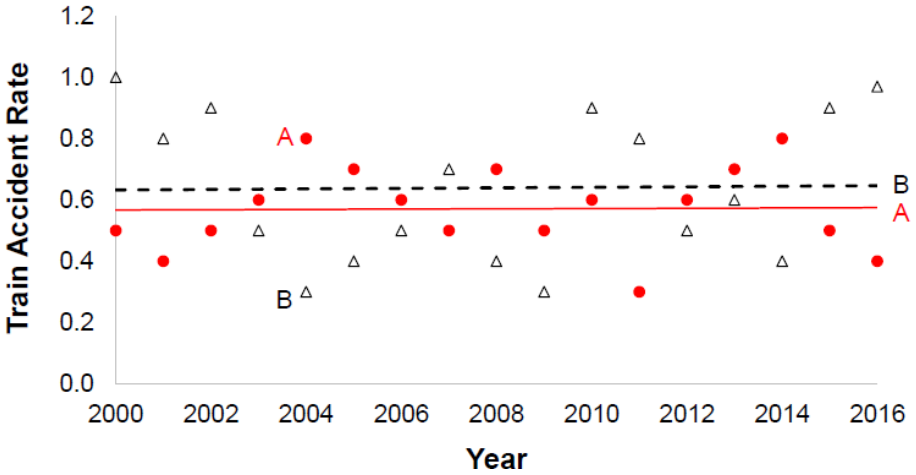


Figure 9: Sample of RTM effect (Liu and Rodriguez, 2016, p. 4)

Some empirical accident rates for Line A appear to be higher than those of Line B for certain years (e.g., 2006 and 2012) although the average accident rate of Line A should be lower than Line B. Such deviation is due to the RTM issue discussed above. The Figure shows that an erroneous conclusion might arise if one identifies a location as 'hotspot' based on one year of data as that year's data could be an extreme value in a given period of time. It is therefore important to consider a particular site for a longer period and to be aware that the accident rate will naturally regress towards the long-term mean value over a longer period (FHWA, 2019).

Considering the effects of RTM in rail safety, some researchers have predicted and compared derailment rates against another railroad or the system average to determine whether safety variation exists (Liu and Rodriguez, 2016). The researchers developed exposure-based models using negative binomial distribution techniques and noted that this was a better approach than merely using the observed data to identify locations of concerns. While Liu and Rodriguez's approach provides an indication of safety deviation from the system average, the approach does not effectively reduce RTM bias as it does not consider the derailment history.

The RTM issue has been intensively studied in road safety studies in which safety performance has been calibrated for different types of roadway in different regions and municipalities (Brimley, Saito and Schultz, 2012; Lu. et al., 2014, Tayki and Park, 2017; Farid, Abdel-Aty and Lee, 2018). However, RTM has not received the same level of attention in rail safety.

2.2.4. Importance of Segmentation

In road safety analysis, SPFs are typically estimated for road segments and for intersections. The quality of the SPFs depends on how the entities analyzed are grouped together to give a reasonable level of homogeneity for each location of interest (Cafiso et al, 2018). Segments, for example, are often defined by the road's function (highway, arterial, collector, or local), characteristics (e.g. number of lanes and lane widths) and exposure (e.g. traffic volumes). The Highway Safety Manual (HSM) provides families of SPFs that can be applied to segments with similar characteristics to improve the statistical inference of the functions.

Although segmentation and its effects on accident predictions have been studied, there is no definitive agreement on the most appropriate segment lengths for safety analysis. Some studies have indicated that both very short and very long segments would create bias in the identification of hotspots (Miaou and Lum, 1993; Ogle et al.,2011; Cafiso and Di Silvestro, 2011). Shorter segments are prone to inaccurate identification of accident locations (Quin and Wellner, 2012) and an overabundance of segments

with zero observations. Both aspects of short segments could yield unreliable results in prediction models. Longer segments could mitigate these short segment issues. Cafiso et al (2018) found that segmentation often depended on data availability and quality. Although no standard minimum segment length has emerged for predictive modeling, a segment length of at least 0.10 miles (0.16 kilometre) has been recommended (Cafiso et al, 2018). These research findings are useful in relation to the segmentation analysis discussed in **Chapter 3.2.4**.

There are currently no studies that have considered nor discussed the best practice for rail track segmentation. For roadway SPFs, segments in roadway network are generally defined by the midblock link between two adjacent intersections. However, for a railway network, the distance between "intersections" or "nodes of interest" such as stations, junctions or at-grade crossings could vary from 10 metres up to over 100 kilometres.

Segmentation determines the sample size of the database for prediction modelling and affects the assignment of attributes (e.g. train volumes and speeds) in the rail network. A database with too many short segments could lead to an excessive number of segments with zero derailments. Longer segments might resolve this issue but would sacrifice homogeneity. Due to the large variation in segment lengths, careful consideration of the methodological approach to rail track segmentation is required for the development of a derailment prediction model.

2.2.5. Length of Analysis Periods

The analysis period plays a major role in statistical modelling. Given the rare nature of train accidents, the analysis periods in previous studies have varied and some covered many years. As an example, **Table 4** shows that a number of studies that involved rail safety assessment have used analysis periods ranging from 8 to 36 years for trend analyses and prediction modelling.

Table 4: Summary of Analysis Periods of Rail Safety Research

Authors	Study Date	Research Topic	Analysis Period	No. of Years
Evans	2007	Rail Safety and Rail Privatisation in Britain	1967 – 2003	36
Moynihan and English	2007	Causes of Accidents and Mitigation Strategies	1999 – 2006	8
Schafer and Barkan	2008	Train Length and Accident Causes and Rates	1990 – 2005	16
Evans	2011	Fatal Train Accidents on Europe's railways	1980 – 2009	29
Saat and Barkan	2012	Causes of Major Train Derailment	2001 – 2010	10
Liu	2015	Statistical Analysis of derailment in the US	2000 – 2012	13
Liu	2016	Collision Risk for Freight Trains in the US	2000 – 2014	16
Li and Barkan et al	2018	Analysis of the Derailment Characteristics	2001 – 2015	15
Khan and Lee	2018	Highway Rail Grade Crossing Prediction Model	2000 – 2016	17

The underlying assumption for using extended analysis periods is that the characteristics of the rail network remain the same. The inclusion of many years of data may present the benefit of providing more data for analysis to compensate for the low frequencies of train accidents. However, the variation in independent variables over a long analysis period (>20 years) can be prone to inconsistent and misleading prediction results.

The rail industry experienced an economic downturn during the recession in 2008 leading to cutbacks in operations and productivity. This resulted in a rapid decline in rail traffic and operating revenues during the recession period, as shown in **Figure 10**. The statistics for both rail productivity and traffic levels are more consistent after the recession. As documented in **Chapter 3.3.1,** the distribution of derailments shows similar patterns over a 10-year period beginning in 2009. Considering these factors, this study uses 10 years of data during the post-economic recession timeframe.

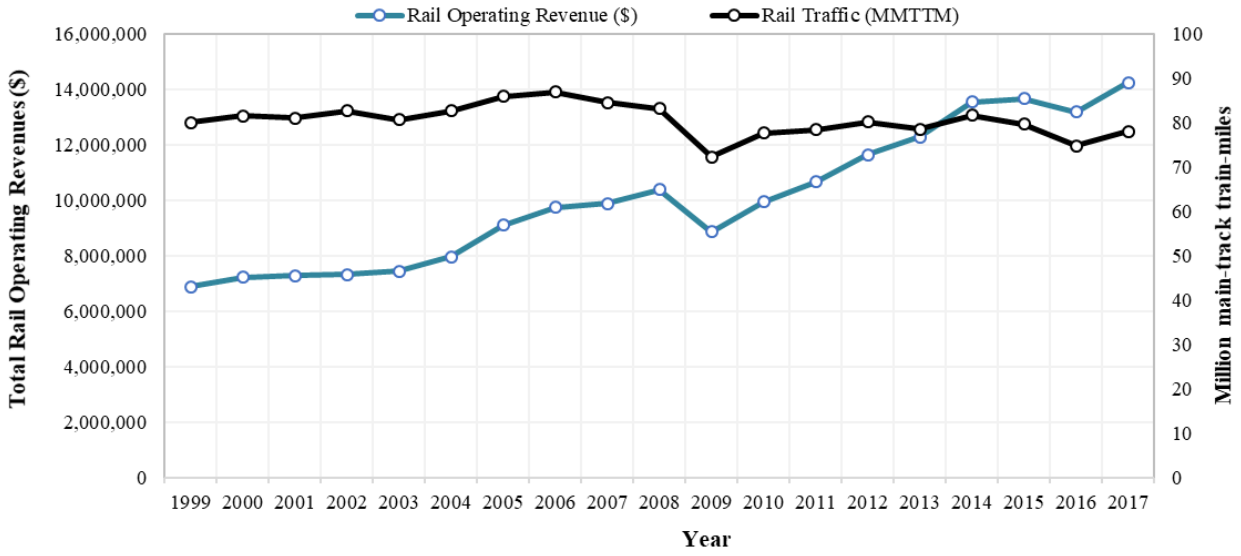


Figure 10: Rail Operating Revenues and Million Main-Track Train-Miles, 1999-2017
(Statistics Canada, 2019; TSB, 2012; TSB, 2018)

2.2.6. Limitations of Existing Rail Models

Many scientific studies have examined the statistical relationship between different rail characteristics (e.g., rail types and classes) and derailment rate. Studies include analyses undertaken using aggregated derailment data (Liu et al., 2012, 2015, 2016; Wang et al., 2017). In all of these cases, none of these studies estimated the number of derailments per segment but rather with aggregated data such as data based on rail track classification.

Given the different structure and organization of US train safety and traffic data and the problem of inconsistent spatial information, segment-level derailment rate analyses are not feasible (Liu et al., 2017).

Since US railroads are privatized, segment-level information would require proprietary information which makes it more challenging for researchers to conduct more detailed statistical analyses. For this reason, previous US studies on derailments are constrained with system-wide data and have not considered, for example, individual regions or segments within a rail system.

As a rail track may traverse different areas inside or outside of Canada and these areas may range from a few kilometres to hundreds of kilometres in length, safety could vary for different segments of the rail track (not just for different classes of rail track). The limitation of using aggregated data is that the results may distort the actual safety performance of a rail network by failing to screen segment-level safety issues. It is therefore crucial to select a network screening method that can evaluate the level of safety of different segments in the rail network.

The extended analysis periods (e.g. > 20 years) adopted in a number of studies can be viewed as a limitation in predictive modelling. The variation in the input data for such long periods requires strong data assumptions which could make models become unstable and produce unstable results.

Taking into consideration the economic recession of 2008 and the distribution of derailments, this study uses the most recent 10 years of complete data for prediction modelling.

2.3. Modelling Methods

This study applies two statistical methods for the purpose of predicting derailments: negative binomial and logistic regression models. The former method predicts frequency of derailment while the latter estimates the probability of derailment as a binary outcome (1 = presence of derailment risk, 0 = absence of derailment risk). Discussion on the applications of these methods is provided in the next sections.

2.3.1. Highway Safety Manual Approach

The Highway Safety Manual (HSM) is an industry guideline widely used by transportation engineers and planners for incorporating quantitative safety assessment in project planning and development processes (HSM, 2010). It defines scientific methodologies for evaluating the safety performance of highways and streets to inform the decision-making process. This guidance document presents best practices that are applicable for different life cycles of a project, from fundamental principles to countermeasure selection and evaluation of design alternatives.

The guiding principles relevant to this study are the application of the network screening process and safety performance functions (SPFs). SPFs are models that can be used to predict accidents for a

specific facility type. They are commonly used in roadway safety studies to reduce the effect of RTM when using negative binomial models (Srinivasan and Bauer, 2013). The development of SPFs consists of collecting and analyzing accident, exposure, geometric and other data for a set of multiple sites for a similar set of circumstances (Farid, 2018; Stapleton et al, 2018). The SPFs estimated functions can be used to predict the number of accidents for a subject site that is similar to the set used to develop the SPF.

The development and application of SPFs are crucial steps in the safety network screening process. In the area of rail safety, an SPF is a set of regression models that can be used to predict the number of derailments as a function of exposure (train volumes) and other railway attributes (e.g. segment length, segment speed, etc.). The HSM approach recommends the Empirical Bayes (EB) method to account for regression-to-the-mean (RTM) bias. The EB method synthesizes the observed and predicted numbers of derailments to estimate the expected safety performance for a subject segment. Note that EB method is only applicable for existing rail networks with historical derailment data. It cannot be used for predicting derailment frequency for a planned rail network. The EB model form is as follows (**Equations 6** and **7)**

$$E[y_i] = w_i \cdot \mu_i + (1 - w_i) y_i \quad \text{(Eq.6)}$$

where:

$E[y]$ represents the EB adjusted number of derailments for segment *i*,

μ_i represents the predicted number of derailments for segment *i*,

y_i represents the observed number of derailments for segment *i*, and

w_i is the EB weight factor for segment *i*, which is estimated in **Equation 7**.

$$w_i = \frac{1}{1 + \alpha \times \sum_{t=1}^{Y} \mu_{ti}} \quad \text{(Eq.7)}$$

where:

α is the dispersion parameter,

μ_{ti} is the predicted number of derailments for train segment or track unit *i* in year *t*, and

Y= 10 years in this study.

Figure 11 provides an illustration of the EB method. Note that the observed number of derailments could be lower or higher than the predicted value.

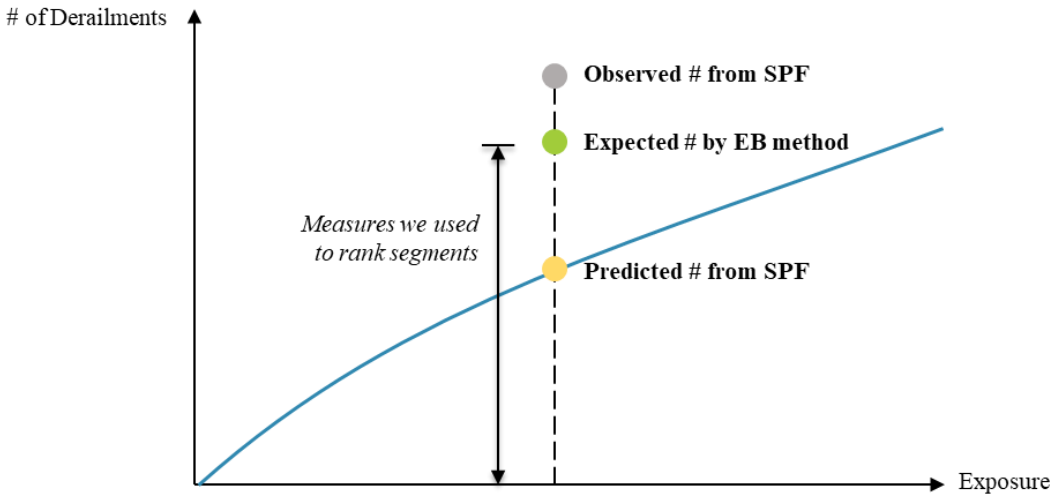


Figure 11: Empirical Bayes Method

SPFs can be used in the network screening process to determine whether the observed safety performance of a segment is higher or lower than average. Consequently, the expected number of derailments can be used as a ranking measure for comparing the potential for safety improvements of different segments.

2.3.2. Logistic Regression Model

A binomial logistic regression predicts the probability that an observation is classified as one of two categories of a dichotomous dependent variable, as a function of one or more independent variables. Logistic regression modelling is often regarded as a potentially acceptable method for handling data with excess zeros (Hu, Li and Lee, 2010; Khan and Khan, 2018; Chen and Fan, 2019; Choi, Chin and Lee, 2019).

In this study, it is expected that many segments in the rail network contain zero observations within the analysis period since derailments are very rare events. Considering its sensitivity in handling excess zeros in a dataset, a binomial logit model can be applied to identify segments with derailment risk. The result is a binary outcome (either zero or positive). Furthermore, since derailments are associated with severe consequences, a segment with any predicted risk of derailment (regardless of the expected frequency of the derailments) should warrant consideration for investigation and/or safety improvements.

Little research that has been done in applying logit modelling techniques in rail safety assessment. A study undertaken in 2018 involved the use of a binomial logistic model to predict accident likelihood at highway-rail grade crossings in North Dakota (Khan, 2018). The rail crossing accident data showed patterns similar to Canada's derailment statistics, for example, 93% of the locations exhibited zero crash frequency

between 2000 and 2016. Standard logistic regression modelling method was employed using the formula given in **Equation 8**.

$$Y = log\left[\frac{P_1}{P_0}\right] = \alpha + \beta_i \chi_i \quad \text{(Eq.8)}$$

where:

Y = binary dependent variable (1 = Presence of Derailment, 0=Absence of Derailment)),

P_1 = probability of accident,

P_0 = probability of no accident,

α = intercept,

β_i = estimated vector of parameters and

χ_i = vector independent variables.

The estimated parameters can be interpreted as odd ratios to understand their statistical association with the dependent variable.

A range of literature has focused on predicting accident likelihood and determining its contributory factors in the area of road safety using logit models (Crocco et al., 2010; Hu et al., 2010; Simoncic, 2000; Al-Ghamdi, 2002; Chen and Fan, 2019). Logit modeling method is particularly common for analyzing accidents involving pedestrians and cyclists as these are rare events (Simoncic, 2000; Chin and Quddus, 2003). Some researchers have adopted logit modeling techniques for analyzing categorical dependent variables such as accident severity (e.g. fatal or non-fatal accidents). For example, Al-Ghamdi (2002) applied a logit regression approach to determining statistically significant variables contributing to severe accidents. The odds ratio concept was used in this study to interpret the model results.

No studies using logit modelling method to examine derailments could be found in the literature search. In a review of accident prediction models, Niveditha et al (2015) concluded that logit models produced better accident prediction results than did negative binomial and Poisson distribution methods. The comparison was conducted based on only two parameters (R^2 and Chi-square tests) which might not provide sufficient evidence on the predictive capabilities and effectiveness of the different modelling methods. Additional goodness-of-fit tests could be adapted to further understand the different model behaviours and performances.

Hosmer and Lemeshow (HL) tests, for example, are often used to evaluate the performance of logit models. The HL test involves grouping entities in a dataset (the default is 10 groups of approximately equal size) and applying the Pearson Chi-square test to compare the predicted and observed values (Hosmer D.W. and Lemeshow S., 1980). Although researchers have stated that the results greatly depend on the number of groups, no real guidance is available for the choice of that number (Allison, 2014). The selection of the

number of groups is completely arbitrary. Another limitation of HL tests, though not applicable to this study, is the tests' poor power in small datasets. For these reasons, other goodness-of-fit tests were applied in this thesis as appropriate to assess the fitting performance of the logit models. See **Chapter 7.3**.

Given the emerging popularity of logit models in accident data analysis, this thesis employed this modelling method for predicting derailments. This study also evaluates whether or not this type of model is more sensitive to the over-representation of segments with zero derailments compared to the NB method. A quantitative comparative analysis was conducted to evaluate the performances of logit and NB models for derailment predictions.

2.4. Chapter Summary

A review of the literature revealed that none of the previous studies considered segment-level prediction models in rail safety. Researchers were constrained by system-wide data limitations and not able to consider, for example, individual segments within a rail network. The limitation of aggregated data is that the results may distort the actual safety performance of a rail network by failing to understand segment-level characteristics.

Since a rail track may traverse different areas nationally or even internationally and as these differences may range from a few kilometres to hundreds of kilometres in length, the level of safety could vary by segment. This issue requires a segment-level network screening method. The goal is to determine the levels of safety of different segments by capturing the variation in traffic exposures and rail characteristics. A methodological approach to segmentation is also required for derailment prediction models.

One of the crucial steps in statistical modelling is to define an appropriate analysis period. It was found that a number of rail safety studies have adopted very long analysis periods with the inclusion of 15 to 36 years of data. This approach may distort the performance of predictive modelling due to variation in input data and facilities over a long period of time. After considering the effects of the 2008 economic recession and the distribution of historical derailments, this thesis uses the most recent 10 years of complete data for prediction modelling.

The HSM is a guidance document developed for roadway safety assessment. There are currently no similar industry guidelines for evaluating the safety of rail facilities. Adopting the HSM approach, this thesis includes the development of safety performance functions (SPFs) for Canada's rail network as part of the safety network screening process. The results can be used to prioritize segments based on expected

levels of safety in terms of derailment risk. This study employs negative binomial and logistic regression modelling methods for derailment prediction.

Since segment-level data is available for Canada, this study has a unique opportunity. A network screening method that can estimate the levels of safety by segment produces greater resolution in derailment prediction. The results allow the identification of segments with the greatest derailment risk the most urgent need for special attention and consideration.

CHAPTER 3. STUDY DATA

This chapter describes the study data including its review and validation and database development process. Given the varying data sources, structure and formats, substantial data-cleaning was undertaken to develop an integrated database. The integrated database contained unique segments of the rail network. Data for each segment's key attributes was used in the prediction modelling. A descriptive data analysis was then conducted to identify patterns and trends in derailments.

3.1. Data Sources

Study data was obtained from three different government agencies in Canada: Transport Canada (TC), the Transportation Safety Board (TSB) and Natural Resources of Canada (NRC). **Table 5** summarizes each dataset. The three sets of data were thoroughly reviewed and validated for the purpose of developing an integrated derailment database containing both statistical and spatial information.

Table 5: Description of Study Data

No.	Dataset	Description	Source
1	Rail Network Shapefile[6]	Junctions, stations, marker posts (indicate distance along a rail track), etc.	Natural Resources of Canada
2	Rail Occurrence Database System (RODS)[7]	Historical derailment records and geospatial information	Transportation Safety Board of Canada (TSB)
3	At-Grade Inventory Database	Crucial traffic information such as daily train volumes and train speed	Transport Canada

A previous rail safety assessment for Canada identified some inconsistencies in the TSB database. Inconsistencies included missing information (coded as N/As) in the database which made it challenging to conduct accurate trend analysis and meant that assumptions had to be made. English and Moynihan (2007) recommended that TC should be more involved in trend analysis and the comparison of safety performance. For instance, a standardized reporting process could provide more accurate and more sufficient information for statistical analysis.

Transport Canada (TC) has since updated the Rail Occurrence Database System (RODS) with supplemental datasets and a standardized data structure. The RODS contains data and statistics associated with all reportable train accidents and incidents in Canada. The supplemental datasets include train data

[6] Canada's Open government Data Portal website: https://open.canada.ca/data/en
[7] Transportation Safety Board of Canada. (2018). Retrieved from http://www.bst-tsb.gc.ca/eng/stats/rail/data-24.asp

(e.g., train type, train control, signal data, etc.), rolling stock properties, injuries/fatalities and track components. The RODS also contains other variables such as environmental conditions, dangerous goods, on/off train injuries, etc., but the availability of such information is not consistent for all records. As an example, only 370 of 2,468 (15%) of main-track derailments have environment data. These omissions could be due to incomplete accident reports or errors in transcribing information to the electronic database.

The TSB also advised that there is still variability in the characteristics of the data due to changes in regulations and definitions over the years (TSB, 2019). It was therefore necessary to conduct a thorough review and validation of the various data sources, as discussed in **Section 3.2**.

3.2. Data Preparation and Validation

Figure 12 shows the study's data processing workflow. The integrated derailment database created in Microsoft Access (accdb) is GIS-compatible which allows efficient post-analysis of model results and visualization. A series of geoprocessing tools were used in resolving data gaps. Information from both RODS and the rail network shapefile was used to correct the locations of derailments in a GIS environment. A segmentation analysis was conducted by spatially joining the locations of point features (station, junction and at-grade crossing). The integrated database contains key variables required for predictive modelling which includes unique segment IDs, segment lengths, train volumes, segment speed and station counts.

Inconsistencies in data structure and format were the key challenges associated with data-cleaning. Cleaning the data involved four steps:

1. Develop a consistent rail network representation
2. Resolve spatial errors in rail occurrence database
3. Data extraction from at-grade crossing inventory database
4. Determine the appropriate track segmentation method

Details regarding these challenges and how they were addressed are provided in the next sections.

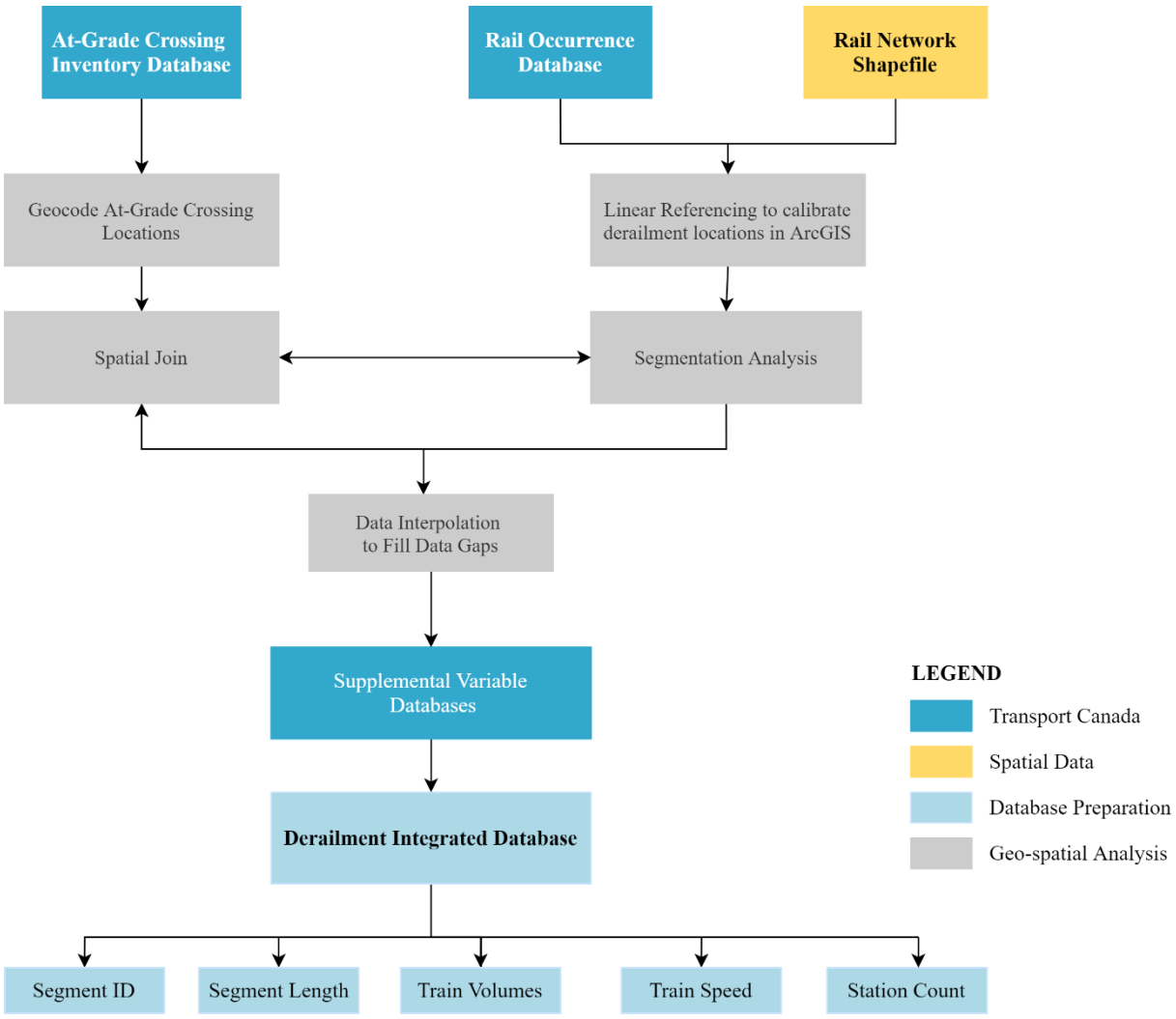


Figure 12: Dataset Development Process

3.2.1. Rail Network and Track Classification

The primary spatial data for this study was the National Railway Network[8] (NRWN) which is published in separate datasets classified by province and territory. In total, there are 11 pre-packaged NRWN datasets covering 9 provinces (BC, AB, SK, MB, ON, QC, NB, NS, and NL) and two territories (YT and NT). The rail network data is supplemented by additional GIS layers including at-grade crossings, junctions, marker posts, stations, and structure lines. These files were merged by feature type to represent Canada's national railway system. **Figure 13** shows the main-tracks and non main-tracks in the rail network.

[8] NRWN shapefiles were downloaded from http://ftp.maps.canada.ca/pub/nrcan_rncan/vector/geobase_nrwn_rfn/

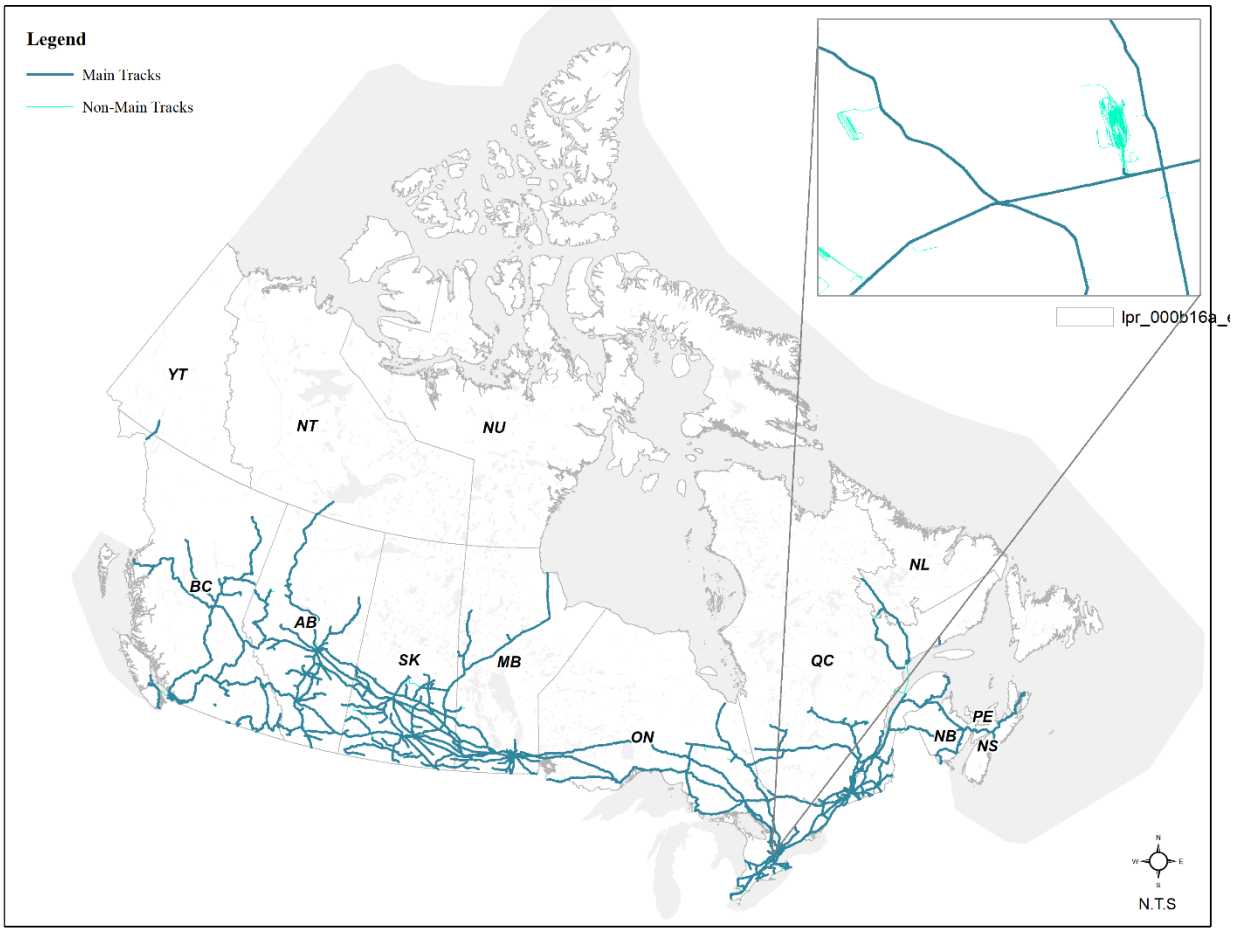


Figure 13: Rail Network in Canada

There are eight track classifications in the rail database: Connecting, Crossover, Ferry Route, Main, Siding, Spur, Wye, and Yard. There is also a small number of unclassified or 'unknown' segments. **Table 6** shows the number of features in each classification. This research included only derailments that occurred on main-track segments.

Table 6: Track Classification and Number of Features

Track Class	**Number of Features**
Connecting	565
Crossover	1,127
Ferry Route	17
Main	11,866
Siding	10,050
Spur	18,080
Wye	363
Yard	22,707
Unknown	17
Total	64,792

The NRWN data includes sections with multi-line tracks which share the same characteristics such as train volumes and locations for rail junctions, stations, marker posts, etc. As the network screening model would ultimately treat all rail tracks as a single feature when predicting the probability of derailments, multi-line tracks were converted to a single line track by using the "Merge Divided Road" tool in GIS. The modification process was applied to multi-line tracks that had the same subdivision name.

This resulted in a base network that is represented by the rail centrelines. **Figure 14** provides an example. **Figure 14 (a)** is an example of the rail network shapefile showing multi-line tracks and **Figure 14 (b)** shows the multi-line tracks converted to a single line track for the purposes of the prediction model.

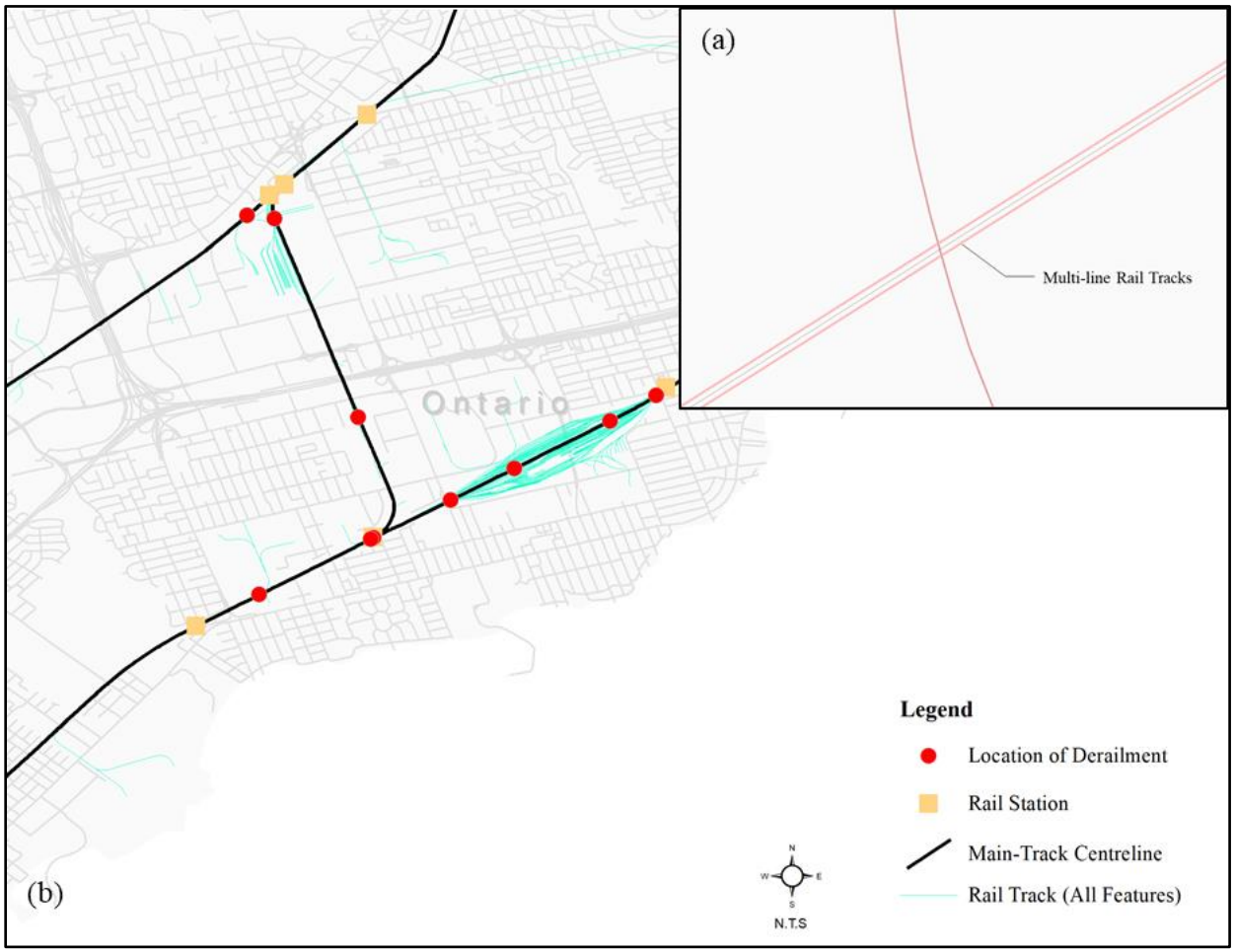


Figure 14: Example of Multi-Lane Tracks into a Single/Centreline Track Conversion

3.2.2. Rail Occurrence Database

The RODS contains accident data for the years 1983 to 2019 (the 2019 data is partially complete). The data includes 31 different types of reportable train related accident. This study used the following criteria to extract the relevant subset:

- Most recent 20 years of complete data (1999 to 2018); and
- Accident types related to main-track derailments that satisfy the following criteria:
    1. Derailment involving track unit[9];
    2. Derailment involving track unit (no damage);
    3. Main-track derailment; and
    4. Main-track derailment (no damage).

Table 7 shows the number of main-track derailments by type and year. There were 2,468 derailments in the 20-year period. In 2014, it appears that TC may have adopted a different classification that placed derailments with no damage into separate categories. This study simply focused on total annual derailments for the prediction modelling.

Table 7: Annual Main-Track Derailment Occurrence, 1999 to 2018

Occurrence Year	Derailment Involving Track Unit	Derailment Involving Track Unit (No Damage)	Main-Track Derailment	Main-Track Derailment (No Damage)	Total Derailment Count
1999	6		119		125
2000	1		122		123
2001	5		132		137
2002	2		124		126
2003	1		156		157
2004	2		160		162
2005	5		198		203
2006	1		139		140
2007	2		160		162
2008	5		129		134
2009	20		66		86
2010	11		82		93
2011	10		110		120
2012	7		67		74
2013	18		84		102
2014	7	7	100	2	116
2015	11	10	75	3	99
2016	7	13	58	6	84
2017	24	5	81	3	113
2018	16	5	87	4	112
Total	147	35	2,280	14	2,468

Figure 15 shows the location of derailments by province. Ontario (664) and British Columbia (511) had the highest numbers of derailments.

[9] A track unit is defined as a vehicle or machine capable of on-track operation utilized for track inspection, track work and other railway activities when on a track. (Transport Canada, 2019)

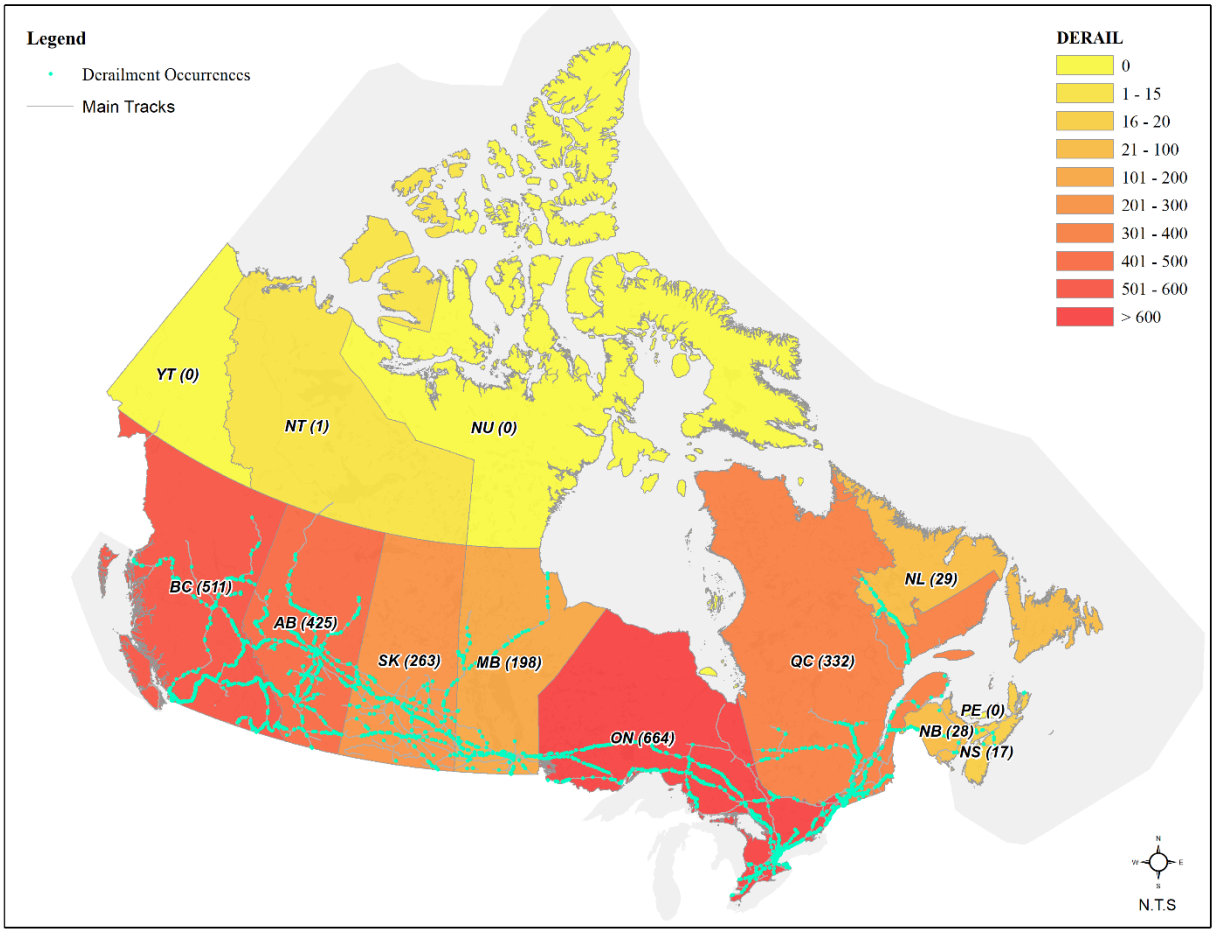


Figure 15: Derailments by Province, 1999 - 2018

The reference location of a derailment is measured according to the progressive distance along the track: each derailment is associated with the linear units that indicate its relative distance (in miles) from the point of origin of the track. Although the derailments could simply be geocoded using the RODS' x, y coordinates, many derailments appeared to be offset from the rail tracks by as far as 5 to 10 metres. To eliminate the offsets, the derailment locations were re-calibrated using linear reference attributes in the database. The linear referencing tool in GIS was used to re-calibrate the rail network as 'routes.' Derailments were coded as 'events' along the tracks. Before geoprocessing, two fields were created within the derailment database. The first was for the route identifier and the second was for the relative distance of a derailment event along a track segment. See **Figure 16**.

In **Figure 16,** the white and yellow columns show the original and processed attributes respectively and the blue columns provide the GIS attributes for linear referencing. The field 'fRoute_SubSE1' in the last column is a unique identifier of each track segment that concatenates the subdivision name ('SUBD1NAME') and the start and end mileage attributes ('SubdStart Mileage Processed' and 'SubdEnd Mileage Processed' fields).

If AH ≠ 0, then AP = AJ - AH

AF	AH	AI	AJ	AL	AM	AP	AQ	AR
SUBDI1NAME	SubdStart Mileage	SubdEnd Mileage	Subd Mileage	SubdStart Mileage Processed	SubdEnd Mileage Processed	Subd Mileage Adjusted	Subd Mileage to Meter	fRoute_SubSE1
Aberdeen	0	147.7	7	0	147.7	7	11,265.4	Aberdeen; 0-147.7
ADIRONDACK	20	49.1	45.4	0	49.1	25.4	40,877.3	ADIRONDACK; 0-49.1
ADIRONDACK	20	49.1	35.1	0	49.1	15.1	24,301.1	ADIRONDACK; 0-49.1
ADIRONDACK	20	49.1	45.5	0	49.1	25.5	41,038.3	ADIRONDACK; 0-49.1
ADIRONDACK	20	49.1	34.9	0	49.1	14.9	23,979.2	ADIRONDACK; 0-49.1
ADIRONDACK	20	49.1	41.9	0	49.1	21.9	35,244.6	ADIRONDACK; 0-49.1
ADIRONDACK	20	49.1	45.4	0	49.1	25.4	40,877.3	ADIRONDACK; 0-49.1
ADIRONDACK	20	49.1	42.02	0	49.1	22.02	35,437.8	ADIRONDACK; 0-49.1
ALBREDA	0	132.3	7.9	0	132.3	7.9	12,713.8	ALBREDA; 0-132.3
ALBREDA	0	132.3	8.5	0	132.3	8.5	13,679.4	ALBREDA; 0-132.3
ALBREDA	0	132.3	6	0	132.3	6	9,656.1	ALBREDA; 0-132.3
ALBREDA	0	132.3	69	0	132.3	69	111,044.7	ALBREDA; 0-132.3
ALBREDA	0	132.3	114.02	0	132.3	114.02	183,497.4	ALBREDA; 0-132.3
ALBREDA	0	132.3	103	0	132.3	103	165,762.4	ALBREDA; 0-132.3
ALBREDA	0	132.3	47.8	0	132.3	47.8	76,926.6	ALBREDA; 0-132.3
ALBREDA	0	132.3	43.7	0	132.3	43.7	70,328.3	ALBREDA; 0-132.3
ALBREDA	0	132.3	128.25	0	132.3	128.25	206,398.4	ALBREDA; 0-132.3
ALBREDA	0	132.3	103	0	132.3	103	165,762.4	ALBREDA; 0-132.3
ALBREDA	0	132.3	51.8	0	132.3	51.8	83,364.0	ALBREDA; 0-132.3
ALBREDA	0	132.3	116.5	0	132.3	116.5	187,488.6	ALBREDA; 0-132.3

"Subd Start Mileage" calibrated to start at 0

Figure 16: Excerpt from Derailments Table

The "Subd Mileage" represents the relative distance of derailment event with respect to a track segment. Since the original distance unit is in miles, the distance has been adjusted to meters (1 mile = 1609.344 meters). The converted distance is stored in the 'Subd Mileage to Meter' field. If the "SubdStart Mileage" started at 0, the conversion was straightforward. If the "SubdStart Mileage" did not start at 0, the 'Subd Mileage Adjusted' value was used for conversion.

Using the linear referencing technique, a route layer was then developed to represent the rail centreline and a 'Derailment Event Point' layer. This geoprocessing technique successfully eliminated the offsets issue. **Figure 17** provides an example of individual derailment points properly referenced to the rail centrelines.

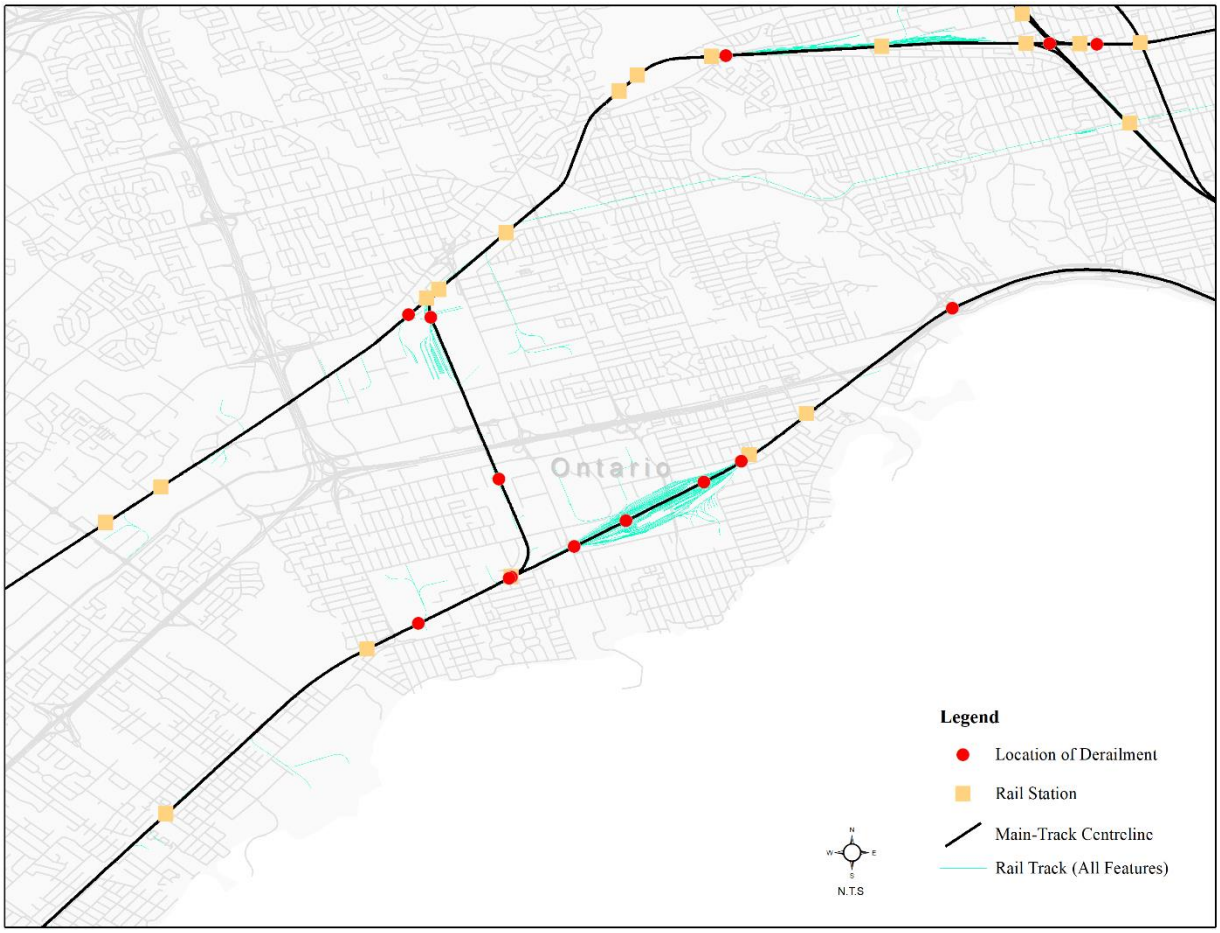


Figure 17: Derailments Geocoded in Reference to the Rail Track Centreline

3.2.3. At-Grade Crossing Inventory Database

The third dataset used in this study was the at-grade crossing inventory published by Transport Canada[10] (TC). This dataset contains information such as daily train traffic, maximum train speed, number of tracks, urban or rural environment, etc.

Table 8 is a snapshot of the data available in this dataset. Daily train volume, for example, is important as it can to be used as the exposure variable in prediction modelling. Relevant attributes from the at-grade crossing dataset were added to the integrated database by spatially joining the at-grade crossing layer to the rail centreline network.

[10] Grade Crossings Inventory, Transport Canada, 2019. Data retrieved from https://open.canada.ca/data/en/dataset/d0f54727-6c0b-4e5a-aa04-ea1463cf9f4c

Table 8: A Screenshot of Grade Crossing Inventory Database

	A	B	F	H	I	L	M	N	Q	R	S	T	U	V	W	X	Y	Z
1	Rank	TC Number	Access	Mile	Subdivision	Location	Latitude	Longitude	Accident	Fatality	Injury	Total Trains Daily	Vehicles Daily	Train Max Speed (mph)	Road Speed (km/h)	Lanes	Tracks	Urban Y/N
6	5	7902	Public	17.52	Kingston - CN	Woodland Avenue	45.42742	-73.88558	1	1	0	54	2700	100	50	2	2	Y
7	6	4861	Public	117.22	Drummondville	Rang Ste-Charlotte	45.6983	-72.775	1	2	0	27	5900	95	90	2	1	N
8	7	4856	Public	109.1	Drummondville	Rang De L'Eglise	45.7784	-72.6532	2	2	0	27	1130	95	90	2	1	N
9	8	8180	Public	263.45	Kingston - CN	D'Arcy St	43.96909	-78.15882	1	1	0	40	5000	95	50	2	3	Y
10	9	8008	Public	146.7	Kingston - CN	Prince St (Cnty 3)	44.40324	-76.01808	1	0	0	40	3500	100	50	2	3	Y
11	10	4840	Public	90.89	Drummondville	Route 122	45.9602	-72.3918	1	0	0	27	3900	95	90	2	1	N
12	11	7994	Public	124.88	Kingston - CN	Bartholomew St	44.6	-75.6816	2	0	0	40	2300	85	50	2	2	Y
13	12	9079	Public	18.51	Chatham - CN	Melbourne Rd - Cnty 9	42.8402	-81.5838	2	3	0	11	2333	95	80	2	1	N
14	13	10200	Public	13.23	Deux-Montagnes - C	11Th Avenue	45.5084	-73.8058	1	0	0	50	11650	65	50	2	2	Y
15	14	8130	Public	221.14	Kingston - CN	Moira St	44.1763	-77.3828	1	0	0	40	3000	95	50	2	2	N
16	15	13971	Public	69.51	St-Hyacinthe	Rue St-Georges	45.5034	-73.4936	1	0	0	55	6020	95	50	2	4	Y
17	16	5034	Public	76.44	Dundas	Egerton St	42.9862	-81.2147	1	0	0	52	11000	70	50	2	7	N
18	17	249	Public	67.98	Alexandria - VIA	Russell Rd 26	45.3751	-75.4968	1	1	0	17.5	3800	95	80	2	1	N
19	18	35213	Public	4.54	Emerson	Bishop Grandin Bouleva	49.84083	-97.07645	5	0	0	4	39984	25	80	2	1	Y

Figure 18 provides an indication of total daily train traffic in an urban area on the west side of Toronto. Note that train traffic is measured as the maximum number of trains (freight or passenger) traversing the at-grade crossing each day.

Figure 18: Sample Grade Crossing Locations with Total Daily Train Traffic

3.2.4. Track Segmentation Method

As discussed in **Chapter 2.2.4**, a methodological approach was needed to determine an appropriate track segmentation method for the study. The National Railway Network (NRWN) shapefile was developed by Natural Resources of Canada (NRC). The NRWN data is segmented by ending a segment when there is a change in attributes such as subdivision name, start and end mileage and administration areas (NRC, 2012). Each segment is associated with linear reference identifiers that correspond to the location attributes of

derailments (refer to **Chapter 3.2.2**). The NWRN also includes data on basic rail network features such as stations, junctions and at-grade crossings. **Table 9** provides the description of these three features.

Table 9: Description of Rail Network Features (NRC, 2012)

Feature	Description	Source
Station	Designated locations for train arrivals and departures	• Point geometry by NRC • Subdivision name and mileage by CN and VIA
Junction	Locations where two or more tracks diverge or converge	• Point geometry by NRC • Junction type and administration area data by Transport Canada
At-grade Crossing	Locations along rail network where a track crosses another network (at-grade only).	• Geometry and attribute by Transport Canada

Using multiple iterations of spatial join in GIS, the rail centrelines were segmented by stations, by junctions and by at-grade crossings and compared with existing segments. The results were analyzed by reviewing segment lengths and the distribution of derailments by segment. **Figure 19** shows the distribution of segment lengths and provide statistics for the four different segmentation approaches.

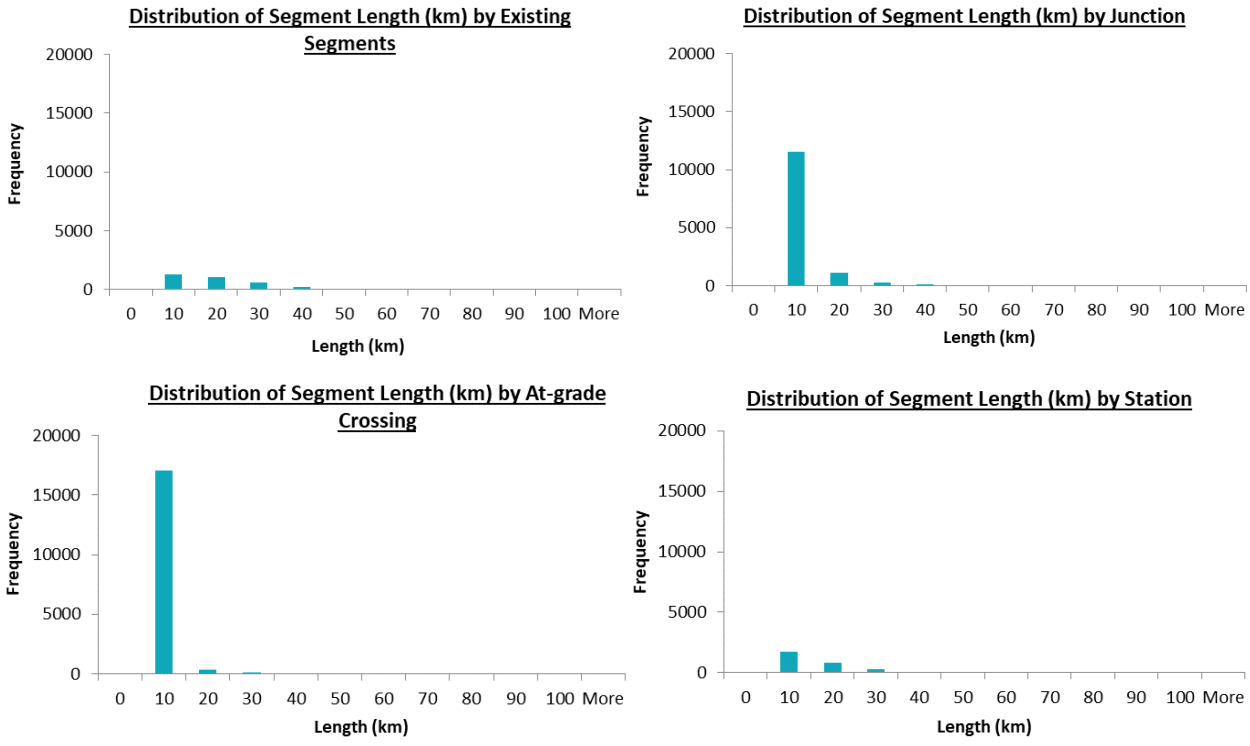


Figure 19: Distribution of Segment Length by Segment Method

Table 10 provides further detail including the number of derailments on the segments.

Table 10: Segmentation Method Analysis Results and Comparison

Method	Segment Length (km)				No. Of Segments		No. of Derailments		Pros	Cons
	Min	Max	Mean	SD	<100m	>10km	Max	Mean		
Existing	0.73	152.6	14.5	11.3	0	1,897	19	2	- low variation in length - least numbers of very short segment length (<100m) - contains measure values (start and end points) that correspond to derailment locations - lowest proportion of segments with zero derailments	- Over half of the segments have zero derailment (63%)
Junction	0.01	199.7	3.5	7.8	3,305	7	11	2	- More distributed locations: junctions are present throughout the network even in rural areas (with no station or crossing)	-Highest number of shorter segments (e.g. <10km) - Large proportion of segments have zero derailment (88%)
At-Grade Crossing	0.10	417.2	2.6	7.9	26	17	37	2	- Less numbers of very short segments compared to the junction method	- Very long segments in rural areas - Two segments with large number of derailments due to excessively long segments (> 150km) - Large proportion of segments have zero derailment (91%)
Station	0.05	417.2	14.7	28.1	58	53	65	2	- Resulted in reasonable number of segments (3,231) - Smaller proportion of segments with no derailment compared to junction and crossing methods	- Larger variation in segment length - Some segments are very long (>10km) due to dispersed stations in rural areas (e.g. Provincial parks, intercity/regional rail lines) - Over half of the segments have zero derailment (66%)

Notes: The analysis is based on 20 years of derailment data.
Existing segmentation (first row of Table) is based on the modified base rail network as documented in Chapter 3.2.1.

Table 10 shows that the split by junction method had the largest number of very short segments (less than 100m) and 88% of the segments have no derailment. Segmentation by at-grade crossings resulted in an overabundance of segments with zero derailments (91%). Segmentation by station shows a smaller percentage (66%) of segments with no derailment (66%) and fewer short segments. This resulted in high variability in segment length with certain entities being very long due to dispersed intercity/regional rail line stations in rural areas and provincial parks. Segmentation by at-grade crossing or station also produced very long segments (>400km) which may sacrifice homogeneity as variables are forced to be aggregated on these entities.

The NRWN's existing segments had the lowest number of very short segments (less than 100m) which is important as longer segments have been found more appropriate for developing safety performance functions (Cafiso et al, 2017). The existing segmentation also had the smallest proportion (63%) of segments with no derailments. In addition, and importantly, existing segments are associated with start and end points that correspond to the location attributes of derailments. This information is crucial for geocoding derailments using a linear referencing technique, as previously discussed in **Chapter 3.2.2**. In view of the pros and cons of each method, the existing segmentation was selected as the most appropriate for model development in this study.

3.2.5. Data Limitations

The study data had several limitations. Firstly, the RODS relies on reports provided by third-party sources. The type and amount of information provided by these sources is inconsistent. This could be due to different rail operators using different policies or standards for their reporting. The RODS database contains many fields with missing or incomplete data (e.g. missing weather, light or track curvature information). In some cases, information for these fields was missing for over 50% of the derailment records.

Secondly, since accident reports are provided directly by third-party sources, the information may not have been validated (TSB, 2019). The TSB conducts investigations on a small percentage of the reported accidents on an as-needed basis. RODS is being continuously updated as these investigations take place.

Thirdly, although each segment in the integrated database is associated with its characteristics (e.g., track characteristics and train specifications), many of the fields are poorly populated and do not provide sufficient samples sizes for prediction modelling. Track segment-related variables such as train weights, track geometry and type of signal operations cannot be included in the prediction models despite their potential effects on derailments.

3.3. Descriptive Data Analysis

This section provides a comprehensive descriptive analysis of the 2,468 main-track derailments (hereafter "derailments") that occurred in Canada over the 20-year study period (1999 to 2018). The analysis includes annual, monthly, seasonal, daily, and hourly trend analyses and analysis of derailment severity and track characteristics (e.g. length, ownership, volumes and speed).

3.3.1. Annual Trend

Figure 20 shows the number of derailments from 1999 to 2018. Only seven of the 2,468 derailments (less than 0.3%) involved passenger trains. None of the passenger train derailments resulted in injuries.

On average, there were 123.8 derailments annually. The highest number of derailments occurred in 2005, accounting for approximately 8% (203) of all derailments recorded in Canada during the study period. With the exception of 2002, the Figure shows an increasing trend from 1999 to 2005 with a below average number of derailments from 2009 to 2018.

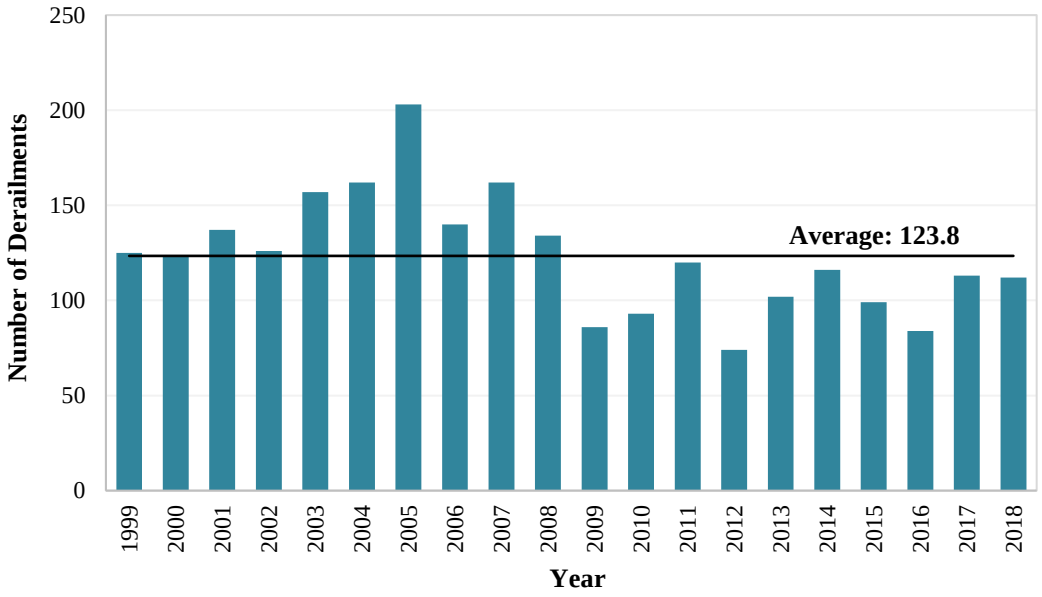


Figure 20: Annual Number of Main-Track Derailments, 1999-2018

3.3.2. Monthly and Seasonal Trends

Figure 21 shows the number of derailments by month. The highest number occurred in July (259 or 10%) closely followed by the winter months from January to March. It is likely that winter derailments were associated with weather conditions such as snow or ice accumulation. October (153 or 6%) and November (152 or 6%) had the lowest number.

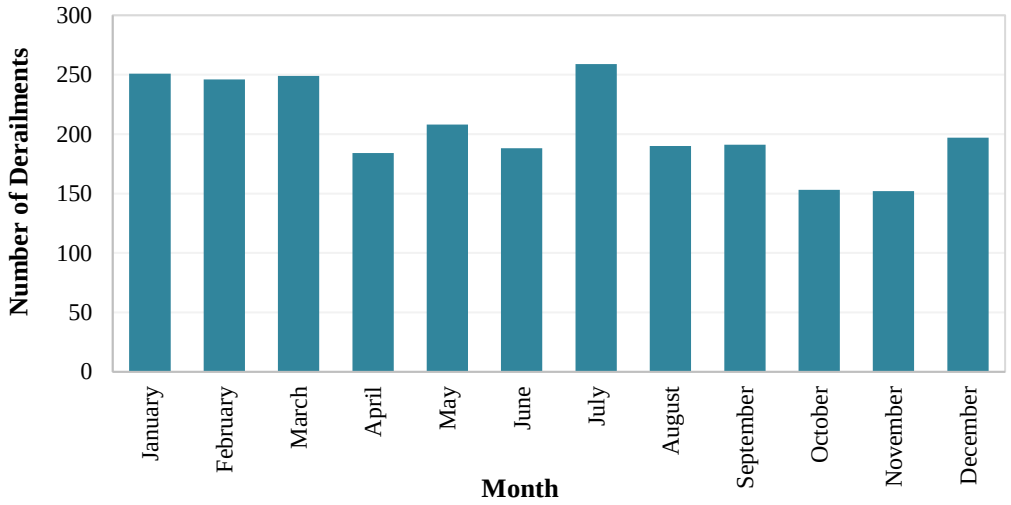


Figure 21: Number of Main-Track Derailments by Month, 1999-2018

Figure 22 shows the seasonal distribution of derailments. The highest number was in winter (694 or 28%), but spring (641 or 26%) and summer (637 or 26%) had nearly as many as winter. Fall had the lowest number of derailments (496 or 20 %).

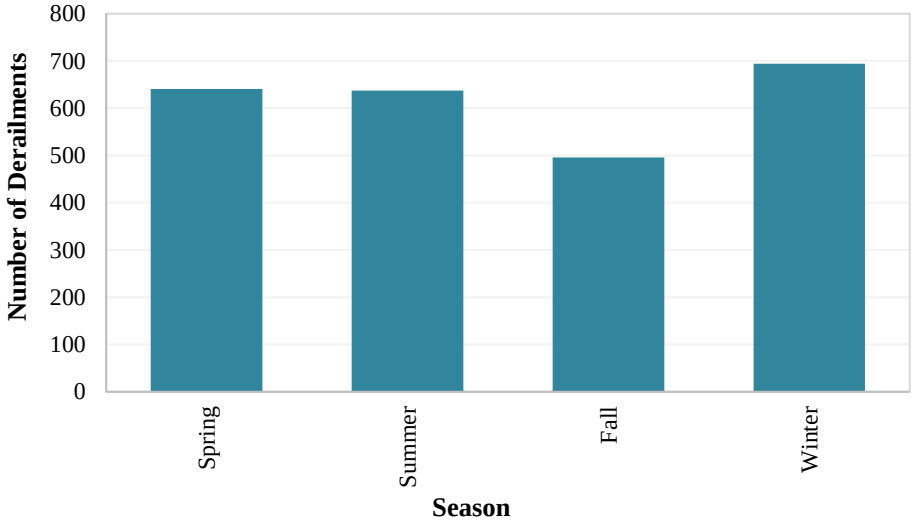


Figure 22: Number of Main-Track Derailments by Season, 1999-2018

3.3.3. Weekly and Hourly Trends

Figure 23 shows the number of derailments by day of the week. Fridays had the highest number (385 or 16%) followed by Tuesdays (361 or 15%) and Wednesday (349 or 14%) derailments. Saturday had the lowest number (324 or 13%).

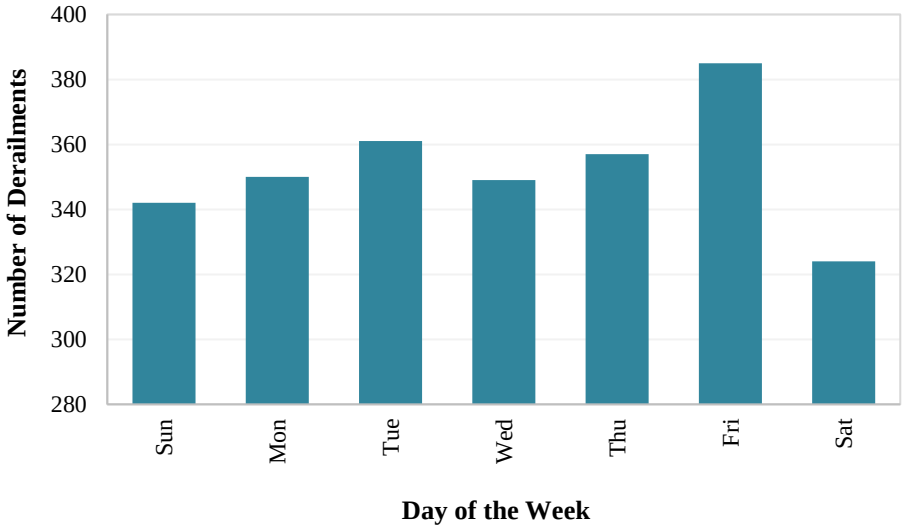


Figure 23: Derailments by Day of the Week

Figure 24 shows the number of derailments by time of the day. Derailments occurred throughout the day with the highest number from around noon to around 4 pm.

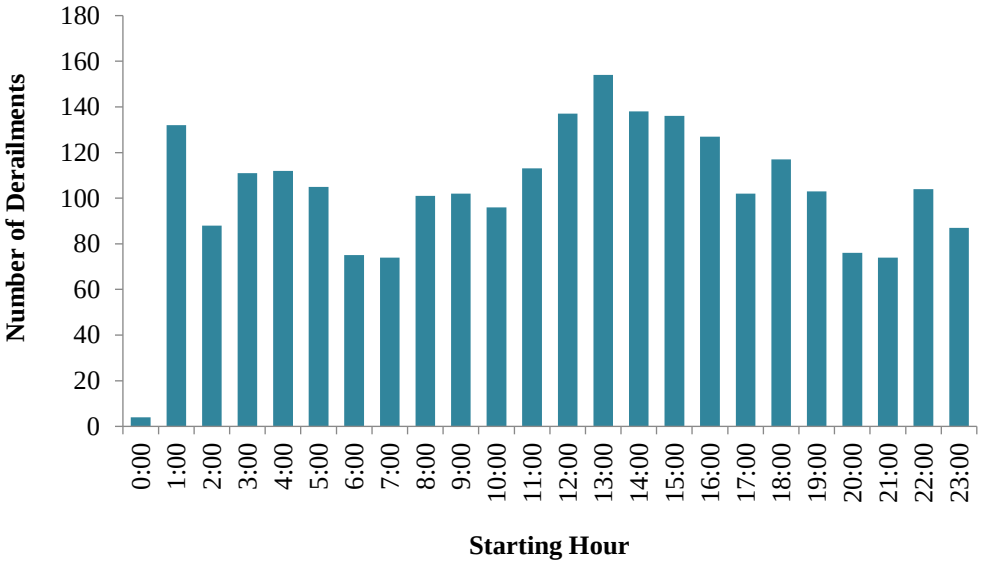


Figure 24: Hourly and Daily Trends of Derailments, 1999-2018

3.3.4. Distribution of Railcars in Derailments

The severity of derailments is typically measured by the number of railcars involved. **Table 11** shows the number of railcars involved in derailments from 1999 to 2018. **Figure 25** presents the same information. Almost half (1,163 or 47%) of the derailments involved only one railcar and the great majority (93%) involved one to 20 railcars. The most severe derailments involved more than 40 railcars, but accounted for only 1% of total derailments.

Table 11: Number of Railcars Involved in Derailments, 1999-2018

Number of Railcars	Frequency	Distribution
1	1,163	47%
2 - 3	346	14%
4 - 5	180	7%
6 - 10	349	14%
11 - 20	270	11%
12 - 30	101	4%
31 - 40	43	2%
>40	16	1%
Total	2,468	100%

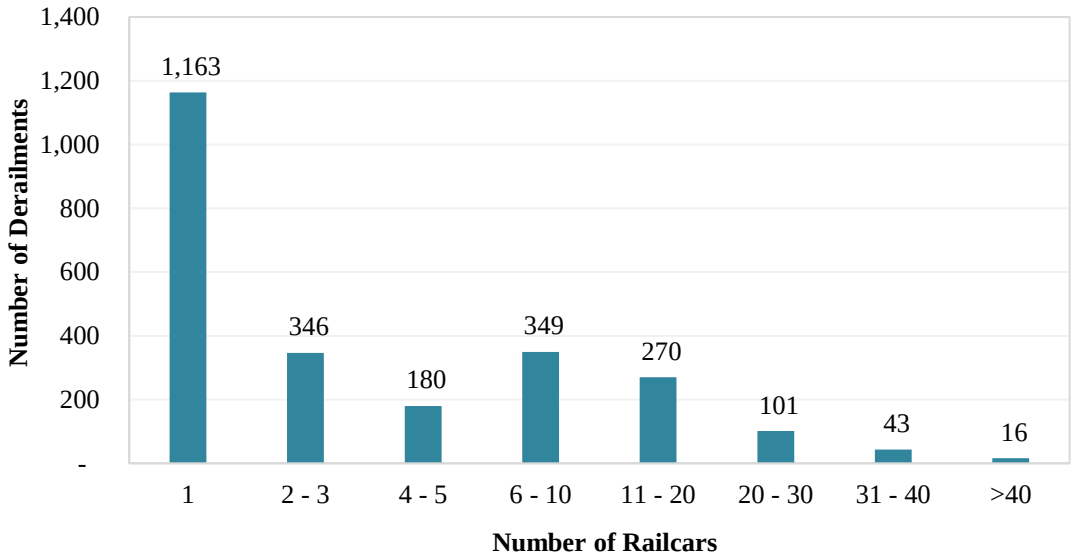


Figure 25: Distribution of Number of Railcars Involved in Derailments, 1999-2018

3.3.5. Derailments by Province and Track Length

Figure 26 shows the number of derailments by province (blue bars) and track length (black line). Ontario had the highest number of derailments (664 or 27%) and highest number of kilometres of rail track. British Columbia had the second higher number of derailments (511 or 21%) followed by Alberta (425 or 17%). The lowest numbers of derailments were noted for northern Canada (Yukon, Northwest Territory and Nunavut) attributable to the lack of rail infrastructure.

In most cases, the number of derailments was proportional to track length. Saskatchewan was an exception as it had the second highest of kilometres of rail track, but a relatively low number of derailments. The flat topography of Saskatchewan may help to explain this. Mountainous provinces such as British

Columbia, have more challenging operating conditions due to changes in elevation (steep slopes) and sharp curvatures.

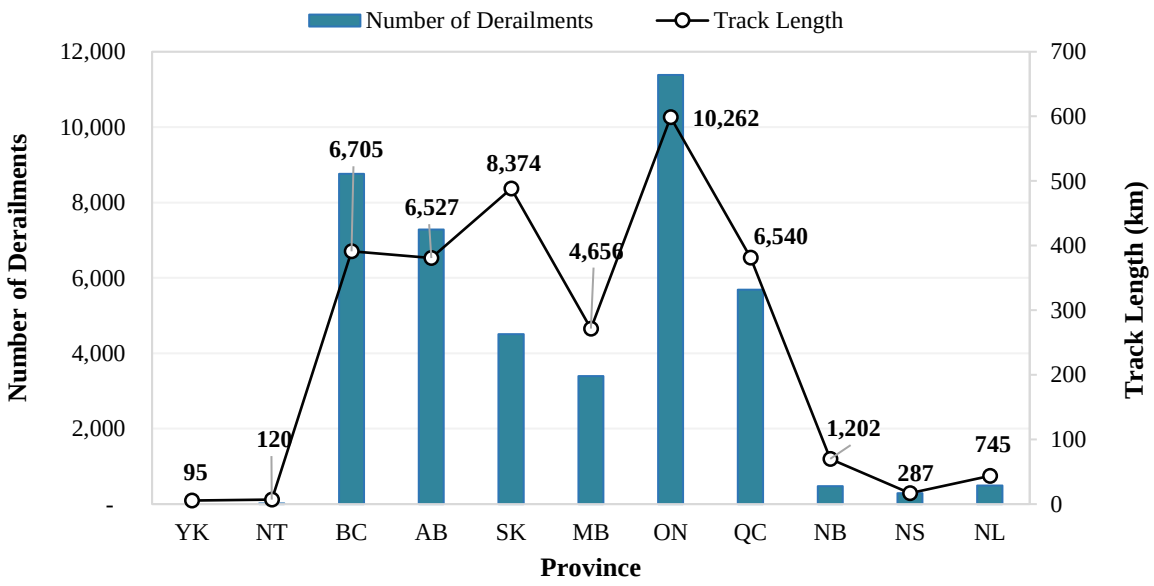


Figure 26: Number of Derailments by Province and Territory, 1999-2018

Figure 27 provides a more accurate representation of the data for Derailments by Province and Territory as the number of derailments is normalized by track length. British Columbia had the highest derailment rate per kilometre of track followed by Ontario and Alberta. The derailment rate per kilometre of track for Saskatchewan was clearly lower the national average, but not as low as the average for New Brunswick, the Northwest Territories or the Yukon. The length of a track segment has a direct relationship with derailments and should be included as one of the independent variables in prediction modelling.

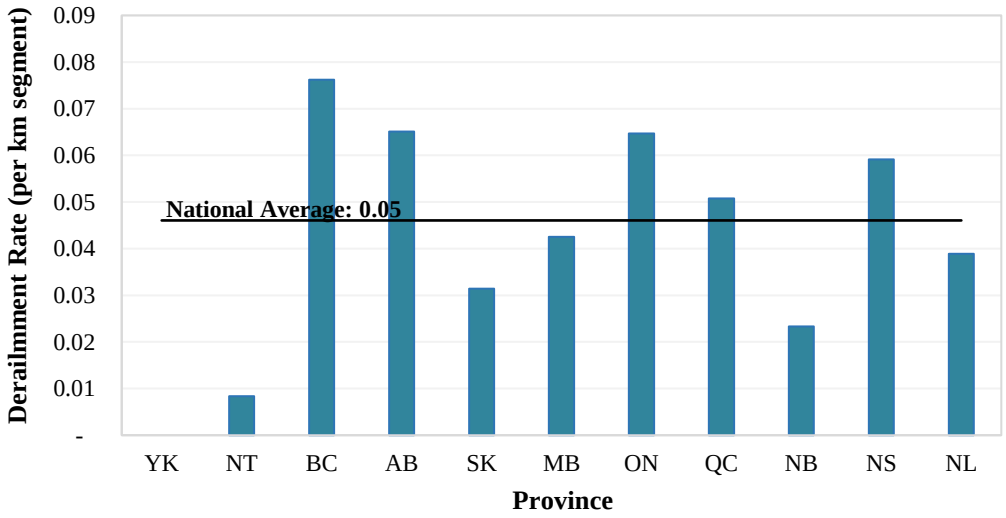


Figure 27: Derailment Rate per Kilometre Track by Province, 1999-2018

3.3.6. Derailments and Track Ownership

Table 12 shows the number of derailments with respect to track ownership. Canada National (1,269 or 51%) and Canadian Pacific Railways (888 or 36%) accounted for most of the derailments. These two companies also own most of the track in Canada's rail network.

Table 12: Derailments by Track Owner, 1999-2018

No.	Rail Company	Derailment
1	CN - Canadian National Railway Co.	1,257
2	CP - Canadian Pacific Railway Co.	898
3	Quebec North Shore and Labrador Railway Co.	73
4	Hudson Bay Railway	64
5	Montreal, Maine and Atlantic Railway	23
6	Ottawa Valley Railway	22
7	Mackenzie Northern	21
8	Okanagan Valley Railway	15
9	Goderich-Exeter Railway Company Ltd.	11
10	St. Lawrence and Atlantic (Quebec) Inc.	9
11	Chemin De Fer De La Matapedia Et Du Golf Inc	8
12	Ottawa Central Railway	8
13	Via Rail Canada Inc.	6
14	Burlington Northern Santa Fe Company	5
15	CSX Transportation	5
16	Kelowna Pacific Railway Ltd.	5
17	Tshiuetin Rail Transportation Inc.	5
18	Essex Terminal Railway Company	4
19	Algoma Central Railway	3
20	Canadian American Railroad Company	3
21	Capital Railway	3
22	Corporation Des Chemins De Fer De La Gaspesie	3
23	Southern Ontario Railway	3
24	Baie Des Chaleurs	2
25	Sydney Coal Railway	2
26	Arnaud Railway	1
27	GO - Metrolinx	1
28	Knob Lake and Timmins Railway Company Inc	1
29	New Brunswick Southern Railway	1
30	Nipissing Central Railway (Ontc)	1
31	Ontario Northland Railway	1
32	Southern Railway of BC	1
33	Toronto Terminals Railway Company	1
34	White Pass and Yukon	1
35	Windsor and Hansport Railway Company Limited	2
	Total	2,468

3.3.7. Derailments by Train Volume

Daily train volumes were estimated from the at-grade crossing inventory database as mentioned in **Chapter 3.2.3**. **Figure 28** shows the number of derailments and number of segments for seven levels of daily train volume. The Figure shows that most segments have daily train volumes of up to 10 trains per day. The Figure also shows that and most derailments (2,405 or 97%) occurred on segments with up to 40 trains per day. Less than one percent of derailments (12 or 0.5%) occurred along segments with a high daily train volume (61 trains or more). Train volume is an exposure variable that is critical for predicting the probability of derailment along segments.

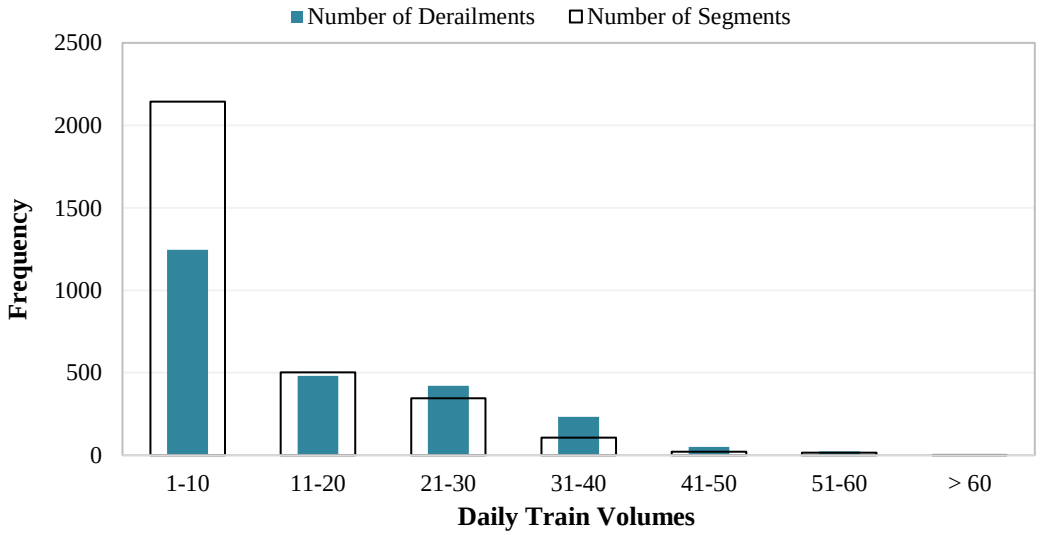


Figure 28: Derailment and Train Volumes on Segments (1999-2018)

3.3.8. Derailments by Train Speed

Segment speeds (mph) were estimated from the at-grade crossing inventory database as discussed in **Chapter 3.2.3**. **Figure 29** shows the estimated segment speeds in the network. The speeds shown are the maximum train speeds recorded at rail crossings in 2019. These speeds represented the speed for each track segment. The high speeds (shown in red) are mainly for freight routes in the Greater Toronto Hamilton Area and southern Ontario and VIA's interprovincial routes across Alberta, Saskatchewan and Manitoba connecting eastern and western Canada.

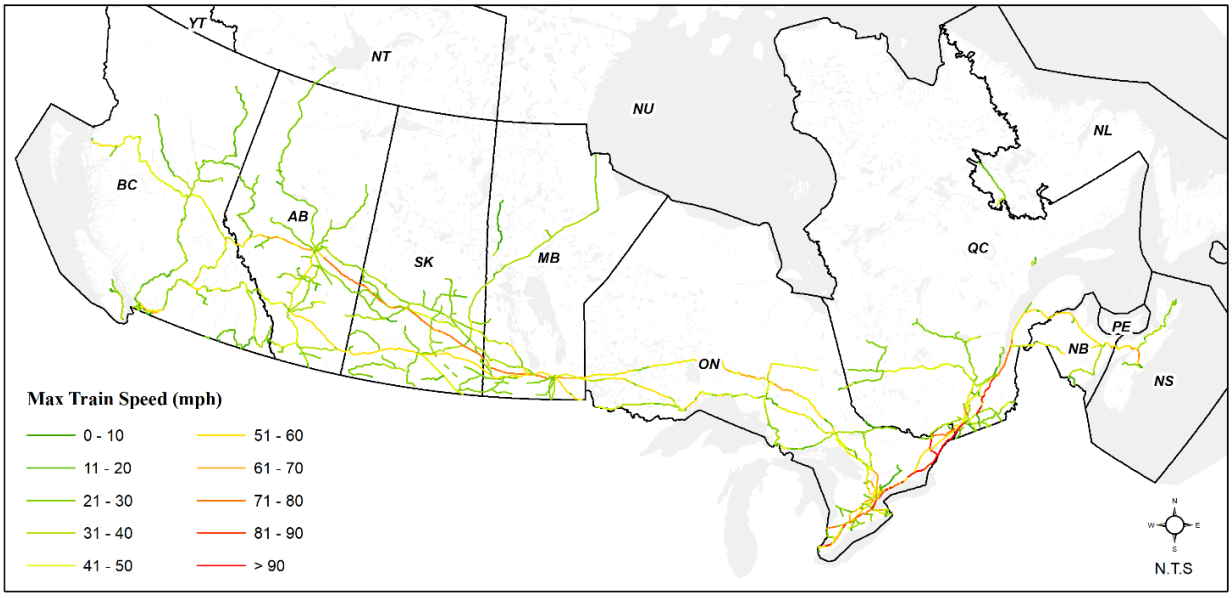


Figure 29: Maximum Train Speeds (mph) Recorded at At-Grade Crossings in Canada, 2019

Figure 30 shows the number of derailments by train speeds along segments. Most derailments (1,855 or 83%) occurred on segments with recorded maximum train speeds of 21 mph to 60 mph. The number of derailments on lower and higher speed segments was much lower. On Class 1 tracks, the allowable speed for freight trains and passenger trains is 10 mph and 15 mph respectively. It is clear that the majority of segments (2,515 of 3144 or 88%) recorded trains operating at excessive speeds.

As operating at a speed unsuitable for the design conditions may elevate derailment risks, segment speed was included as an independent variable in the prediction model.

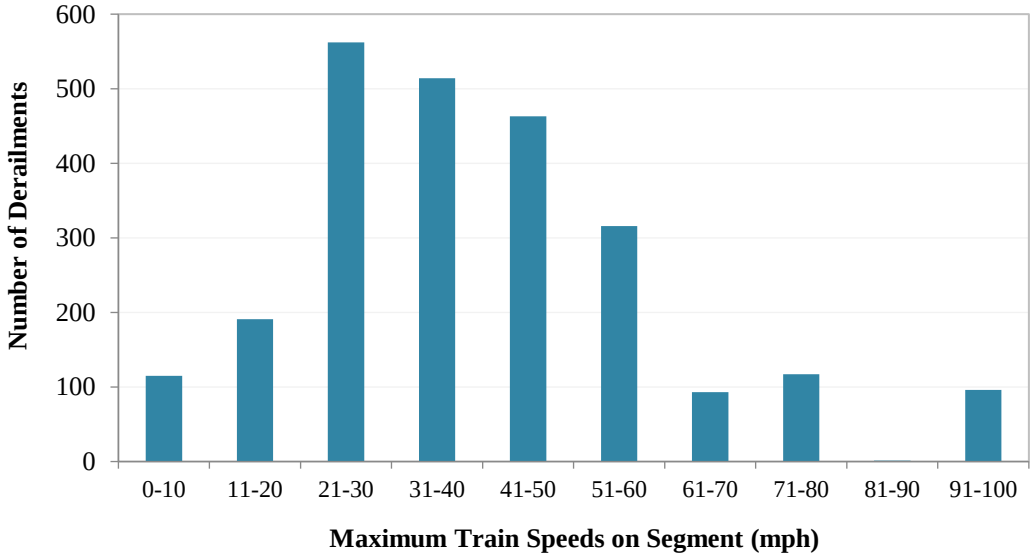


Figure 30: Derailment and Segment Train Speeds

3.4. Chapter Summary

The study data used in this research was published by Natural Resources of Canada, Transport Canada and the Transportation Safety Board of Canada (TSB). Significant effort was spent on data quality control, testing and validation in order to synthesize the different datasets into an integrated database. Challenges with the data integration included spatial data discrepancies (resolved through geoprocessing tools), missing/incomplete data (resolved through data analysis), and the absence of an appropriate segmentation method (also resolved through data analysis).

The integrated database contained 3,144 segments. In the case of multi-line tracks, the segments were represented by the centreline. Each segment in the database was associated with key attributes such as segment identifier, total derailment count, train volume, train speed, and number of stations.

Several data limitations arose. For example, missing/incomplete information limited the number of independent variables available for the predictive modelling. Some data was inconsistent. For example, the Rail Occurrence Database System (RODS) obtains its data from accident reports submitted by individual rail operators. As only a small portion of the accidents warrant investigation by TSB, much of the RODS data RODS has not been verified.

A descriptive analysis was conducted to examine patterns and trends in main-track derailments. The analysis of derailments included consideration of the month, the season, the day of the week, the hour of the day, severity (in terms of the number of railcars involved), the province, track length, track owner, train volume, train speed

From 1999 to 2018, the total number of main-track derailments in Canada was 2,468, an average of 123.8 per year. Almost all (97%) were freight trains. The number of the derailments was higher than average from 1999 to 2008, and lower than average from 2009 to 2018. The number of the derailments tended to be higher in July and from January to March, higher in winter, higher on Fridays, and from around noon to around 4 pm. The great majority of derailments (93%) involved one to 20 railcars, and 47% involved only one railcar. Ontario and British Columbia accounted for 48% of the derailments. Ontario had the most track. Saskatchewan had the second most track (but a relatively low number of derailments). Canada National and Canadian Pacific Railways own most Canadian rail track and had most (87%) of the derailments. Most derailments (97%) occurred on segments with from one to 40 trains per day. Most derailments (83%) occurred on segments with recorded maximum train speeds of 21 mph to 60 mph.

The data analysis revealed that track length, segment speed and train volume may be key variables that contribute to derailment potential.

CHAPTER 4. MODELLING METHODOLOGY

This Chapter details the steps taken prior to model development. The main issues were input data preparation, outlier detection, model descriptions, model form, model calibration and validation, goodness-of-tests, and variable selection. The analysis period for model development is 10 years as discussed in **Chapter 2.2.5**.

4.1. Input Data Preparation

This section provides a description of the input dataset and outlier detection methods.

4.1.1. Descriptive Statistics

The integrated database consisted of 3,144 segments. Each segment had a unique identifier, derailment count, and information on the segment's attributes. The total derailments over the 10-year analysis period was used as the dependent (response) variable. In addition to the three variables mentioned in previous Chapter, two other variables were considered: average train speeds and the number of stations along a segment. The number of stations along segment has implications on derailment potential due to variation in train speed caused by acceleration and/or deceleration. In total, five independent (explanatory) variables were considered for predicting derailments. **Table 13** provides descriptions of the model variables.

Note that "VL" in the variable names for Maximum Daily Train Volumes (VL_Count) and Maximum Train Speed (VL_TrainSpeed) simply refers to the at-grade crossing dataset that provided the data on train volume and speed.

Table 13: Model Variable Names and Descriptions

Variables	Description
Dependent	
DerailCount	Number of Derailments
Independent	
Seg_Length	Segment Length (km)
VL_Count	Maximum Daily Train Volumes
VL_TrainSpeed	Maximum Train Speed (mph)
Avg_Train	Average Daily Train Volumes
Stn_Count	Number of Stations on Segment

4.1.2. Outlier Detection

Anomalies in the database can significantly alter or bias the fit of a prediction model. It is important to identify any outliers in the dataset and determine a robust method for rectifying them. **Figure 31** is a set of boxplots for the five independent variables: (Segment Length (Seg_Length), Maximum Train Speed (VL_TrainSpeed), Maximum Daily Train Volume (VL_Count), Average Daily Train Volume (Avg_Train), and Station Count (Stn_Count).

Outliers were defined as values that were greater than the upper bound and lower than the lower bound of each independent variable. **Figure 31** shows a number of outliers for each variable.

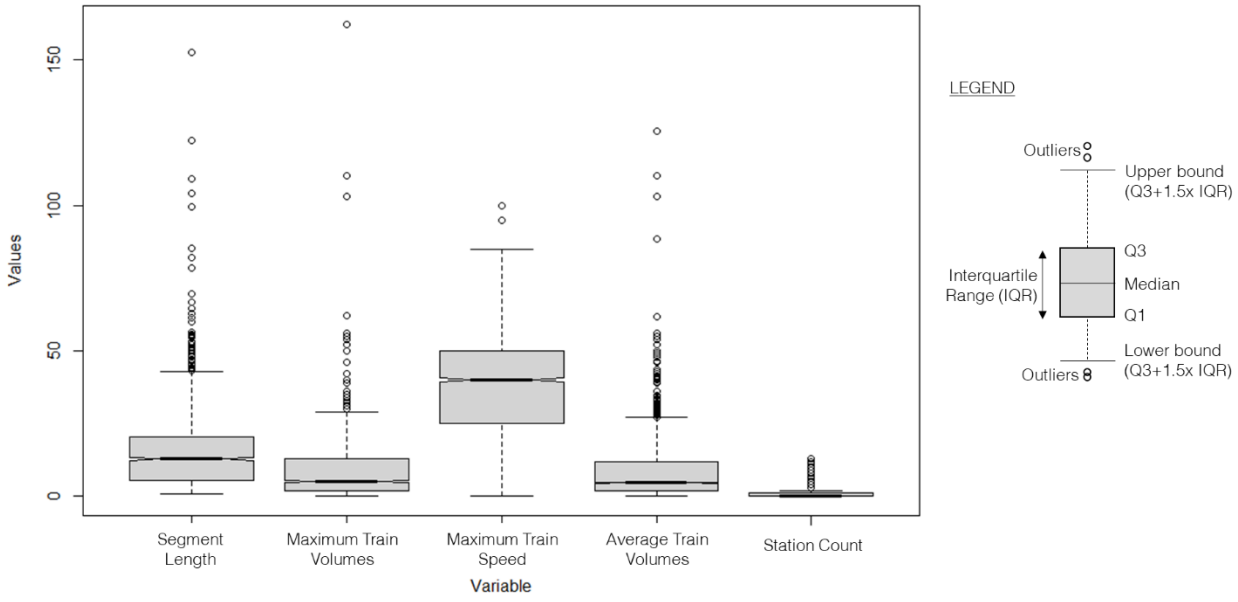


Figure 31: Box Plot for each Independent Variable

Table 14 provides a summary of the statistics used in Figure 31 for the five independent variables.

Table 14: Summary Statistics for Independent Variables

	Seg length (km)	**Train Speed (mph)**	**Max Train Vol**	**Avg Train Vol**	**Station Count**
Minimum	1	0	0	0	0
Quartile 1 (Q1)	5	25	2	2	0
Median	13	35	5	5	0
Quartile 3 (Q1)	20	50	13	12	1
Maximum	153	100	162	126	13
Mean	14.48	37.86	9.47	8.72	0.80
Range	152	100	162	126	13
Interquartile Range (IQR = Q3-Q1)	15	25	11	10	1
Upper Bound (Q3-1.5 x IQR)	42.99	87.50	29.50	27.00	2.50
Lower Bound (Q1-1.5 x IQR)	-17.08	-12.50	-14.50	-13.00	-1.50

4.2. Model Descriptions

Homogeneity plays an important role in developing derailment prediction models. This is because infrastructure with similar characteristics is assumed to have a similar level of safety performance which can be characterized by a set of independent variables. Given the great contrasts in the geographic and track characteristics of a vast rail system like Canada's, developing one model for the whole country might give a misleading impression of the level of safety of the network. Exposure (e.g., daily train volumes) and segment length are examples of independent variables that vary greatly across the country.

As a single model form is clearly unlikely to be able to account for the variation in characteristics in different provinces and territories, multiple derailment prediction models were developed to capture the variation in rail network characteristics more precisely. Discussion below provides a quantitative assessment of the number of prediction models to be developed and the justification for each.

Figure 32 shows the number of derailments that occurred on the network's 3,144 segments in the 10-year study period. The number of derailments per segment ranged from 0 to 8 with a mean of 0.31 and a variance of 0.56. A variance greater than the mean indicates over-dispersed data. Although the great majority of segments (79%) had no derailments, 470 segments had two, and a few segments had three or more. The study's statistical analysis and model development was designed to investigate how the independent variables might be associated with potential derailment risks on all 3,144 segments.

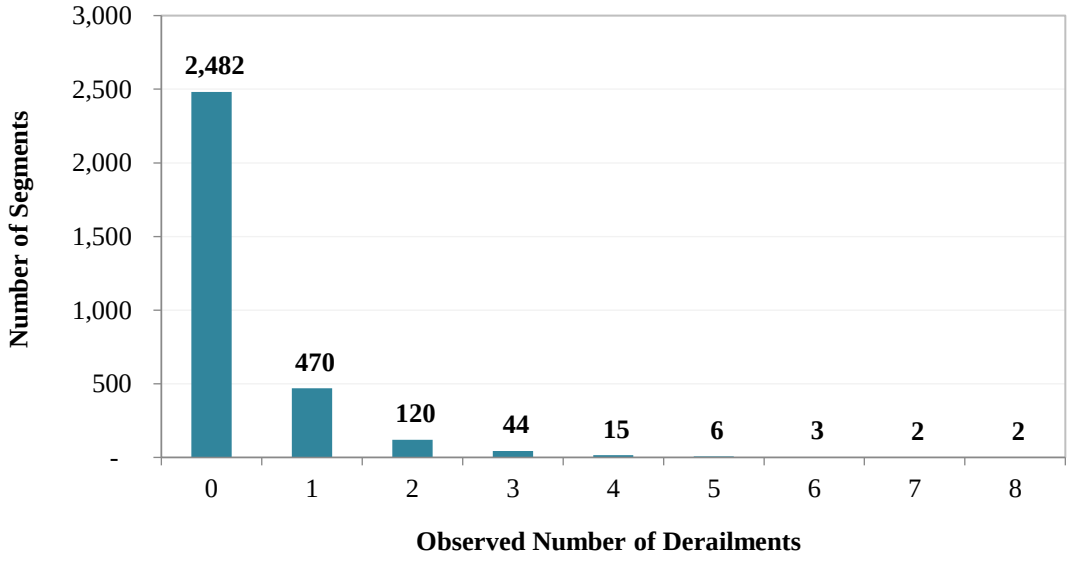


Figure 32: Number of Derailments by Number of Track Segments, 2009-2018

A large number of entities (track segments) and outcome events (derailments) in a dataset is ideal as it maximizes the performance of the prediction model. Small sample sizes may yield biased estimates of the regression coefficients and compromise the predictive power of the model. For example, if we develop derailment prediction models separately for each province/territory, some provinces and territories would have such small sample that we cannot develop statistically sound derailment prediction models. For example, the Northwest Territories, Nova Scotia and New Brunswick had only 11 derailments during the study period. To ensure a good sample size, the study developed models for eastern Canada and western Canada. **Table 15** and **Figure 33** summarize the main rail system characteristics of Eastern and Western Canada. **Table 15** also provides a breakdown for the provinces and territories.

Table 15: Rail System Characteristics of Eastern and Western Canada

Province and Territories	Track Length (km)	Daily Train Traffic	Daily Train Millions KM Travelled	Derailments (2009-2018)
New Brunswick	1,202	589	0.7	8
Newfoundland and Labrador	287	40	0.0	19
Nova Scotia	745	399	0.3	3
Ontario	10,262	9,416	96.6	230
Quebec	6,540	3,167	20.7	119
Alberta	6,527	3,704	24.2	196
British Columbia	6,705	6,193	41.5	203
Manitoba	4,656	2,404	11.2	80
Northwest Territories	120	4	0.0	-
Saskatchewan	8,374	3,848	32.2	122
Yukon	95	4	0.0	-
Eastern Canada	19,036	13,612	259	379
Western Canada	26,477	16,158	428	601
Canada Total	45,513	29,769	1,354.9	980

Eastern Canada: New Brunswick, Newfoundland and Labrador, Nova Scotia, Ontario and Quebec.
Western Canada: Alberta, British Columbia, Manitoba, Northwest Territories, Saskatchewan and Yukon.

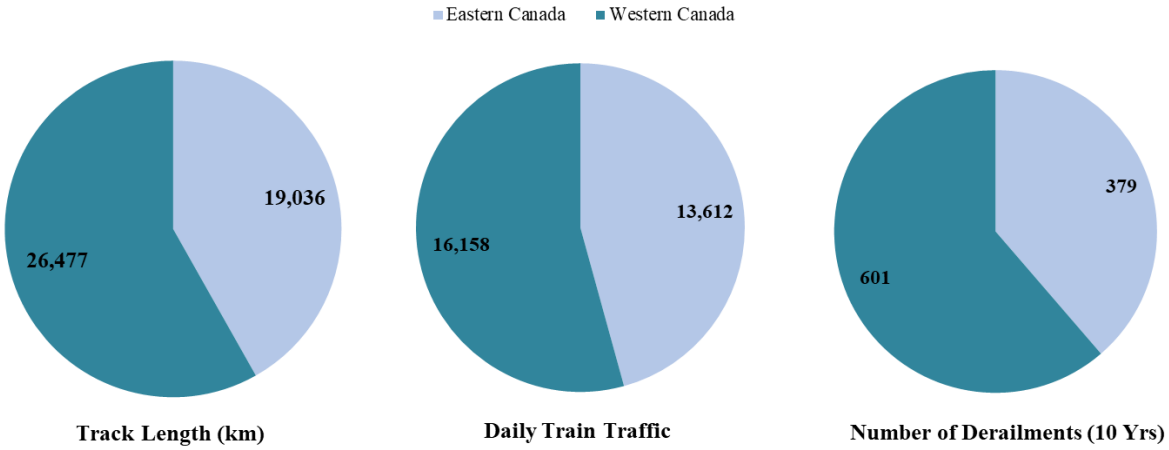


Figure 33: System Characteristics for Eastern and Western Canada

Western Canada has more rail track, higher train traffic and a higher number of derailments than does Eastern Canada. Both regions provide enough derailments and segments for prediction modelling.

Since the rail network in Canada is primarily owned by CN and CP, this study developed prediction models for these two railway companies. **Figure 34** shows the maps of rail track ownership in Canada. Approximately 86% of the derailments occurred on CN (47%) or CP (39%) owned tracks during the 10-year analysis period. These percentages might lead to a false perception of safety performance if simply taken at face value. Derailments can be normalized by track length to check whether the proportion of derailments is associated with track ownership.

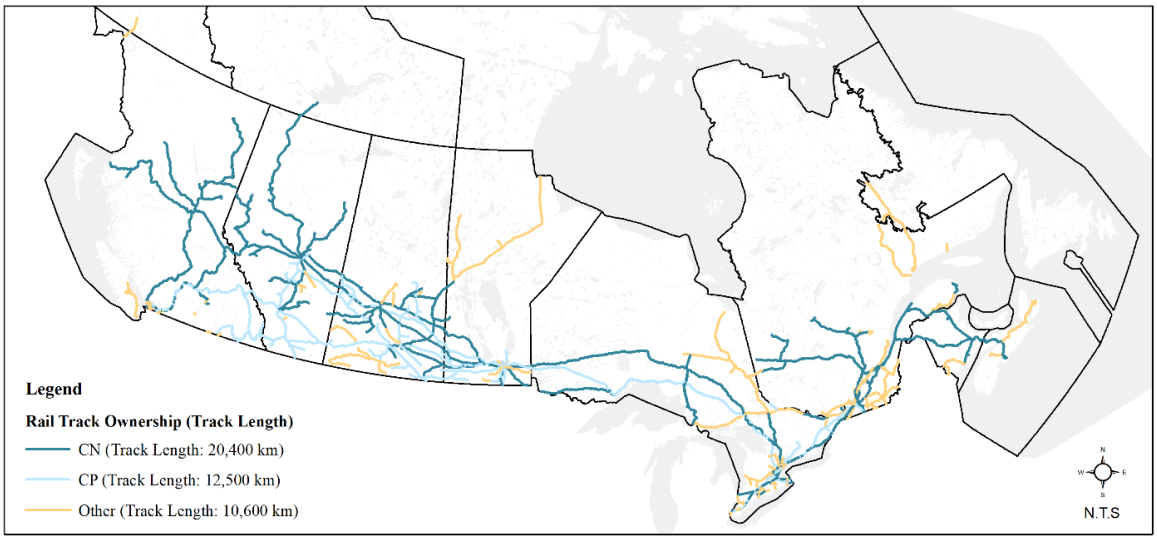


Figure 34: Rail Track by Ownership

Figure 35 shows derailments normalized by track length. The Figure shows that CP had higher derailment rates than CN in every year except 2009, and that both CN and CP had higher numbers of derailment rates other railway companies in all 10 years.

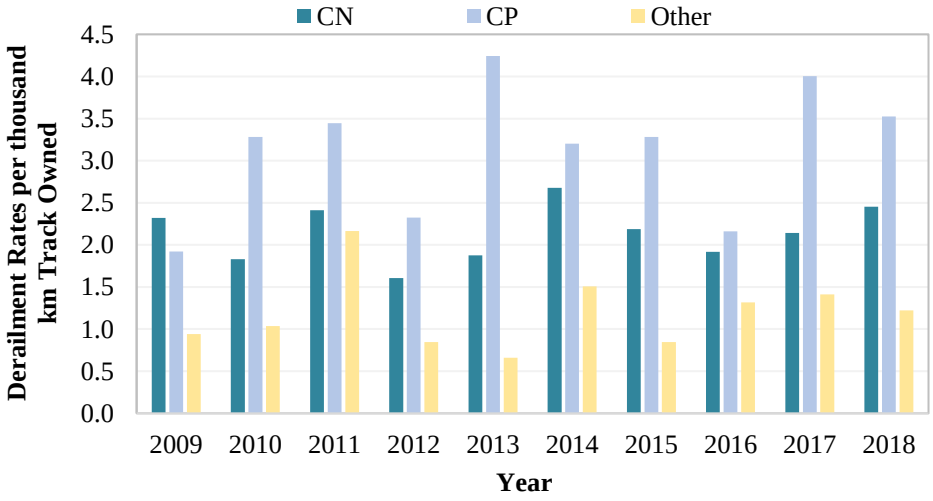


Figure 35: Derailments per Thousand Kms of Track Owned (2009-2018)

As CN and CP are likely to have different policies and standards for maintenance and operations, it appeared useful to examine the safety performance of their rail networks separately. The results obtained from the prediction models could provide useful information for improving their safety improvement strategies.

Consideration of the data and the sample sizes available suggested that the study should develop five models:

1. Canada
2. Eastern Canada
3. Western Canada
4. Canadian National Railway
5. Canadian Pacific Railway

4.3. Model Form

This study used negative binomial (NB) regression distribution to evaluate the safety performance of each segment. The functional form of the NB models for each track segment *i* is shown in **Equation 9** for the expected mean (μ_i) which is then re-expressed in **Equations 10** and **11** (Tayki et al., 2018):

$$\ln(\mu_i) = x_i\beta + \varepsilon_i \quad \text{(Eq.9)}$$

$$\mu_i = exp\left(x_{i1}\beta_1 + x_{i2}\beta_2 \ldots x_{ij}\beta_j + \varepsilon_i\right) \quad \text{(Eq.10)}$$

$$y_i \sim x_{ij}\beta_k \quad \text{(Eq.11)}$$

where:

y_i is the observed value of the dependent variable (i.e., number derailment) for segment i,

x_i is the independent variable for segment i,

β is the regression coefficient estimated from the model,

ε_i is an error term with a mean of 1 and variance α, and

y_i follows a negative binomial distribution.

An error term (ε_i) is introduced to account for the issue of overdispersion as NB regression distribution assumes that the variance and the mean are not equivalent.

The probability density function for the NB model can be written as **Equation 12** with dispersion parameter α.

$$P(y_i|\mu_{i,}\alpha) = \frac{\Gamma\left(\frac{1}{\alpha}+y_i\right)}{\Gamma(y_i+1)\Gamma\left(\frac{1}{\alpha}\right)}\left(\frac{1}{1+\alpha\mu_i}\right)^{\frac{1}{\alpha}}\left(\frac{\alpha\mu_i}{1+\alpha\mu_i}\right)^{y_i} \quad \text{(Eq.12)}$$

The likelihood function of the NB model is shown in **Equation 13**:

$$\ln(\mu|y,\alpha) = \prod_{i=1}^{n} exp\left\{y_i ln\left(\frac{\alpha\mu_i}{1+\alpha\mu_i}\right) - \frac{1}{\alpha}ln(1+\alpha\mu_i) + ln\Gamma\left(\frac{1}{\alpha}+y_i\right) - ln\Gamma(y_i+1) - ln\Gamma\left(\frac{1}{\alpha}\right)\right\} \quad \text{(Eq.13)}$$

The log likelihood function of the NB model is shown in **Equation 14**.

$$\ln(\beta|y,\alpha) = \sum_{i=1}^{n} y_i ln\left(\frac{\alpha exp(x_i\beta)}{1+\alpha exp(x_i\beta)}\right) - \frac{1}{\alpha}ln(1+\alpha exp(x_i\beta)) + ln\Gamma\left(\frac{1}{\alpha}+y_i\right) - ln\Gamma(y_i+1) - ln\Gamma\left(\frac{1}{\alpha}\right) \quad \text{(Eq.14)}$$

The goal is to maximize the likelihood function for obtaining the coefficient estimates for β and α. R-Language (R Core Team, 2018) was used to estimate the parameters of the NB models.

The models (SPFs) developed in this study represent the average conditions for segments in the Canada's rail network using segment-level data. For application in other areas, it might be necessary to calibrate the SPFs to better reflect local conditions and different study periods.

4.4. Empirical Bayes (EB) Approach

Reducing regression-to-the-mean (RTM) bias is important in accident prediction models to reduce the effects of natural fluctuations in accident data. This study applied the industry standard method known as the Empirical Bayes (EB) technique (Persaud and Lyon, 2016; Sun, 2002; Yu et al., 2014). **Equation 15** shows the formula for EB method. The formula uses information (predicted and observed values) obtained from the NB model outputs to estimate the long term mean values (i.e., the EB adjusted values) of the number of derailments for a certain segment:

$$\mathrm{E[y_i]} = \mathrm{w_i \cdot \mu_i} + (1 - \mathrm{w_i})\mathrm{y_i} \quad \text{(Eq.15)}$$

where:

$E[y]$ represents the EB adjusted number of derailments for segment i,

μ_i represents the predicted number of derailments for segment i,

y_i represents the observed number of derailments for segment i, and

w_i is the EB weight factor for segment i.

w_i is estimated in **Equation 16**:

$$w_i = \frac{1}{1+\alpha \times \sum_{t=1}^{Y} \mu_{ti}} \quad \text{(Eq.16)}$$

where:

α is the dispersion parameter,

μ_{ti} is the predicted number of derailments for segment i in year t, and

$Y = 10$ years in this study.

4.5. Model Calibration and Validation

A cross-validation method with data partitioning was used for model calibration and validation. The data was randomly partitioned into two sets. A larger portion of the observation set was used to calibrate the model parameters. The remaining portion was then used to validate the predictive power of the model. Different splits between the calibration portion and the validation portion (90:10, 80:20, 70:30, and 50:50) were tested to determine which split yielded the best performance for the derailment prediction models. As the results were similar for all four splits, a 70:30 split was selected as this approach provided a reasonable amount of data for the calibration and validation of the five models developed for the study.

Using 30% of the dataset, the models were further validated by predicting the number of derailments using coefficient estimates and comparing the results with observed values. This approach was effective for evaluating the predictive performance of the models as the validation dataset was not included in the original estimation of the coefficient process. Data partitioning also allowed for the use of goodness-of-fit tests to assess model performance.

4.6. Goodness-of-fit Tests

Candidate models were developed using different combination and permutations of independent variables. Models that produced intuitive signs for coefficient estimates and acceptable Cumulative Residual (CURE) plots were then shortlisted and evaluated by goodness-of-fit tests. Nine goodness-of-fit (GOF) tests were performed to compare and evaluate the relative performance of the shortlisted models. The GOF tests were: Akaike's Information Criterion (AIC), Bayesian Information Criterion (BIC), Overdispersion Parameter, Mean Squared Error (MSE), Mean Squared Prediction Error (MSPE), Freeman Tukey R-Squared (R2FT), Mean Prediction Bias (MPB), Mean Absolute Deviation (MAD) and CURE plots.

4.6.1. Akaike Information Criteria

Akaike Information Criteria (AIC) is a common metric used for model selection. It is used to select models that best explain the data with minimum free parameters. AIC penalizes models with a higher number of parameters (Burnham and Anderson, 2004; Young and Park, 2013). **Equation 17** shows the formula for estimating AIC:

$$\mathrm{AIC} = -2\log L + 2(p) \qquad \text{(Eq.17)}$$

L and p represent the model log likelihood and the number of parameters in the model.

4.6.2. Bayesian Information Criterion

Bayesian Information Criterion (BIC) is closely related to AIC, but also considers sample size (number of segments in this study). BIC is a panelized likelihood criterion (as is AIC). A lower BIC implies fewer independent variables and/or a better fit. **Equation 18** shows the formula for estimating BIC:

$$\mathrm{BIC} = p \times \log(n) - 2\log \mathrm{L} \qquad \text{(Eq.18)}$$

L and p are the same as **Equation 17** while *n* represent the sample size.

4.6.3. Dispersion Parameter

Dispersion parameter is generated by the NB model as part of the variance equation. Larger dispersion indicates larger variance (greater variability) and larger variance increases the standard errors of a model's estimates (Lyon C. et al, 2016). A model with a low dispersion parameter is preferred. **Equation 19** shows the formula for the variance of the NB distribution, and the re-arranged formula for estimating the dispersion parameter is shown in **Equation 20** (Lyon et al, 2016):

$$\mathrm{Var}\,\{m\} = \mathrm{E}\{m\} + f(k)E\{m\}^2 \qquad \text{(Eq.19)}$$

$$f(k) = \frac{\mathrm{Var}\,\{m\} - \mathrm{E}\{m\}}{E\{m\}^2} \qquad \text{(Eq.20)}$$

where:

$\mathrm{Var}\,\{m\}$ is the estimated variance of mean,

$\mathrm{E}\{m\}$ is the estimated mean, and

$f(k)$ is the estimated dispersion parameter function.

4.6.4. Mean Square Error and Root Mean Squared Error

Mean Square Error (MSE) measures how closely the fitted value represents the observed value: a smaller MSE indicates a fit that is closer to the actual value (Washington et al., 2005; Washington et al., 2007; Young and Park, 2013). It is applied to the calibration dataset. A model with a small MSE is preferred. **Equation 21** shows the formula for calculating MSE:

$$\mathrm{MSE} = \frac{\sum_{i=1}^{n}\left(\mu_i - y_i\right)^2}{n - p} \quad \text{(Eq.21)}$$

where:

Y_i is the observed number of derailments at segment i,

μ_i is the number of derailments predicted by the model,

n is the total sample size, and

p is the same as **Equation 17.**

The square root of MSE is the Root Mean Squared Error (RMSE) which is the standard deviation of the prediction errors. The RMSE measures the data spread around the regression. Values closer to zero indicate less error in model prediction.

4.6.5. Mean Square Prediction Error

Mean Square Prediction Error (MSPE) measures the errors associated with predicted values. Values closer to zero indicate less error in model prediction. It is applied to the validation dataset. **Equation 22** shows the formula for calculating MSPE (Hamidi et al. 2010):

$$\mathrm{MSPE} = \frac{\sum_{i=1}^{n}\left(\mu_i - y_i\right)^2}{n} \quad \text{(Eq.22)}$$

where:

Y_i , μ_i and n are the same as **Equation 21**.

The calculation method for MSPE is similar to MSE. These two GOF tests can be compared to provide more information on the predictive capability of a model. If the MSPE value is greater than MSE value, it is an indication of an overfitting model. The less difference between MSPE and MSE values implies better model fit.

4.6.6. Freeman Tukey R squared

Freeman Tukey R squared (R^2TF) is a goodness-of-fit test commonly used for count models mainly because of its variance-stabilizing transformations for binomial distribution (Freeman and Tukey, 1950). R^2TF is widely used to measure model fits for SPFs (Lu, 2014). It is applied to both the calibration and the validation datasets and measures how well a model fits the data. Larger R^2TF values imply better model fit. R^2TF is estimated using **Equation 23**:

$$R^2_{FT} = \frac{\sum_{i=1}^{n}(f_i - \bar{f}) - \sum_{i=1}^{n}\hat{e}_i^2}{\sum_{i=1}^{n}\left((f_i - \bar{f})\right)^2} \quad \text{(Eq.23)}$$

where:

$f_i = \sqrt{Y_i} + \sqrt{\mu_i + 1}$

$\bar{f} = \frac{\sum_{i=1}^{N}(\sqrt{Y_i} + \sqrt{\mu_i + 1})}{N}$, and

$\hat{e}_i = \sqrt{Y_i} + \sqrt{Y_i + 1} - \sqrt{4 \times \mu_i + 1}$

Y_i, μ_i, and n are the same as **Equation 21**.

4.6.7. Mean Prediction Bias

Mean Prediction Bias (MPB) measures the magnitude and direction of the average model bias (Hamidi et al., 2010; Washington et al., 2005). It is applied to the validation dataset to calculate the differences between the predicted and observed values. Smaller MPB values indicate better prediction. MPB values can be positive (indication of over-estimation) or negative (indication of under-estimation). **Equation 24** shows the formula for calculating MPB:

$$MPB = \frac{\sum_{i=1}^{n}(\hat{Y}_i - Y_i)}{n} \quad \text{(Eq.24)}$$

Y_i, μ_i, and n are the same as **Equation 21.**

4.6.8. Mean Absolute Deviation

MAD measures the average magnitude of variability of prediction in absolute values (Lyon C. et al, 2016). It is applied to the validation data. Smaller MAD values indicate better prediction. **Equation 25** shows the formula for calculating MAD:

$$MAD = \frac{\sum_{i=1}^{n}|\hat{Y}_i - Y_i|}{n} \quad \text{(Eq.25)}$$

Y_i, μ_i, and n are the same as **Equation 21.**

4.6.9. Cumulative Residual Plots

Residuals are the differences between the observed and predicted values and can be used to measure the predictive power of a model. A CURE plot is a graphical representation of the model fit. It can be useful for understanding the performance of a model and for identifying where biased estimates occur.

Figure 36 provides an example of a CURE plot from this study. Residuals were plotted against exposure (number of derailments against daily train volume in this example). The zero line on the graph corresponds to the region where the estimates are unbiased (the observed values) and the black lines above and below zero are the residuals. The area above the zero line represents underestimations and the area below the zero line represents overestimations. The green and red lines indicate two standard deviations above and below zero. Residuals that fall within the green and red lines fall within the 95% confidence level and indicate a good fitting model. Generally, a CURE plot that shows cumulative residuals that fall within the two standard deviations and converge at 0 indicate to a good fitting model (Lu et al., 2014). **Figure 36** shows a good fit.

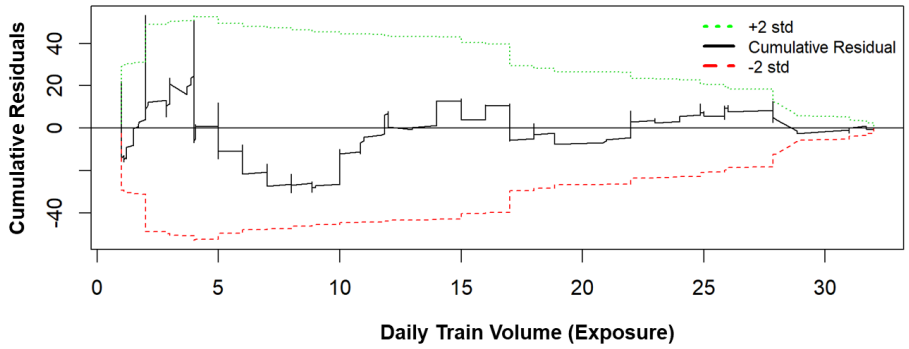


Figure 36: Sample Cumulative Residual (CURE) Plot

Figure 37 shows an example of a CURE plot for a poor fitting model. The increasing line that runs above zero indicates that this model was consistently underestimating the number of derailments. Selecting a different functional form, and/or adding or removing variables in the model might improve the fitting of the estimates.

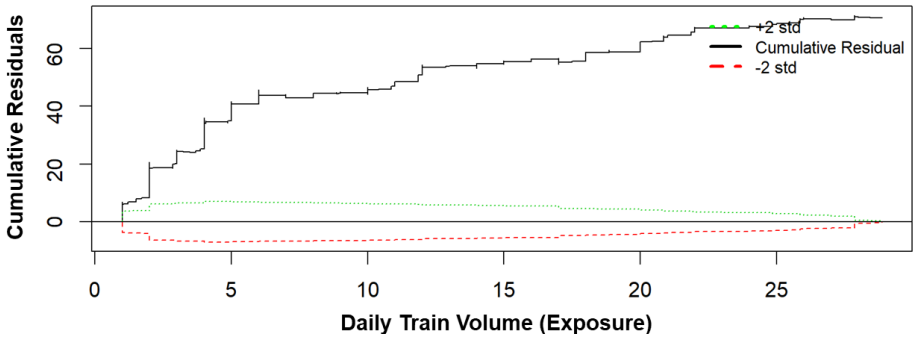


Figure 37: Sample CURE Plot Showing a Poor Fitting Model

4.7. Variable Selection

The first step of the model development was variable selection which involved a correlation analysis of response and independent variables. **Table 16** is the correlation matrix developed to pre-screen the statistical relationships between these variables. Five independent variables of interest were considered: daily train volume (VL_Count), maximum train speed (VL_TrnSpd), segment length (Seg_Length), average daily train volume (Avg_Train) and number of stations on a segment (Stn_Count).

Table 16: Correlation Matrix of Model Variables

	DerailCount	Seg_Length	VL_Count	VL_TrainSpeed	Avg_Train	Stn_Count
DerailCount	1.00	0.31	0.19	0.20	0.18	0.19
Seg_Length	0.31	1.00	-0.01	0.07	-0.02	0.29
VL_Count	0.19	-0.01	1.00	0.64	0.98	0.31
VL_TrainSpeed	0.20	0.07	0.64	1.00	0.63	0.20
Avg_Train	0.18	-0.02	0.98	0.63	1.00	0.31
Stn_Count	0.19	0.29	0.31	0.20	0.31	1.00

Segment length had a low linear association (0.31) with the number of derailments and is to contribute to improving the fitting performance of the model. The other independent variables had a low correlation with the number of derailments (less than 0.20). Despite the low correlations with the number of derailments, all the independent variables are carried forward as their association with derailments might improve the overall predictive capabilities of different model forms.

The issue of multilinearity was further examined. Most of the variables are not highly correlated with each other (>0.5), but maximum train speed and daily train volume showed a moderate correlation with each other (0.64). These two variables were carried forward for modelling as they are expected to have direct effects on derailment potential. Since maximum and average daily train volumes are both exposure variables (highly correlated), these were not included in the same model forms simultaneously. Variance inflation factors (VIF) were calculated to assess the magnitude of multicollinearity for the best-fitting models (provided in **Appendix C**). In this study, all independent variables in the best-fitting models had VIF values less than 5, an indication that multicollinearity is not a concern.

Statistical significance and/or predictive ability can be used to evaluate the variables in a statistical model. Much scientific research uses statistical significance based on the p-values of the independent variables. However, model results should also be interpreted in the context of the study and not solely on

p-values (Wasserstein and Lazar, 2016). This issue leads to some controversies on statistical null hypothesis significance testing, particularly in road safety research.

For example, variables that may not appear to be statistically significant may still have a strong effect in predicting a dependent variable (Hauer, 2004). Hauer suggested that more attention should be given to the estimates of effect magnitude and standard errors as some variables may have predictive capability although their p-values are not within the 95% confidence level.

This thesis focuses on predicting the output value (μ_i) using a set of independent variables (X) rather than developing an explanatory model. Explanatory modelling is interested in identifying a set of statistically significant variables to test a theoretical hypothesis. Predictive modelling constructs the function from the dataset by capturing the association between μ_i and X. In predictive modelling in the context of transportation safety, the predictive power of the model is more important than the statistical significance of the independent variables. This study considered both statistical significance and the predictive power of the models developed.

During the model development process, independent variables were introduced to check the performance of the model with different combinations of functional forms. This was an iterative process that continued until a best performing model was obtained. The overall predictive performance of the model was then tested using the GOF tests described in **Chapter 4.6**. The selection of the best-fitting model was based on the results of GOF tests and CURE plots.

4.8. Chapter Summary

This chapter described the input data and modelling scheme for the study. Given the great variety of geographic and track characteristics in the study data, five derailment prediction models were developed to capture the variation in the rail network. The five models were: Canada, Eastern Canada, Western Canada, Canadian National Railway, and Canadian Pacific Railway.

Negative binomial distribution was the first modelling approach to predicting the number of derailments in this study. The goal was to maximize the likelihood function for obtaining the coefficient estimates in the prediction models. To account for the regression-to-the-mean (RTM) bias, Empirical Bayes (EB) method was applied to calculate the expected safety performance of each segment in the rail network. In terms of variable selection, forward selection was used to account for the inclusion of model variables. The modelling approach applied a cross-validation process in which the database was randomly split into two separate sets calibration and validation. A set of candidate models were developed with different

combination of independent variables. The models that have intuitive signs for coefficient estimates and acceptable CURE plots were then shortlisted.

Nine goodness-of-fit (GOF) tests evaluated the relative performance of the shortlisted models: Akaike's Information Criterion (AIC), Bayesian Information Criterion (BIC), Overdispersion Parameter, Mean Squared Error (MSE), Mean Squared Prediction Error (MSPE), Freeman Tukey R-Squared (R2FT), Mean Prediction Bias (MPB), Mean Absolute Deviation (MAD) and CURE plots. The CURE plots provided a visual illustration of the differences between the predicted and observed values. **Table 17** provides a summary of the GOF measures and their preferred values.

Table 17: Summary of GOF Measures and Preferred Values

GOF Measure	**Preferred Value**
Akaike's Information Criterion	Smaller Values
Bayesian Information Criterion	Smaller Values
Dispersion Parameter	Smaller Values
Mean Squared Error/ Root Mean Squared Error	Smaller Values
Mean Squared Prediction Error	Smaller Values
Freeman Turkey R-Squared	Larger Values
Mean Prediction Bias	Smaller Values
Mean Absolute Deviance	Smaller Values

CHAPTER 5. DEVELOPMENT AND ANALYSIS OF PREDICTION MODELS

This chapter discusses the development, analysis and results of the five prediction models using negative binomial modelling method. The five models were: Canada, Eastern Canada, Western Canada, Canadian National Railway (CN) and Canadian Pacific Railway (CP). As train traffic is an exposure variable, only segments with train volumes greater than zero were used for the modelling.

The methodology described in **CHAPTER 4** was used. Each model followed a four-step process: identification of outliers, development and comparison of shortlisted models, selection of the best-fitting model, and discussion of results. **Figure 38** summarizes the process.

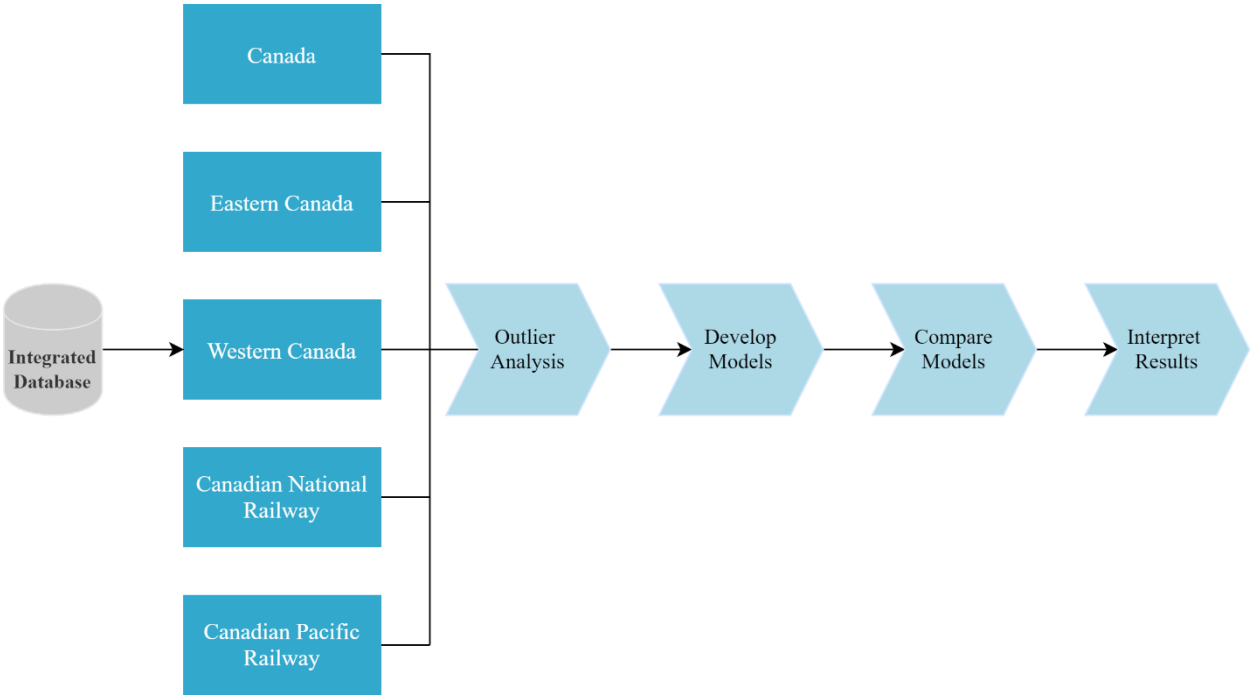


Figure 38: Organization of Model Analysis Discussion

Before model development, an outlier analysis was conducted to identify values greater than the upper bound and less than the lower bound of each independent variable. The model calibration process was based on a forward selection of variables. This approach begins with a model involving only the exposure variable and gradually expanding it by adding independent variables until the introduction of another variable did not improve the model performance. The candidate models with good predictive capability were shortlisted on the basis of an intuitive interpretation of the signs of the coefficient estimates

and acceptable CURE plots. The candidate models on the shortlist were compared, and the best-fitting model was selected.

Appendix A provides the model forms of the candidate models. **Appendix B** presents the function forms, model statistics and CURE plots for the shortlisted models. **Appendix C** presents the variance inflation factors (VIF) for the best-fitting models to examine any issue of multicollinearity.

5.1. Derailment Prediction Model for Canada

The Canada model consists of 3,144 track segments. **Figure 39** shows all track segments in Canada.

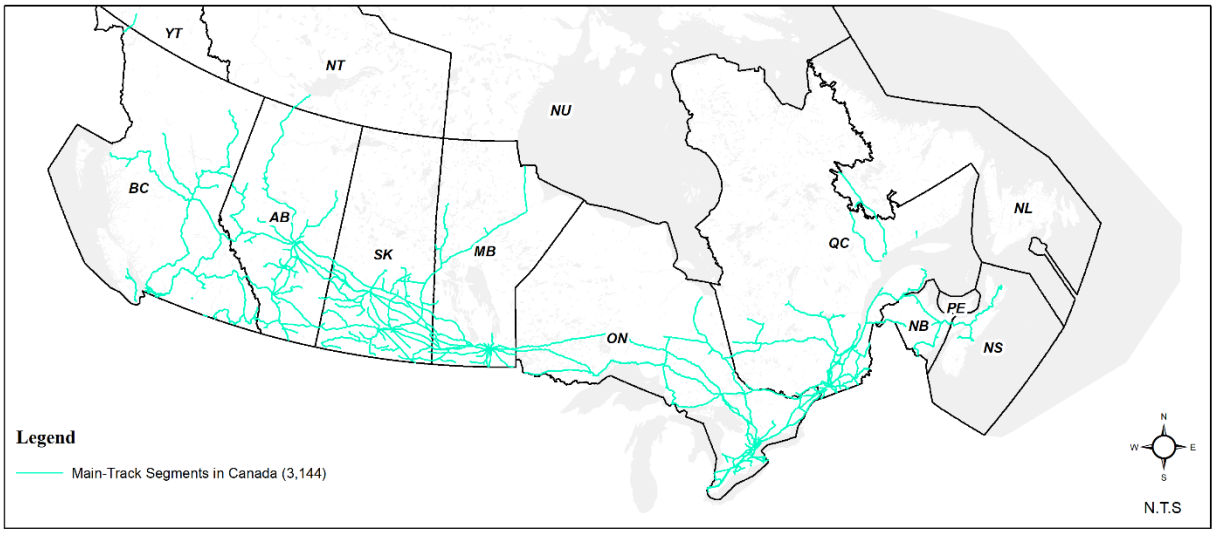


Figure 39: Track Segments in Canada

5.1.1. Outlier Analysis for Canada Model

Table 18 shows the summary statistics for the outlier analysis conducted for the Canada dataset.

Table 18: Summary Statistics for Key Independent Variables – Canada Model

	Seg_length	Train Speed	Max Train Vol	Avg Train Vol	Station Count
Minimum	1	10	1	1	0
Q1	5	30	2	2	0
Median	12	40	5	5	0
Q3	19	50	13	12	1
Maximum	43	80	62	62	2
Mean	13	41	9	8	1
Range	42	70	61	62	2
Interquartile Range	14	20	11	10	1

Prior to model development, values that are greater than the upper bound and less than the lower bound of each independent variable were identified as outliers. After outliers were removed, the Canada dataset has 2,553 observations (or 81% of the original data). **Figure 40** is a map showing the segments being included (dark blue lines) in the model.

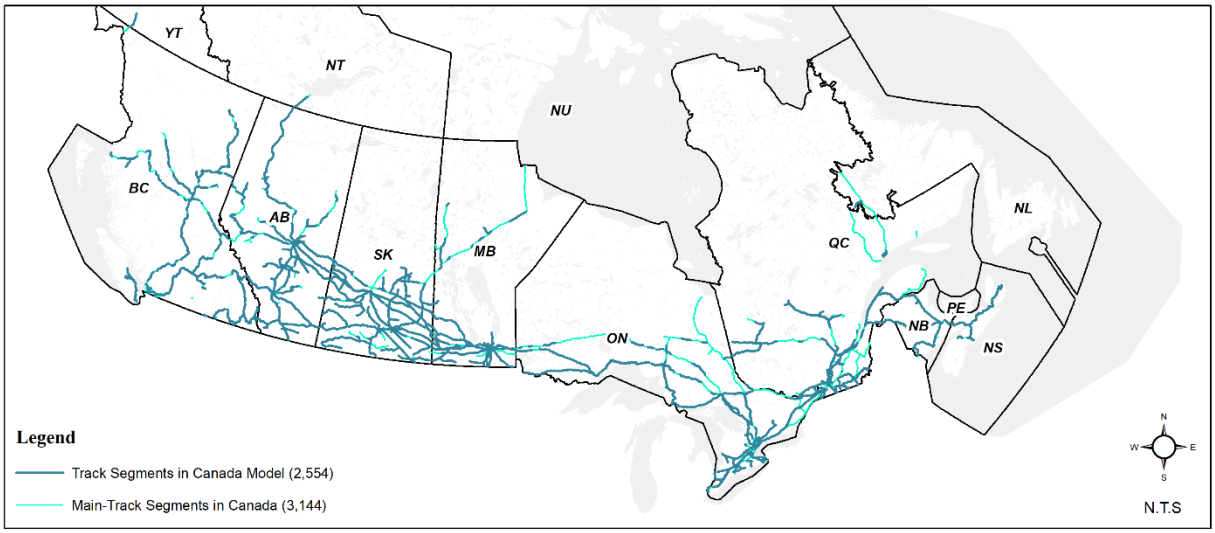


Figure 40: Track Segments in Canada Model

5.1.2. Model Comparison of Candidate Models for Canada

In total, 70 candidate models were developed for predicting the number of derailments in Canada. **Table 19** compares the results of the GOF tests applied to the calibration and validation data for the shortlisted Canada models. Model ID#7 (shown in bold) was selected as the best-fitting model. **Chapter 5.1.3** continues the discussion of the results.

Table 19: Summary of Goodness-of-fit Test Results for Shortlisted Canada Models

Shortlisted Model	AIC	BIC	Dispersion	Calibration Data (70%)		Validation Data (30%)			
				MSE	R2FT	MSPE	MPB	MAD	R2FT
4	2214.30	2247.34	1.06	0.40	8.7%	0.43	-0.02	0.42	3.7%
7	**2191.04**	**2224.07**	**1.04**	**0.41**	**8.8%**	**0.43**	**-0.01**	**0.41**	**4.0%**
8	2192.24	2225.27	0.98	0.41	7.8%	0.43	-0.01	0.42	3.9%
9	2212.13	2245.17	1.02	0.41	8.1%	0.43	-0.01	0.42	3.9%
65	2182.12	2220.66	1.01	0.41	8.3%	0.42	-0.02	0.41	5.5%
67	2210.93	2249.46	1.06	0.40	9.2%	0.53	-0.01	0.42	3.6%
68	2210.97	2255.01	1.07	0.40	9.2%	0.50	-0.01	0.42	4.2%
69	2210.93	2249.46	1.06	0.40	9.4%	0.43	-0.02	0.42	3.6%
70	2213.30	2251.84	1.06	0.38	14.6%	0.41	-0.03	0.39	4.5%

5.1.3. Discussion of Results of Model Selected for Canada

This section discusses the results of the CURE plot for the model selected for Canada, presents the functional form of the selected model, and considers the statistical relationships between the independent variables and the number of derailments predicted.

Figure 41 shows the CURE plot for the selected model for Canada. When train traffic was less than 42 trains per day, the model's prediction performance was good with predictions within ±2 standard deviations of the zero line. (The zero line on the graph corresponds to the region where estimates are unbiased (the observed values)). When train traffic was higher than 42 trains per day, the model underestimated the derailment potential. The segments with more than 42 trains per day appeared to be the shorter segments in the dataset.

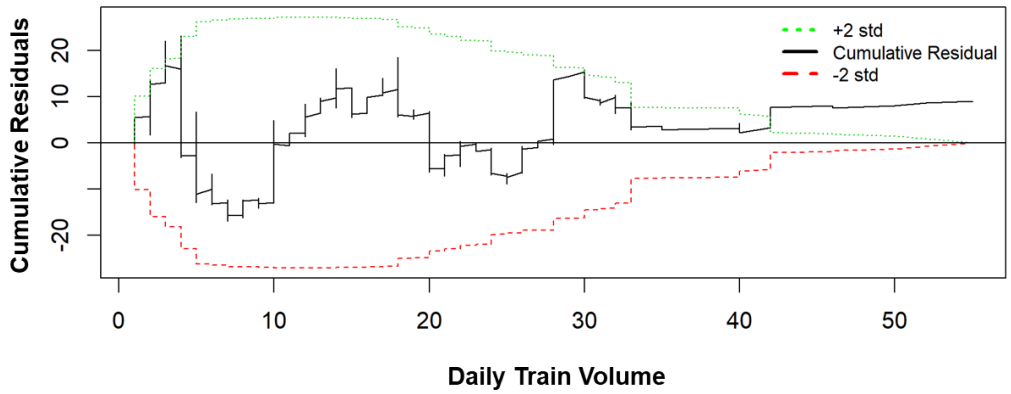


Figure 41: Cumulative Plot of Negative Binomial Model for Canada

Equation 26 shows the functional form of the selected model for Canada. This equation was used to predict the number of derailments in Canada for a 10-year period. The number of years is represented by *N* in the model form.

$$\mu_i = N \times exp(-9.0773) \times exp\left[\begin{matrix}(0.2573 \times \text{logVL\_Count}) + (1.0891 \times \text{logVL\_TrnSpd}) \\ +(-0.0170 \times \text{Stn\_Count}) + (0.0576 \times \text{Seg\_Length})\end{matrix}\right] \quad \text{(Eq. 26)}$$

Table 20 presents the coefficient estimates and model statistics.

Table 20: Negative Binomial Model Results for Canada

Variable	Estimate	Std. Error	z value	Pr(>|z|)	Model Statistics
(Intercept) ***	-9.0773	0.6944	-13.0721	0.0000	
log_VL_Count***	0.2573	0.0722	3.5648	0.0004	Dispersion Parameter: 1.0362
log_VL_TrnSpd***	1.0891	0.2046	5.3233	0.0000	Standard Error.: 0.197
Stn_Count	-0.0170	0.0725	-0.2337	0.8152	Log Likelihood: -2179.036
Seg_Length***	0.0576	0.0055	10.4306	0.0000	Observations: 2,553

****significance level of less than 0.0001*

Table 20 shows that, as expected, an increase in exposure (daily train traffic) showed a positive association with the number of derailments. Higher rail traffic increases track damage which may then increase the risk of derailments.

Higher train speed also showed a positive association with the number of derailments and also increases track damage. High speed can damage the foundation layer of the track due to high vibration and track deformation (Kish and Clark, 2009; Sayeed and Shahin, 2016) and may contribute to an increased risk of derailment. Several studies have noted that train speed has often been a major contributing factor in derailment and in the number of fatalities resulting from a derailment (Anderson and Barkan, 2005; Bagheri et al., 2011; Bibel et al., 2016; Liu et al., 2013; Wang and Li, 2012). Britton et al. (2017) reported that about 60% of all derailments could have been affected by train speed.

The number of stations on a track segment showed a negative association with the number of derailments. Both frequent and rare stops along a segment could impact train speed. In general, a segment with fewer stops provides more opportunities for a train to accelerate and maintain its travel speed, possibly increasing the risk of derailment.

The Canada model produced statistically significant (95% confidence level) coefficient estimates for all the independent variables except station count. Although its p-value was not within the 95% confidence level, the inclusion of the station count variable helped to improve the model performance.

5.2. Derailment Prediction Model for Eastern Canada

The Eastern Canada model used 1,311 segments from New Brunswick, Newfoundland and Labrador, Nova Scotia, Ontario and Quebec. **Figure 42** shows the segments in Eastern Canada.

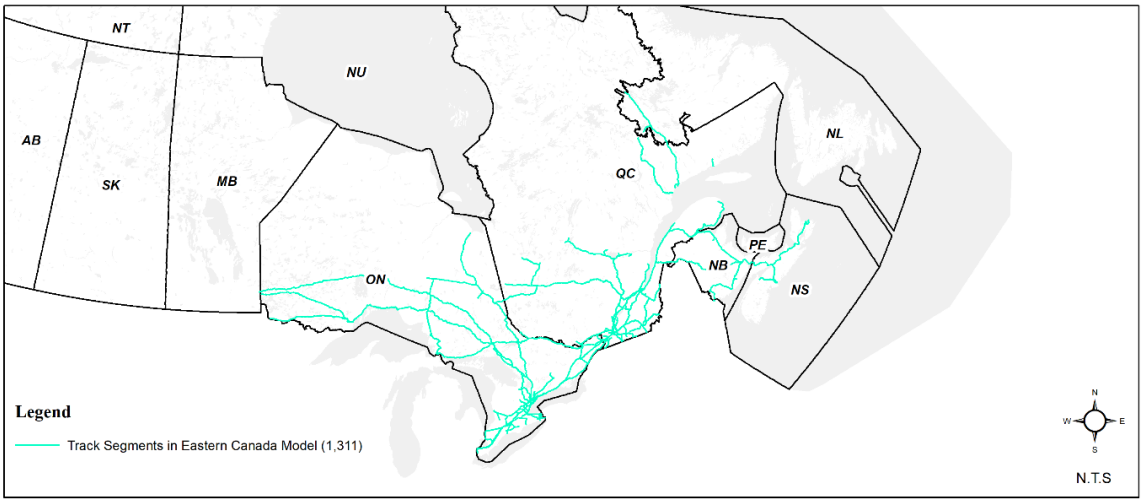


Figure 42: Track Segments in Eastern Canada Model

5.2.1. Outlier Analysis for Eastern Canada Model

Table 21 shows the summary statistics for the outlier analysis conducted for the Eastern Canada dataset.

Table 21: Summary Statistics for Key Independent Variables – Eastern Canada

	Seg_length	Train Speed	Max Train Vol	Avg Train Vol	Station Count
Minimum	1	10	1	1	0
Q1	4	35	4	3	0
Median	12	45	8	6	0
Q3	19	60	14	12	1
Maximum	43	80	36	36	2
Mean	13	45	9	8	1
Range	42	70	35	36	2
Interquartile Range	15	25	10	9	1

After outliers were removed, the dataset comprised 925 segments (71% of the original Eastern Canada dataset), as shown in **Figure 43**.

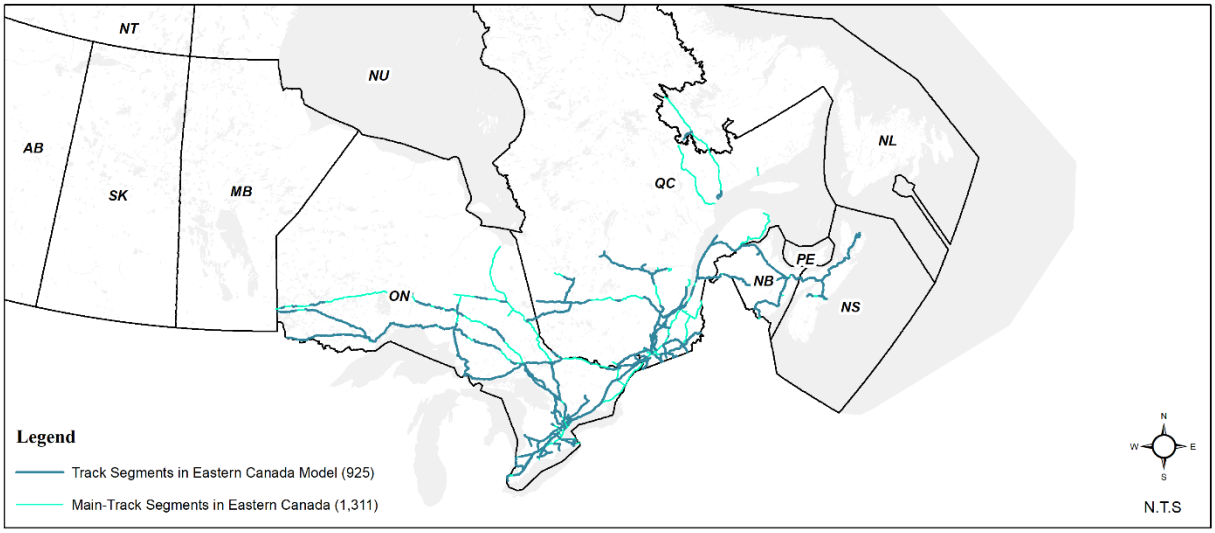


Figure 43: Track Segments in Eastern Canada Model

5.2.2. Model Comparison for Eastern Canada

In total, 72 candidate models were developed for predicting the number of derailments in Eastern Canada. **Table 22** compares the results of the GOF tests applied to the calibration and validation data for the shortlisted Eastern Canada models. Model ID#69 (shown in bold) was selected as the best-fitting model. **Chapter 5.2.3** continues the discussion of the results.

Table 22: Summary of Goodness-of-fit Tests for Shortlisted Eastern Canada Models

Shortlisted Model	AIC	BIC	Dispersion	Calibration Data (70%) MSE	R2FT	Validation Data (30%) MSPE	MPB	MAD	R2FT
4	776.65	803.49	0.66	0.38	7%	0.27	0.03	0.34	-3%
8	776.09	802.93	0.60	0.39	4%	0.26	0.04	0.34	-2%
9	782.37	809.20	0.60	0.39	5%	0.26	0.04	0.34	0%
69	**775.43**	**806.74**	**0.67**	**0.38**	**7%**	**0.27**	**0.03**	**0.34**	**-2%**
71	779.55	815.33	0.67	0.38	7%	0.27	0.02	0.33	-2%
72	778.65	809.96	0.66	0.38	7%	0.27	0.03	0.34	-3%

5.2.3. Model Results for Eastern Canada

Figure 44 shows the CURE plot for the selected model for Eastern Canada. The cumulative residual line almost converges at zero, indicating that the predicted numbers of derailments were very close to the observed number of derailments. The model predictions were within ±2 standard deviations over almost the entire range of train traffic. The exception was segments with between 15 and 20 trains per day where the number of derailments was underestimated. The majority of these segments (63%) had no derailments which may have contributed to the underestimation.

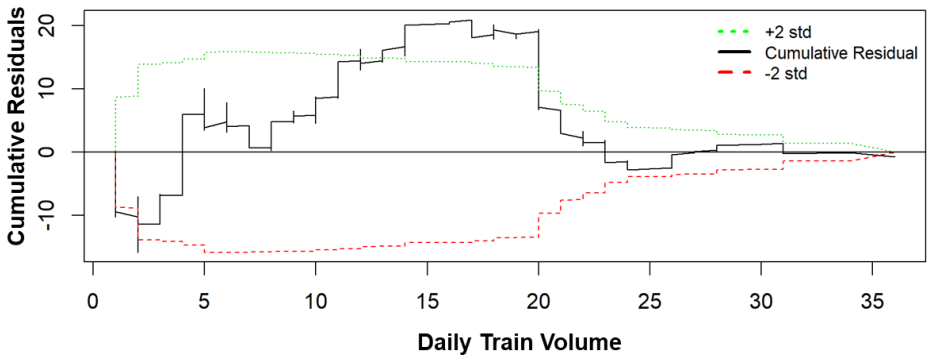


Figure 44: Cumulative Plot of Negative Binomial Model for Eastern Canada

Equation 27 shows the functional form of the selected model for Eastern Canada. This equation was used to predict the number of derailments in Eastern Canada for a 10-year period. The number of years is represented by *N* in the model form.

$$\mu_i = N \times exp(-5.5095) \times exp\left[\begin{array}{c}(0.0769 \times \text{VL\_Count}) + (0.0183 \times \text{VL\_TrnSpd}) + \\ (0.2365 \times \text{Stn\_Count}) + (0.0537 \times \text{SEg\_Length}) \\ +(0.0537 \times (\text{VL\_Count} \times \text{Seg\_Length}))\end{array}\right] \quad \text{(Eq.27)}$$

Table 23 presents the coefficient estimates and model statistics.

Table 23: Negative Binomial Model Results for Eastern Canada

Variable	Estimate	Std. Error	z value	Pr(>|z|)	Model Statistics
(Intercept) ***	-5.5095	0.4346	-12.6764	0.0000	
log_VL_Count	0.0769	0.1307	0.5880	0.5565	Dispersion Parameter: 0.6687
VL_TrnSpd	0.0183	0.0094	1.9541	0.0507	Standard Error.: 0.192
Stn_Count	0.2365	0.4570	0.5175	0.6048	Log Likelihood: -764.117
Seg_Length ***	0.0537	0.0091	5.8946	0.0000	Observations: 925
TrnSpd × StnCount	-0.0068	0.0087	-0.7722	0.4400	

****significance level of less than 0.0001*

Like the Canada model, daily train volume and maximum train speed showed positive associations with the number of derailments. Unlike the Canada model, the number of stations on a track segment also shows a positive association with derailments. Both frequent and rare stops along a segment could impact train speed, acceleration and deceleration. Frequent or sudden changes in train speed while approaching or leaving a station could contribute to a derailment (Snow et al., 2013).

The interaction term between train speed and station count changed the direction of the effect of train speed from negative to positive allowing for a more intuitively acceptable interpretation of the sign of the model parameters. It also appears train speed and station count have a negative association (a smaller number of stations along a segment contributes to higher train speed) as might be expected. The introduction of the interaction term enhanced the Eastern Canada model's performance.

In terms of statistical significance, segment length had a p-value within the 95% confidence level. The p-value of the other variables were less than the 0.05 threshold for the 95% confidence level, but were retained as they improved the overall predictive performance of the model.

5.3. Derailment Prediction Model for Western Canada

The Western Canada model used 1,833 derailments from Alberta, British Columbia, Manitoba, and Saskatchewan. **Figure 45** shows the segments in Western Canada.

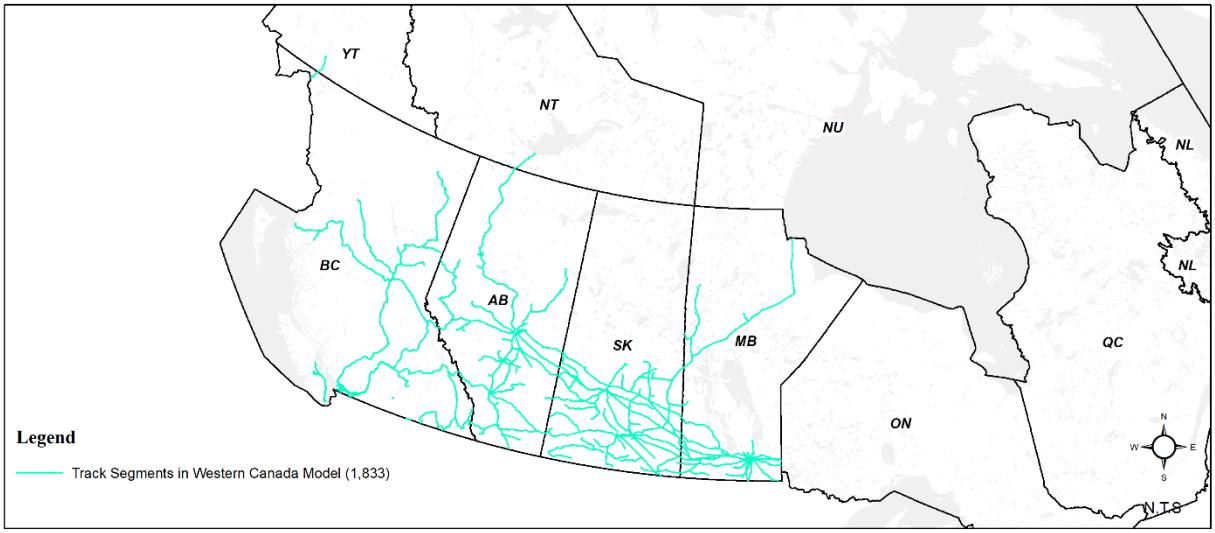


Figure 45: Track Segments in Western Canada

5.3.1. Outlier Analysis for Western Canada Model

Table 24 shows the summary statistics for the outlier analysis conducted for the Western Canada dataset.

Table 24: Summary Statistics for Key Independent Variables – Western Canada

	Seg_length	Train Speed	Max Train Vol	Avg Train Vol	Station Count
Minimum	1	10	1	1	0
Q1	5	25	2	2	0
Median	12	40	4	4	0
Q3	19	40	9	8	1
Maximum	42	65	26	26	2
Mean	13	36	6	6	0
Range	41	55	25	26	2
Interquartile Range	14	15	7	6	1

After outliers were removed, the dataset comprised 1,447 segments (79% of the original dataset for Western Canada), as shown in **Figure 46**.

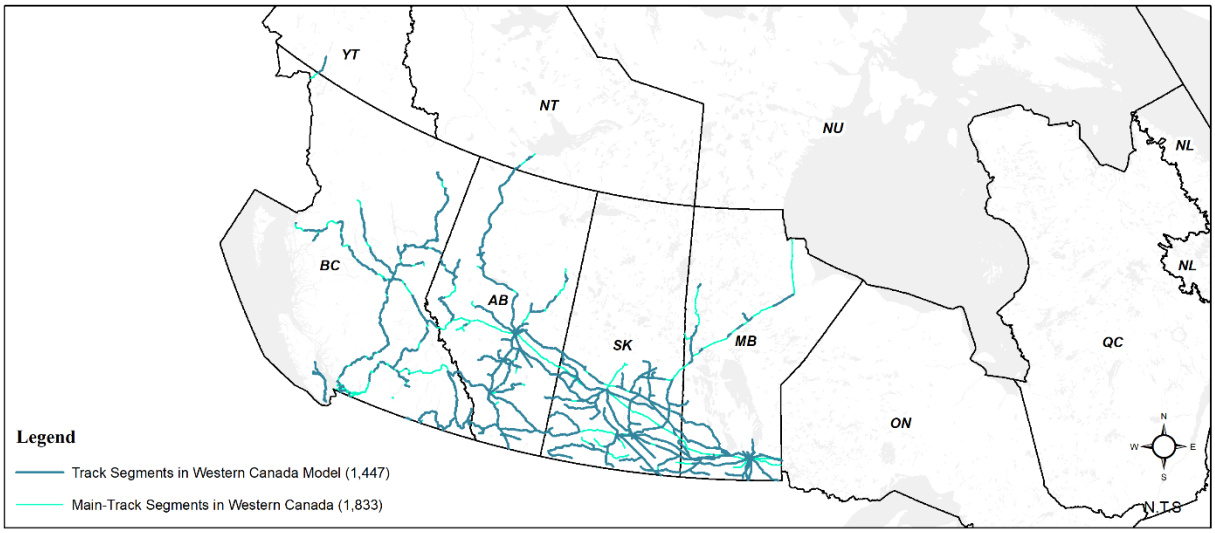


Figure 46: Track Segments in Western Canada Model

5.3.2. Model Comparison for Western Canada

In total, 72 candidate models were developed for predicting the number of derailments in Western Canada. **Table 25** compares the results of the GOF tests applied to the calibration and validation data for the shortlisted Western Canada models. Model ID#4 (shown in bold) was selected as the best-fitting model. **Chapter 5.3.3** continues the discussion of the results.

Table 25: Summary of Goodness-of-fit Tests for Shortlisted Western Canada Models

Shortlisted Model	AIC	BIC	Dispersion	Calibration Data (70%) MSE	Calibration Data (70%) R2FT	Validation Data (30%) MSPE	Validation Data (30%) MPB	Validation Data (30%) MAD	Validation Data (30%) R2FT
4	**1198.05**	**1227.64**	**1.01**	0.37	6%	0.49	-0.03	0.39	12%
9	1193.84	1223.42	0.98	0.39	4%	0.49	-0.03	0.39	13%
30	1199.73	1234.25	1.02	0.37	6%	0.49	-0.03	0.39	12%
67	1197.85	1227.44	1.02	0.37	6%	0.49	-0.03	0.39	12%
70	1199.16	1223.68	1.18	0.35	9%	0.50	-0.03	0.40	9%

5.3.3. Model Results for Western Canada

Figure 47 shows the CURE plot for the best-fitting model. The predictions are within ±2 standard deviations over the entire range of daily train volume except for segments with more than 25 trains per day. Majority (85%) of the segments with train volumes of more than 25 trains per day had no observed derailments which led to underestimation in the model.

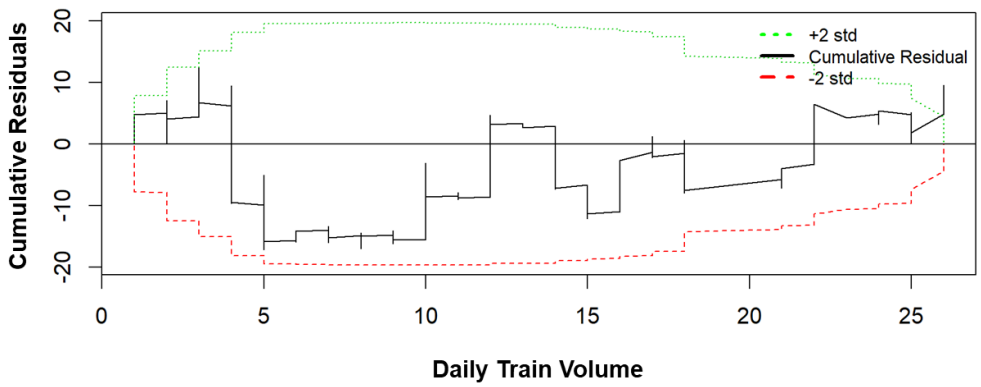


Figure 47: Cumulative Plot of Negative Binomial Model for Western Canada

Equation 28 shows the functional form of the selected model for Western Canada. This equation predicts the number of derailments for a 10-year period. The number of years is represented by "N" in the model form.

$$\mu_i = N \times exp(-7.0619) \times exp\begin{bmatrix}(0.0200 \times \text{logVL\_Count}) + (0.0622 \times \text{VL\_TrnSpd}) + \\ (0.2256 \times \text{Stn\_Count}) + (0.0518 \times \text{Seg\_Length})\end{bmatrix} \quad \text{(Eq.28)}$$

Table 26 presents the coefficient estimates and model statistics.

Table 26: Negative Binomial Model Results for Western Canada

Variable	Estimate	Std. Error	z value	Pr(>|z|)	Model Statistics
(Intercept) ***	-7.0619	0.3379	-20.9020	5.1E-97	
log_VL_Count	0.0200	0.1096	0.1827	0.8550	Dispersion Parameter: 1.0134
VL_TrnSpd***	0.0622	0.0090	6.9008	0.0000	Standard Error: 0.253
Stn_Count*	0.2256	0.1005	2.2457	0.0247	Log Likelihood: -1186.048
Seg_Length***	0.0518	0.0079	6.5536	0.0000	Observations: 1,447

**significance level of less than 0.01*

****significance level of less than 0.0001*

All the independent variables (daily train traffic, maximum train speed, station count, and segment length) showed a positive association with the number of derailments. This finding indicates that increases in these variables have the effect of increasing the risk of derailment.

In terms of statistical significance, maximum train speed and segment length had p-values that are within 95% confidence level. The p-value of other variables was less than the 0.05 threshold for the 95% confidence level, but were retained as they improved the overall predictive performance of the model.

5.4. Derailment Prediction Model for the Canadian National Railway

The Canadian National (CN) Railway model used the 1,459 segments owned by the CN Railway. **Figure 48** shows the segments.

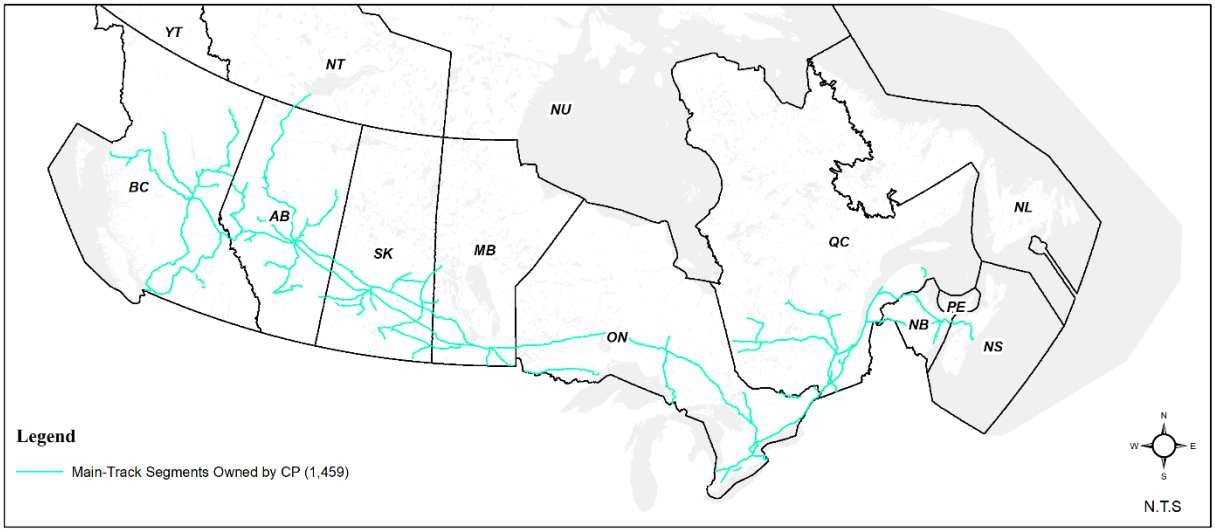


Figure 48: Track Segments Owned by CN

5.4.1. Outlier Analysis for Canadian National Railway Model

Table 27 shows the summary statistics for the outlier analysis conducted for the CN dataset.

Table 27: Summary Statistics for Key Independent Variables in CN Model

	Seg_length	Train Speed	Max Train Vol	Avg Train Vol	Station Count
Minimum	1	10	1	1	0
Q1	7	30	3	3	0
Median	14	40	5	5	1
Q3	21	50	14	13	2
Maximum	43	80	40	40	5
Mean	15	43	10	9	1
Range	42	70	39	40	5
Interquartile Range	14	20	11	10	2

After outliers were removed, the database comprised 1,332 observations (91% of the original CN dataset), as shown in **Figure 49**.

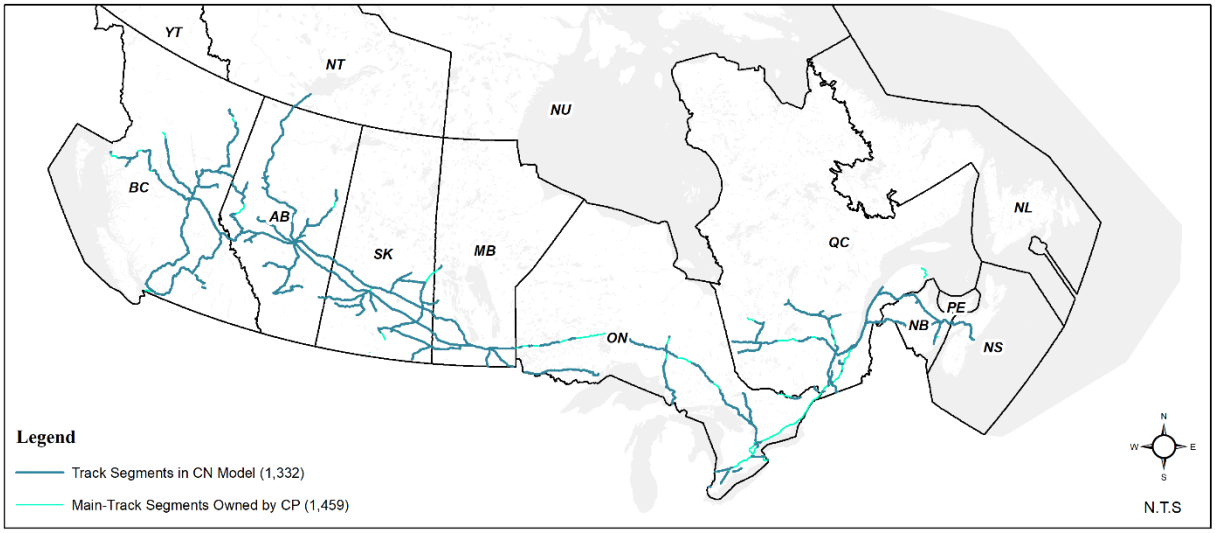


Figure 49: Track Segments in CN Model

5.4.2. Model Comparison for Canadian National Railway

In total, 72 candidate models were for predicting the number of derailments on CN-owned tracks. **Table 28** compares the results of the GOF tests applied to the calibration and validation data for the shortlisted CN models. Model ID#61 (shown in bold) was selected as the best-fitting model. **Chapter 5.4.3** continues the discussion of the results.

Table 28: Summary of Goodness-of-fit Tests for Shortlisted CN Models

Shortlisted Model	AIC	BIC	Dispersion	Calibration Data (70%) MSE	Calibration Data (70%) R2FT	Validation Data (30%) MSPE	Validation Data (30%) MPB	Validation Data (30%) MAD	Validation Data (30%) R2FT
61	**1264.93**	**1299.10**	**1.78**	**0.38**	**8%**	**0.40**	**0.01**	**0.41**	**8%**
62	1286.12	1315.41	1.78	0.38	6%	0.40	0.01	0.41	8%
63	1287.28	1316.57	1.67	0.37	6%	0.40	0.01	0.41	8%
68	1261.38	1300.43	1.87	0.37	8%	0.40	0.01	0.42	4%
69	1264.73	1298.90	1.69	0.38	7%	0.39	0.01	0.41	7%

5.4.3. Model Results for Canadian National Railway

Figure 50 shows the CURE plot for the best-fitting model. The model predictions were within ±2 standard deviations over the entire range of daily train volume. The cumulative residual line almost converges at zero, indicating that the predicted numbers of derailments were very close to the observed number of derailments.

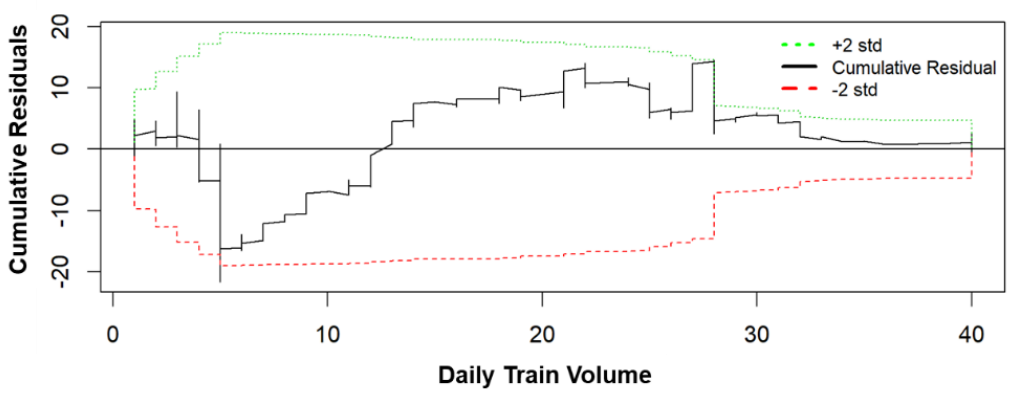


Figure 50: Cumulative Plot of Negative Binomial Model for CN

Equation 29 shows the functional form of the selected model for CN. This equation was used to predict the number of derailments on the CN network for a 10-year period. The number of years is represented by *N* in the model form.

$$\mu_i = N \times exp(-5.4929) \times exp\begin{bmatrix} (0.0774 \times \text{logVL\_Count}) + (0.0181 \times \text{VL\_TrnSpd}) \\ +(0.3403 \times \text{Stn\_Count}) + (0.0370 \times \text{logSeg\_Length}) \\ +(-0.0082 \times (\text{VL\_Count} \times \text{Stn\_Count})) \end{bmatrix} \quad \text{(Eq.29)}$$

Table 29 presents the coefficient estimates and model statistics.

Table 29: Negative Binomial Model Results for CN

Variable	Estimate	Std. Error	z value	Pr(>|z|)	Model Statistics
(Intercept) ***	-5.4929	0.2599	-21.1326	0.0000	
log_VL_Count	0.0774	0.1066	0.7258	0.4680	Dispersion Parameter: 1.7835
VL_TrnSpd***	0.0181	0.0048	3.7331	0.0002	Standard Error: 0.6020
Stn_Count***	0.3403	0.0794	4.2843	0.0000	Log Likelihood: -1250.93
Seg_Length***	0.0370	0.0074	5.0312	0.0000	Observations: 1,332
VLCount × StnCount*	-0.0082	0.0038	-2.1299	0.0332	

**significance level of less than 0.01*
****significance level of less than 0.0001*

The coefficient estimates showed statistical relationships similar to those of the previous model. All the independent variables showed a positive association with derailments. The introduction of the interaction term between daily train volume and station count changed the direction of effect of daily train traffic from negative to positive and enhanced the CN model's performance.

In terms of statistical significance, maximum train speed, station count and segment length all had p-values within the 95% confidence level. The p-value for daily train traffic was less than the threshold of 0.05 for the 95% confidence level, but daily train traffic was retained as it acted as an exposure term and its inclusion improved the predictive power of the model.

5.5. Derailment Prediction Model for the Canadian Pacific Railway

The Canadian Pacific (CP) Railway model used the 982 segments owned by the CP Railway. **Figure 51** shows the segments.

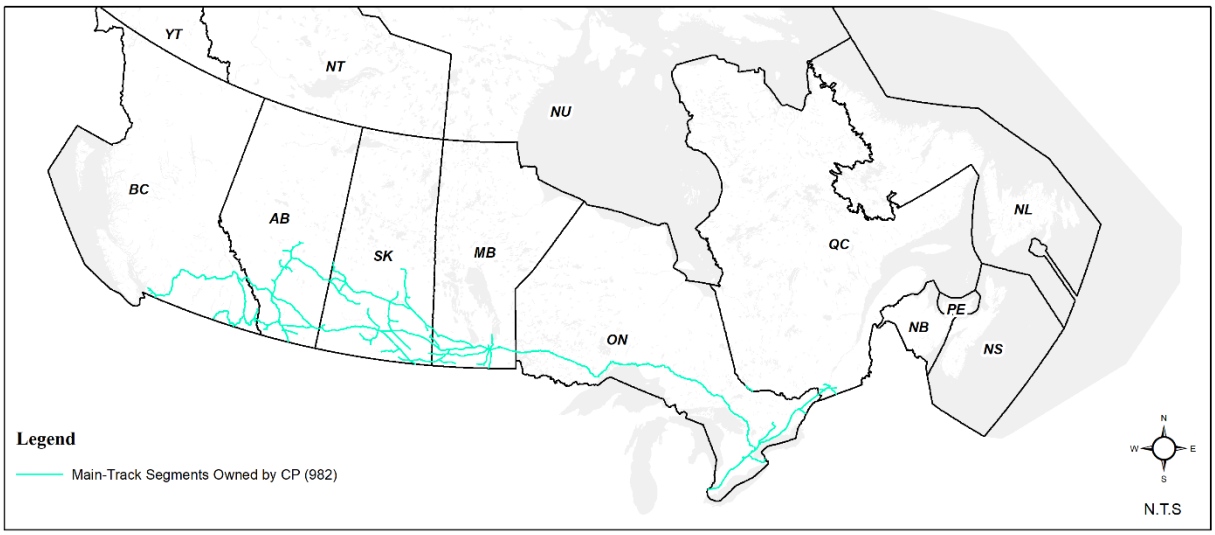


Figure 51: Track Segments Owned by CP

5.5.1. Outlier Analysis for Canadian Pacific Railway Model

An outlier analysis was performed on the CP database. When the outlier detection method was applied, the dataset was left with segments with no data for station count. In initial model testing, station count was omitted, but the models then produced no intuitive signs. As including the station count variable helped to maximize the number of permutations and combinations of model forms, it was decided that outliers associated with station count should be retained in the dataset. The same outlier detection approach was applied to the other independent variables.

Table 30: Summary Statistics for Key Explanatory Variables in CN Model

	Seg_length	Train Speed	Max Train Vol	Avg Train Vol	Station Count
Minimum	1	10	1	1	0
Q1	5	30	3	2	0
Median	11	40	10	8	0
Q3	17	55	16	13	0
Maximum	38	75	35	33	13
Mean	12	42	11	9	0
Range	37	65	34	33	13
Interquartile Range	13	25	13	10	0

The final CP dataset comprised 940 segments (96% of the original dataset for CP), as shown in **Figure 52**.

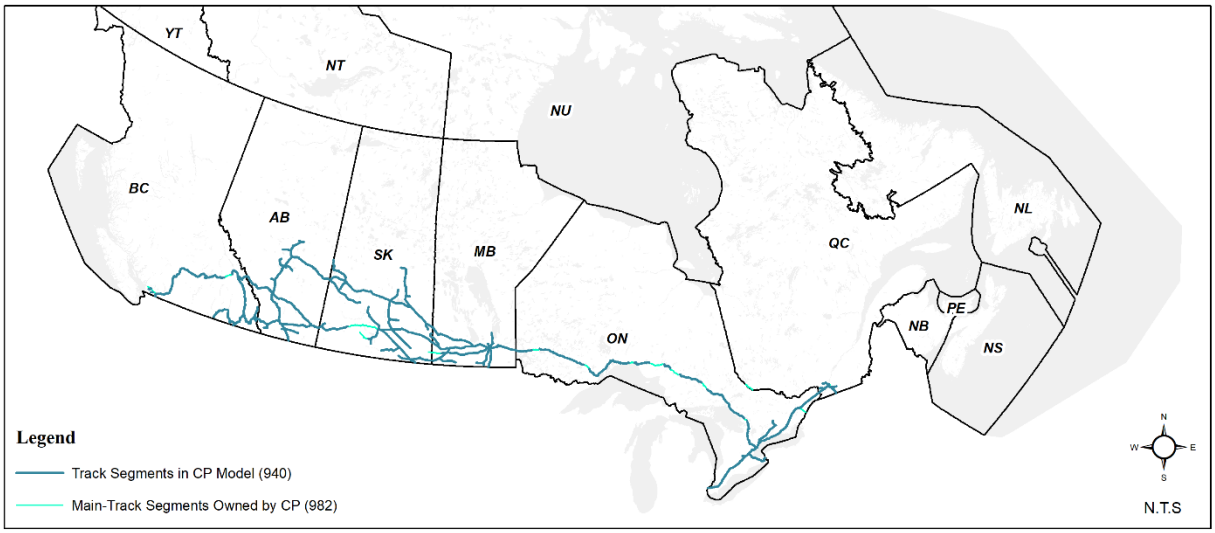


Figure 52: Track Segments in the CP Model

5.5.2. Model Comparison for Canadian Pacific Railway

In total, 88 candidate models were developed for predicting the number of derailments on CP-owned tracks. **Table 31** compares the results of the GOF tests applied to the calibration and validation data for the shortlisted CP models. Model ID#61 (shown in bold) was selected as the best-fitting model. **Chapter 5.5.3** continues the discussion of the results.

Table 31: Summary of Goodness-of-fit Tests for Shortlisted CP Models

Shortlisted Model	AIC	BIC	Dispersion	Calibration Data (70%) MSE	Calibration Data (70%) R2FT	Validation Data (30%) MSPE	Validation Data (30%) MPB	Validation Data (30%) MAD	Validation Data (30%) R2FT
4	900.95	927.90	1.59	0.43	14%	0.73	-0.11	0.51	17%
6	895.68	922.63	1.52	0.43	14%	0.74	-0.11	0.51	17%
23	899.80	926.75	1.33	0.45	11%	0.75	-0.10	0.51	16%
61	**900.03**	**931.48**	**1.63**	**0.43**	**15%**	**0.76**	**-0.12**	**0.51**	**15%**
62	902.02	937.96	1.63	0.43	15%	0.76	-0.12	0.51	15%
63	902.33	933.78	1.62	0.42	15%	0.73	-0.11	0.51	17%
64	900.97	932.41	1.60	0.43	15%	0.73	-0.11	0.51	17%
65	903.69	935.13	1.53	0.43	13%	0.75	-0.12	0.51	16%
68	905.46	936.91	1.44	0.44	13%	0.73	-0.09	0.51	17%
69	904.99	940.93	1.47	0.44	13%	0.73	-0.09	0.51	17%
72	903.76	930.71	1.53	0.43	13%	0.75	-0.11	0.51	15%
73	905.71	937.15	1.53	0.43	13%	0.76	-0.11	0.51	14%

5.5.3. Model Results for Canadian Pacific Railway

Figure 53 shows the CURE plot for the best-fitting model. The model predictions within the ±2 deviations over the entire range of daily train volume. The cumulative residual line almost converges at zero, indicating that the predicted numbers of derailments were very close to the observed number of derailments.

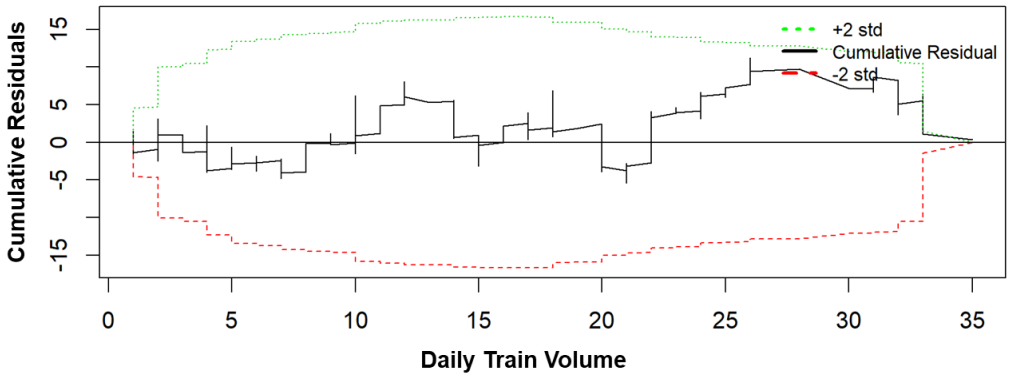


Figure 53: Cumulative Plot of Negative Binomial Model for CP

Equation 30 shows the functional form of the selected model for CP. This equation was used to predict the number of derailments for a 10-year period. The number of years is represented by *N* in the model form.

$$\mu_i = N \times exp(-5.4520) \times exp\begin{bmatrix} (0.4033 \times \text{VL\_Count}) + (0.0069 \times \text{VL\_TrSpd}) + \\ (-0.5734 \times \text{Stn\_Count}) + (0.0589 \times \text{Seg\_Length}) \\ +(0.0210 \times (\text{VL\_Count} \times \text{Stn\_Count}) \end{bmatrix} \quad \text{(Eq. 30)}$$

Table 32 presents the coefficient estimates and model statistics.

Table 32: Negative Binomial Model Results for CP

Variable	Estimate	Std. Error	z value	Pr(>|z|)	Model Statistics
(Intercept) ***	-5.4520	0.3231	-16.8749	0.0000	
log_VL_Count***	0.4033	0.1078	3.7402	0.0002	Dispersion Parameter: 1.632
VL_TrnSpd	0.0069	0.0081	0.8550	0.3925	Standard Error: 0.624
Stn_Count	-0.5734	0.3849	-1.4899	0.1363	Log Likelihood: -886.032
Seg_Length***	0.0589	0.0083	7.1354	0.0000	Observations: 940
VL_Count×Stn_Count	0.0210	0.0130	1.6194	0.1054	

****significance level of less than 0.0001*

In terms of statistical relationships, all the independent variables except station count showed a positive association with the number of derailments. Daily train traffic and segment length had p-values within the 95% confidence level. The p-value for the other variables were less than the threshold of 0.05 for the 95% confidence level, but the variables were retained as they helped improve the overall predictive power of the model.

5.6. Chapter Summary

This Chapter considered the development and results of the five prediction models using negative binomial regression: Canada, Eastern Canada, Western Canada, Canadian National Railway (CN), and Canadian Pacific Railway (CP). Five variables were used to predict the number of derailments: daily train volume, maximum train speed, segment length, average train volume and station count. Most of the models' predictions for train volumes in the lower range were poor, probably due to the zero observations on segments with low train volumes.

For each of the five models, the shortlisted models were compared using a set of GOF tests and CURE plots to determine the best-fitting models.

The statistical relationships between the response variable (derailments) and the independent variables showed a positive association between almost all the independent variables and the number of derailments. This finding indicated that increases in these variables elevated the risk of derailment. In the Canada model, the number of stations on a segment showed a negative association with the number of derailments. This finding was attributed to the relationship between train speed and the distribution of station locations. Train speed on a segment can be affected by whether stations are frequent or rare. When station locations are far apart, trains have more opportunities to accelerate and maintain speed which may increase the risk of derailment.

The inclusion of one or more interaction terms helped to produce parameters with intuitively acceptable signs and helped to improve overall model performance for the Eastern Canada, CN and CP models. In particular, the interaction between train speed and station count indicated that train speed and station count had a negative association (a lower number of stations on a segment contributed to higher train speed).

Chapter 6 uses the findings of this chapter together with the Empirical Bayes method to conduct a hotspot analysis.

CHAPTER 6. SAFETY NETWORK SCREENING

This chapter presents the hotspot analysis by employing the EB method as described in **Chapter 4.4**. Hotspots were identified based on the expected number of derailments for all five models.

6.1. Hotspot Analysis for Canada

The expected numbers of derailments for 10 years were calculated using the EB method. The EB estimates (shown as Exp. # of Derailments in **Table 33)** were used to rank the top 10 segments in Canada. Given the discussion of model performance discussed in **Chapter 5.1.3**, the selected model underestimates the number of derailments.

Figure 54 shows the locations of the segments. These segments are located in four provinces; Alberta, British Columbia, Ontario and Saskatchewan. Five segments are located in British Columbia, particularly in the mountainous regions. This suggests that regions with more severe terrains may pose challenges to track geometry and train handling because of steeper slopes and excess speeds.

The owners of the top 10 segments are also presented in **Table 33** and **Figure 55**. CP owns nine out of 10 segments.

Table 33: Safety Network Screening Results of Canada

Seg ID	Owner Name	Province	Subdivision Name	Obs. # of Derailments	Exp. # of Derailments	Rank	
						Obs	Exp
12258	Canadian Pacific	British Columbia	Mountain	6	4.02	2	1
12520	Canadian Pacific	British Columbia	Mountain	5	3.18	5	2
10481	Canadian Pacific	Ontario	Heron Bay	4	3.05	8	3
12344	Canadian Pacific	British Columbia	Mountain	7	2.92	1	4
10154	Canadian Pacific	Ontario	Heron Bay	4	2.86	8	5
10967	Canadian Pacific	Saskatchewan	Swift Current	3	2.74	15	6
12836	Canadian Pacific	Alberta	Laggan	6	2.67	2	7
12294	Canadian National	British Columbia	Chetwynd	4	2.31	8	8
12441	Canadian Pacific	British Columbia	Mountain	3	2.20	15	9
10829	Canadian Pacific	Saskatchewan	Maple Creek	3	2.16	15	10

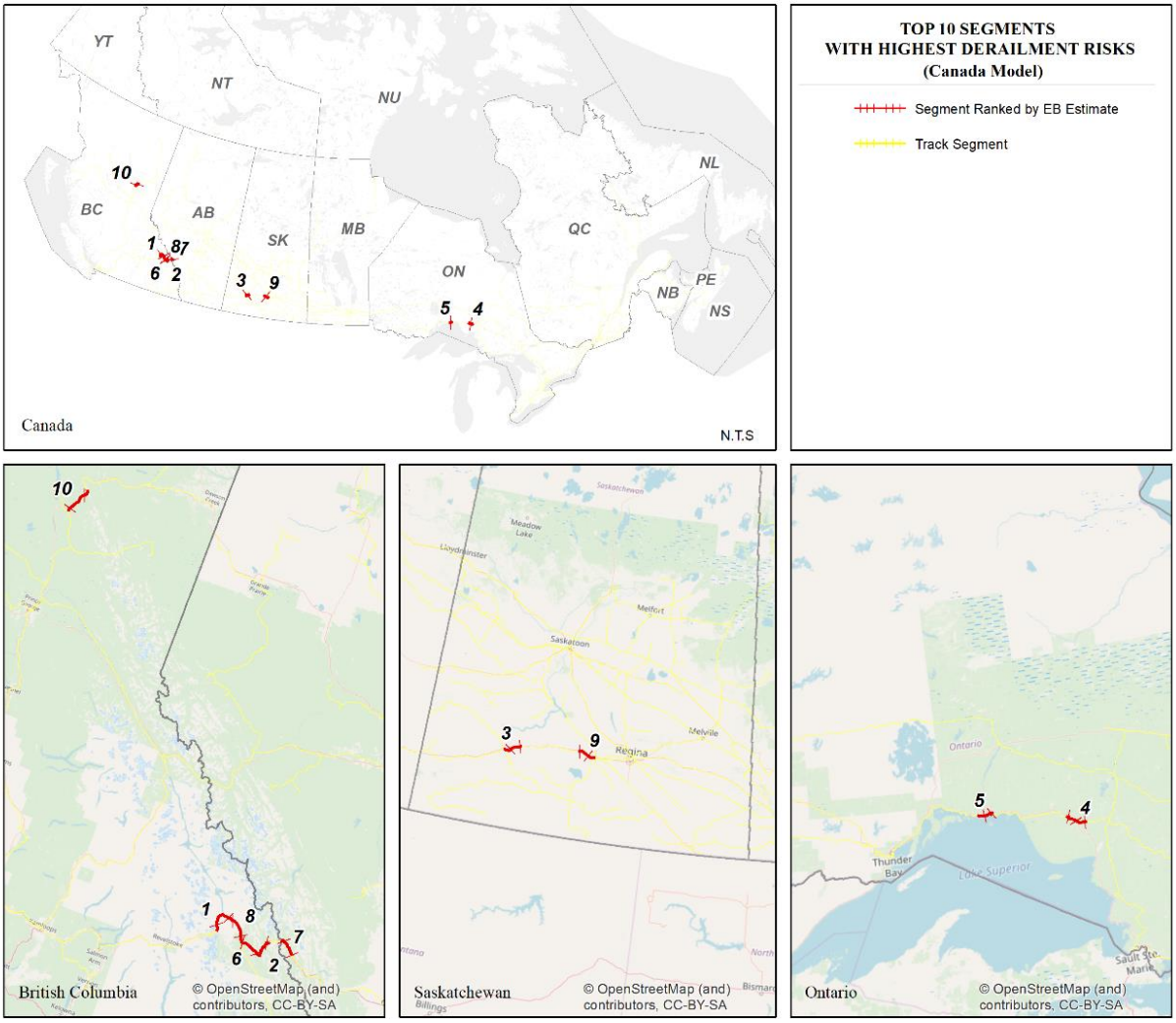


Figure 54: Top 10 Segments with Highest Derailment Risks in Canada

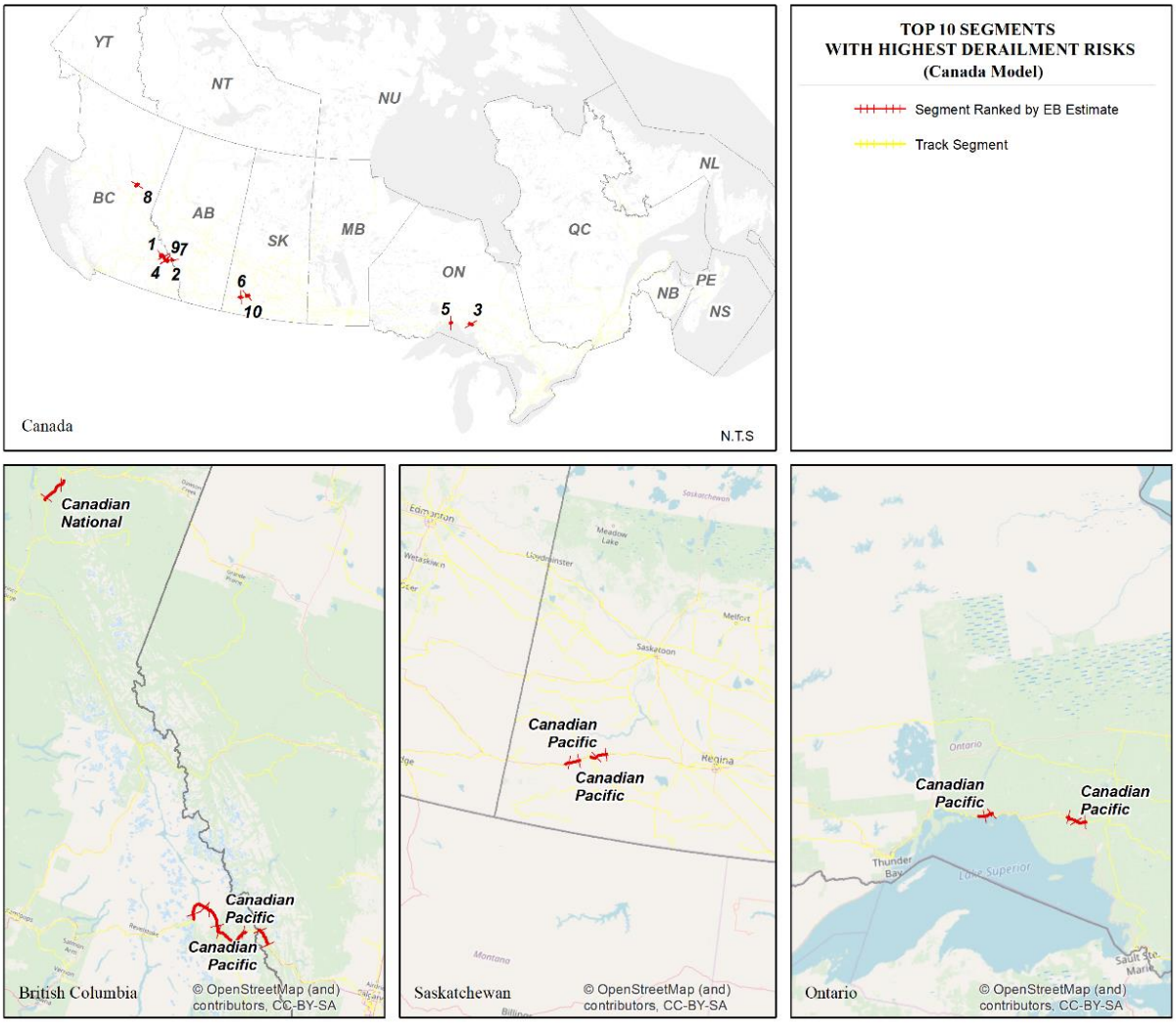


Figure 55: Track Owners of Top 10 Segments with Highest Derailment Risks in Canada

6.2. Hotspot Analysis for Eastern Canada

Table 34 summarizes the network screening results for Eastern Canada. The top 10 segments were ranked based on the expected numbers of derailments. As evident from model results presented in **Chapter 5.2.3**, the selected model underestimates the number of derailments.

Figure 56 shows the locations of these segments. The top 10 segments are located in Ontario (8) and Quebec (2). With respect to track ownership, CN owns seven of the segments as presented in **Figure 57**. The remaining segments are owned by CN (2) and Iron Ore Company of Canada (1).

Table 34: Safety Network Screening Results for Eastern Canada

Seg ID	Owner Name	Province	Subdivision Name	Obs. # of Derailments	Exp. # of Derailments	Rank Obs	Rank Exp
10481	Canadian Pacific	Ontario	Heron Bay	4	2.65	3	1
10154	Canadian Pacific	Ontario	Heron Bay	4	2.59	3	2
12948	Canadian Pacific	Ontario	Winchester	3	1.93	7	3
11585	Canadian Pacific	Quebec	Temiscaming	5	1.89	2	4
10254	Canadian Pacific	Ontario	White River	3	1.75	7	5
11578	Iron Ore Company of Canada	Quebec	Wacouna	4	1.74	3	6
11374	Canadian National	Quebec	Lac St-Jean	6	1.74	1	7
12950	Canadian Pacific	Ontario	Winchester	2	1.60	12	8
10538	Canadian Pacific	Ontario	White River	2	1.60	12	9
10373	Canadian National	Ontario	Ruel	2	1.58	12	10

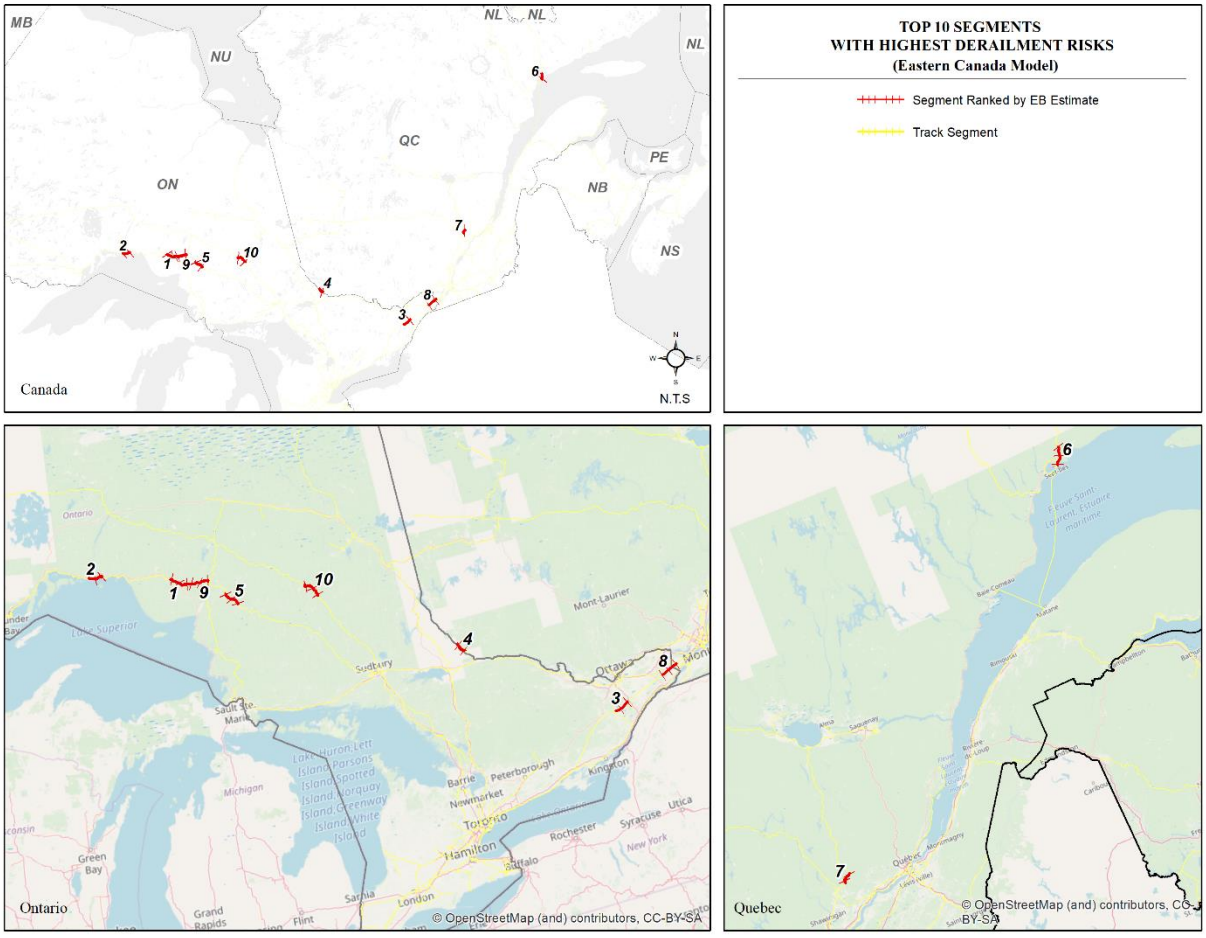


Figure 56: Top 10 Segments with Highest Derailment Risks of Eastern Canada Model

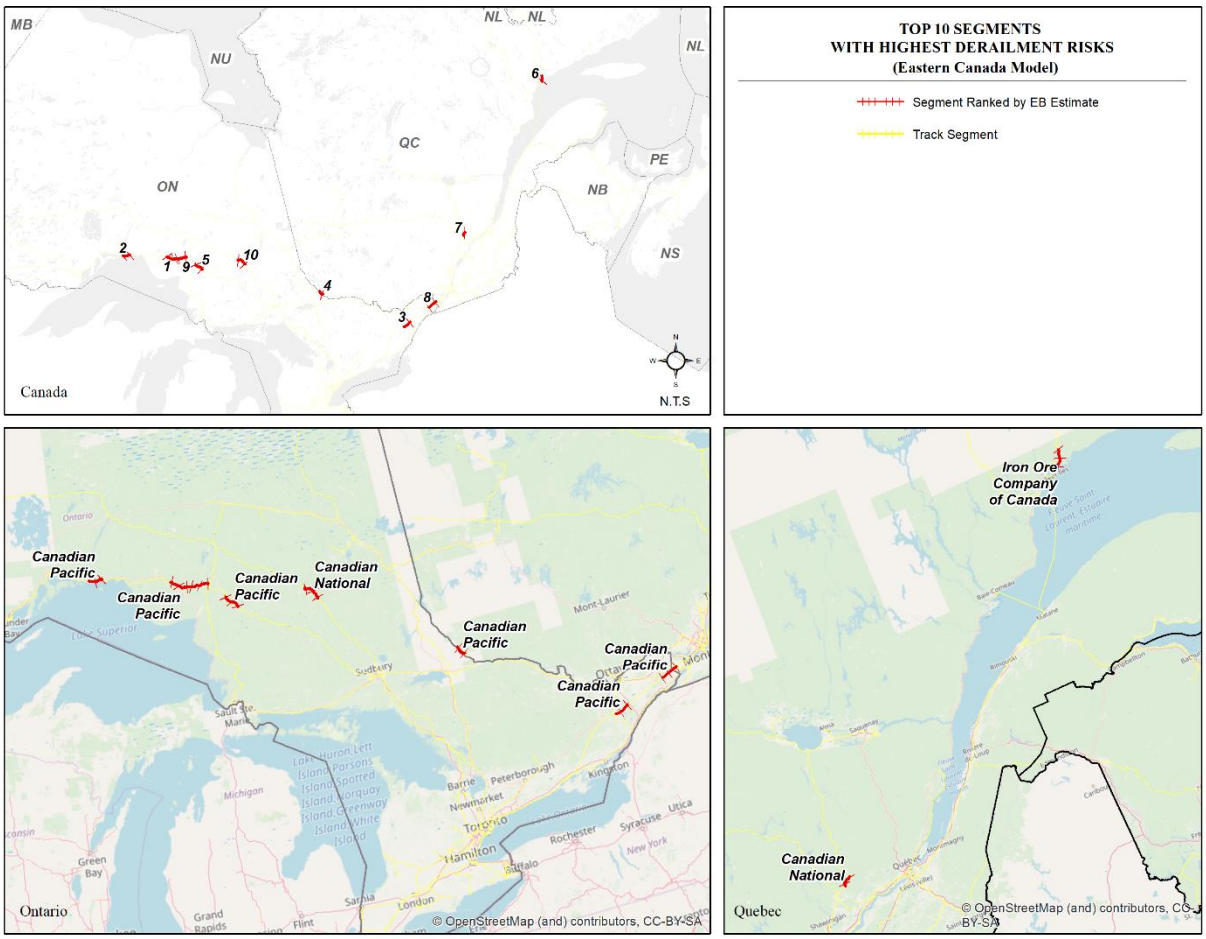


Figure 57: Track Owners of Top 10 Segments with Highest Derailment Risks in Eastern Canada

6.3. Hotspot Analysis for Western Canada

Table 35 summarizes the network screening results for Western Canada. The top 10 segments were ranked based on the expected numbers of derailments. As previously discussed in **Chapter 5.3.3**, the selected model underestimates the number of derailments.

Figure 58 shows the locations of these segments. The segments are located in three provinces; British Columbia, Alberta and Saskatchewan. Six of the top 10 segments are situated in British Columbia. The remaining four segments are located in Saskatchewan (2) and Alberta (2). As observed in the Canada model, segments in British Columbia are situated in mountainous areas which indicates that terrain is likely an influential factor affecting derailments.

Figure 59 presents a map of EB estimates by ownership. The owners of the top 10 segments are CP (5), CN (4) and Burlington Northern Santa Fe Railway (1).

Table 35: Safety Network Screening Results for Western Canada

Seg ID	Owner Name	Province	Subdivision Name	Obs. # of Derailments	Exp. # of Derailments	Rank Obs.	Rank Pre.
12344	Canadian Pacific	British Columbia	Mountain	7	3.07	1	1
12294	Canadian National	British Columbia	Chetwynd	4	2.79	5	2
12520	Canadian Pacific	British Columbia	Mountain	5	2.58	3	3
12316	Canadian National	British Columbia	Chetwynd	3	2.37	8	4
10829	Canadian Pacific	Saskatchewan	Maple Creek	3	2.35	8	5
11173	Canadian Pacific	Saskatchewan	Maple Creek	3	2.24	8	6
12359	Canadian National	British Columbia	Chetwynd	3	2.21	8	7
12836	Canadian Pacific	Alberta	Laggan	6	2.05	2	8
12622	Canadian National	Alberta	Lac La Biche	3	1.91	8	9
12447	Burlington Northern Santa Fe Railway	British Columbia	New Westminster	3	1.86	8	10

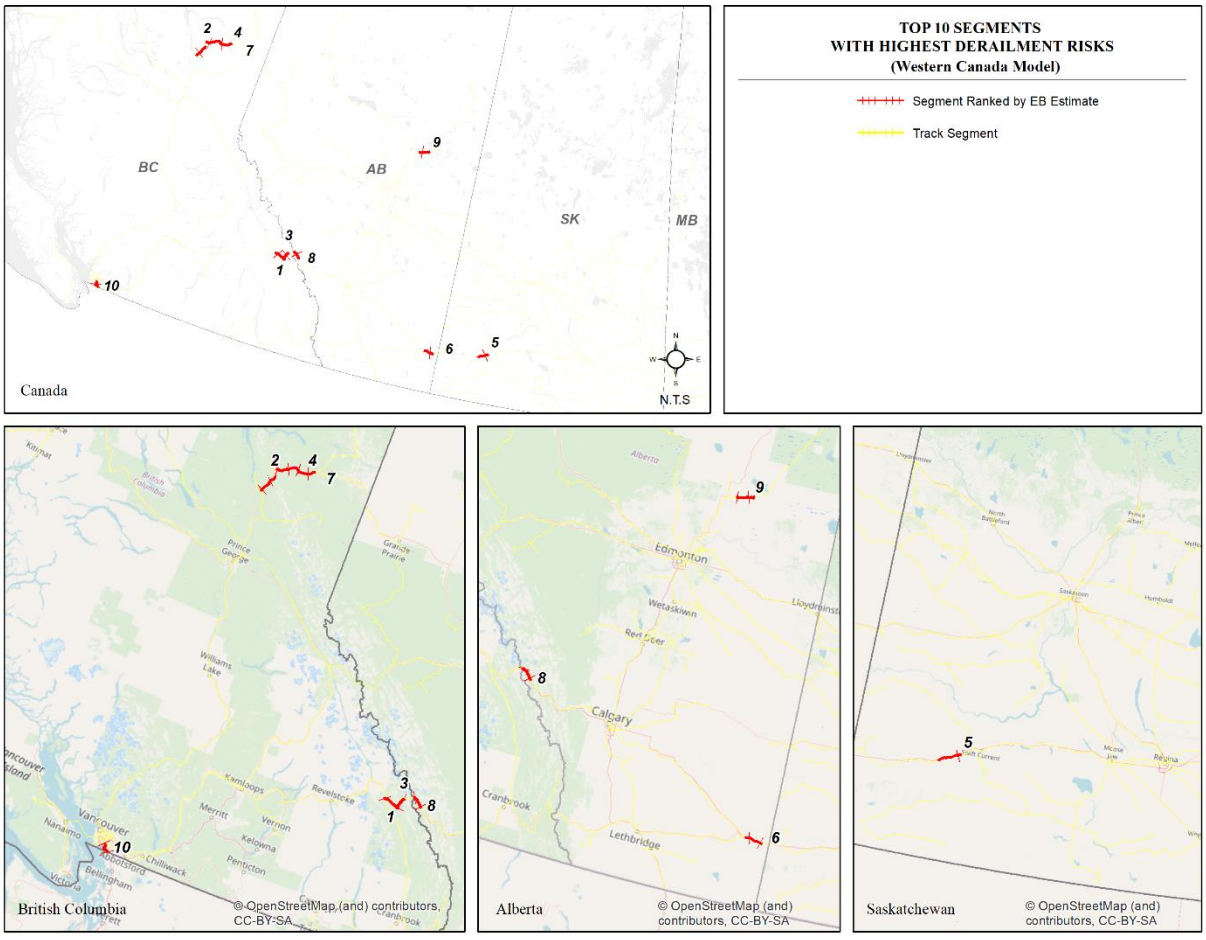


Figure 58: Top 10 Segments with Highest Derailment Risks in Western Canada

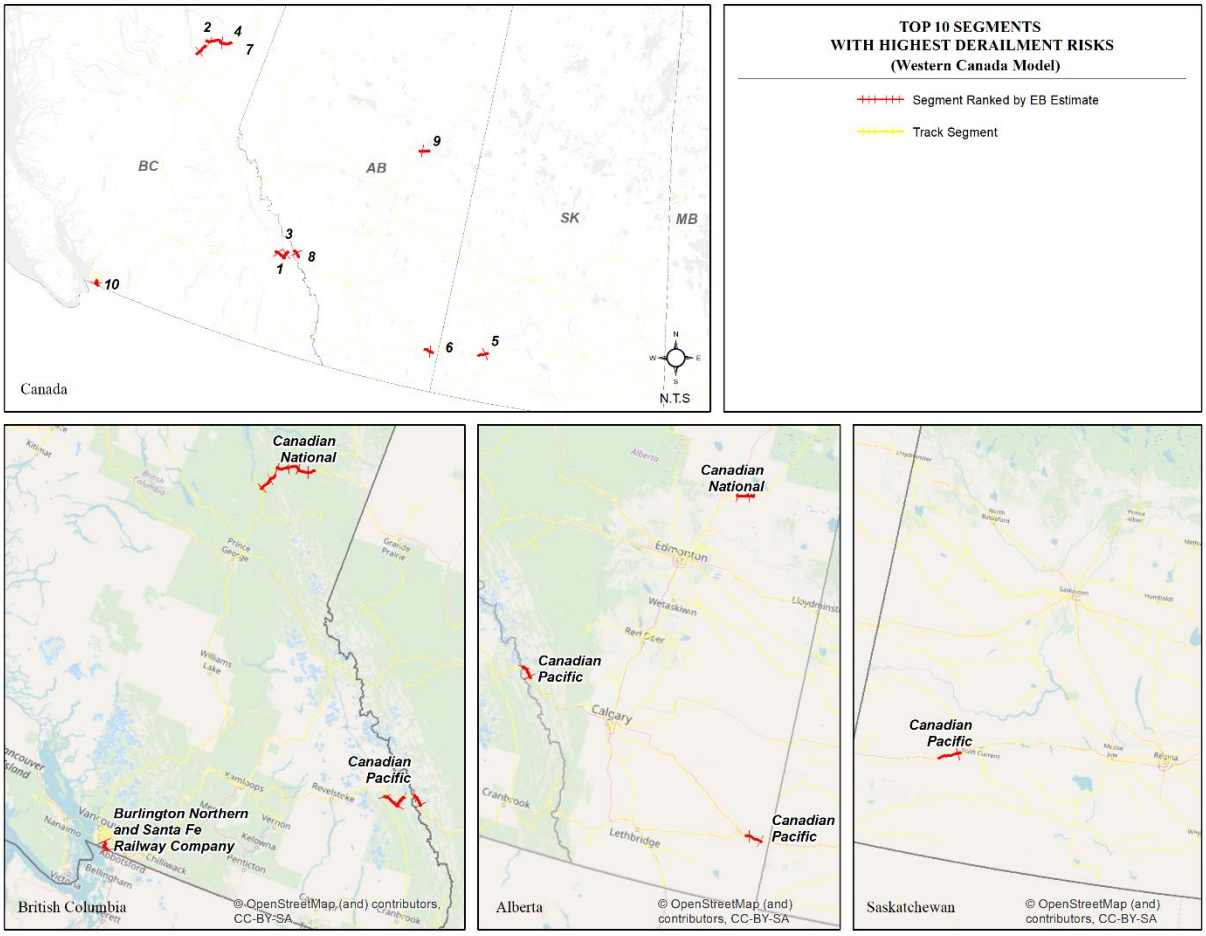


Figure 59: Track Owner of Top 10 Segments with Highest Derailment Risks in Western Canada

6.4. Hotspot Analysis for Canadian National Railway

Table 36 summarizes the network screening results for CN's rail network. The top 10 segments were ranked based on the expected numbers of derailments. The model underestimated the number of derailments in most cases except for two segments located in Alberta (Seg IDs: 12714 and 12571). There are higher numbers of stations along these two segments which led to overestimations by the model.

Figure 60 shows the locations of these segments. The segments are found in four provinces, nine of these are in Western Canada (British Columbia (4), Alberta (4) and Manitoba (1)). One segment is located in Ontario.

Table 36: Safety Network Screening Results for CN

Seg ID	Owner Name	Province	Subdivision Name	Obs. # of Derailments	Exp. # of Derailments	Rank Obs.	Rank Pre.
10081	Canadian National	Ontario	Kingston	4	2.23	2	1
12120	Canadian National	British Columbia	Squamish	4	1.72	2	2

Seg ID	Owner Name	Province	Subdivision Name	Obs. # of Derailments	Exp. # of Derailments	Rank	
						Obs.	Pre.
12294	Canadian National	British Columbia	Chetwynd	4	1.67	2	3
11769	Canadian National	Manitoba	Redditt	2	1.65	21	4
12788	Canadian National	Alberta	Lac La Biche	3	1.60	8	5
12316	Canadian National	British Columbia	Chetwynd	3	1.48	8	6
12714	Canadian National	Alberta	Mountain Park	1	1.48	69	7
12806	Canadian National	Alberta	Edson	4	1.43	2	8
12571	Canadian National	Alberta	Foothills	1	1.41	69	9
12359	Canadian National	British Columbia	Chetwynd	3	1.37	8	10

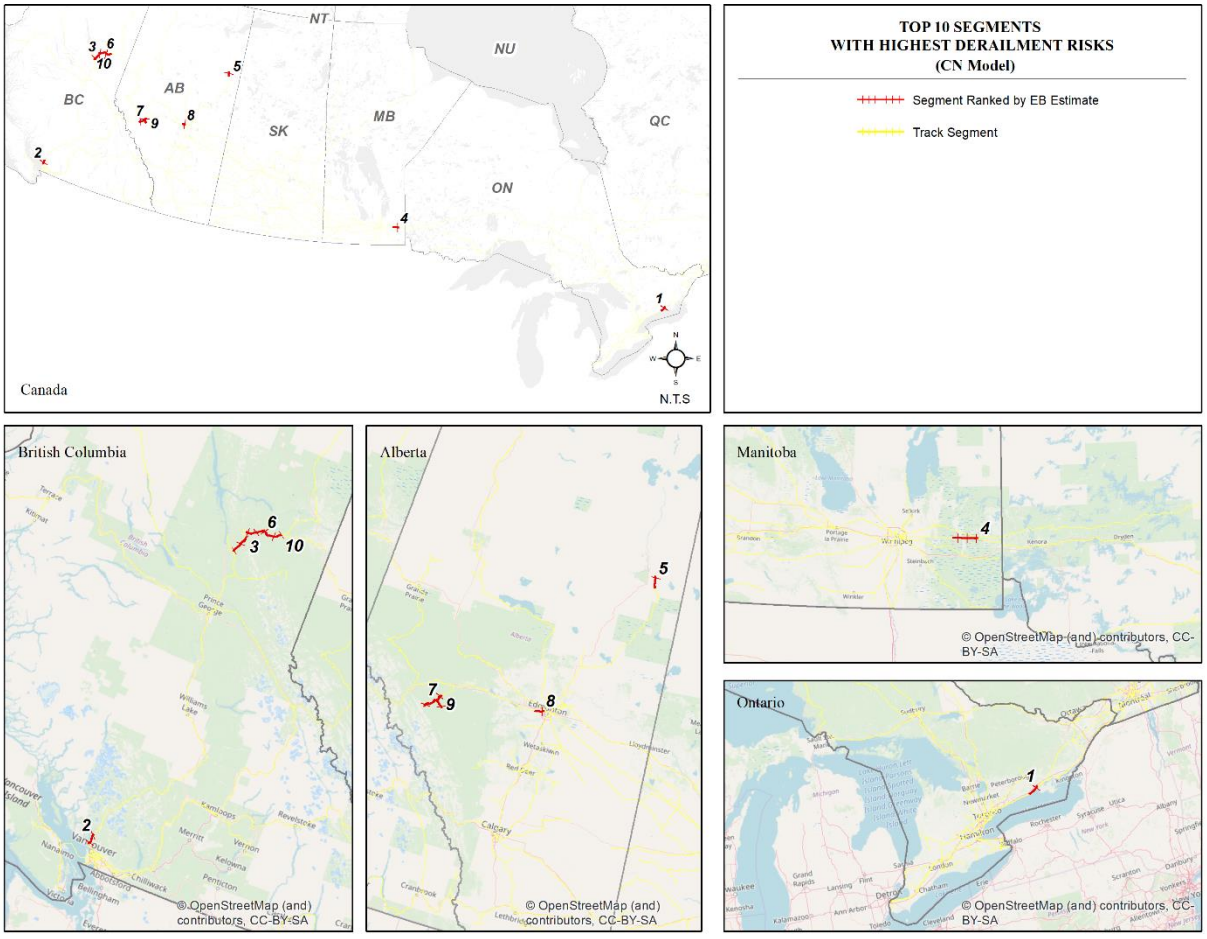


Figure 60: Top 10 Segments with Highest Derailment Risks for CN

6.5. Hotspot Analysis for Canadian Pacific Railway

Table 37 summarizes the network screening results for CN's rail network. The top 10 segments were ranked based on the expected numbers of derailments.

Figure 61 shows that seven segments are located in Western Canada; British Columbia (5) and Saskatchewan (2). The other three segments are located in Eastern Canada; Ontario, Alberta and Quebec. The continuous segments in British Columbia are in mountainous regions. The rough terrains in these regions and the sharp curvatures may increase the derailment potential along these segments.

Table 37: Safety Network Screening Results for CP

Seg ID	Owner Name	Province	Subdivision Name	Obs. # of Derailments	Exp. # of Derailments	Rank Obs.	Rank Pre.
12258	Canadian Pacific	British Columbia	Mountain	6	4.52	2	1
11496	Canadian Pacific	Quebec	Vaudreuil	4	4.10	6	2
12520	Canadian Pacific	British Columbia	Mountain	5	3.49	4	3
12836	Canadian Pacific	Alberta	Laggan	6	2.94	2	4
10154	Canadian Pacific	Ontario	Heron Bay	4	2.79	6	5
12344	Canadian Pacific	British Columbia	Mountain	7	2.76	1	6
12441	Canadian Pacific	British Columbia	Mountain	3	2.42	11	7
12464	Canadian Pacific	British Columbia	Cascade	2	2.34	27	8
10829	Canadian Pacific	Saskatchewan	Maple Creek	3	2.22	11	9
11173	Canadian Pacific	Saskatchewan	Maple Creek	3	2.10	11	10

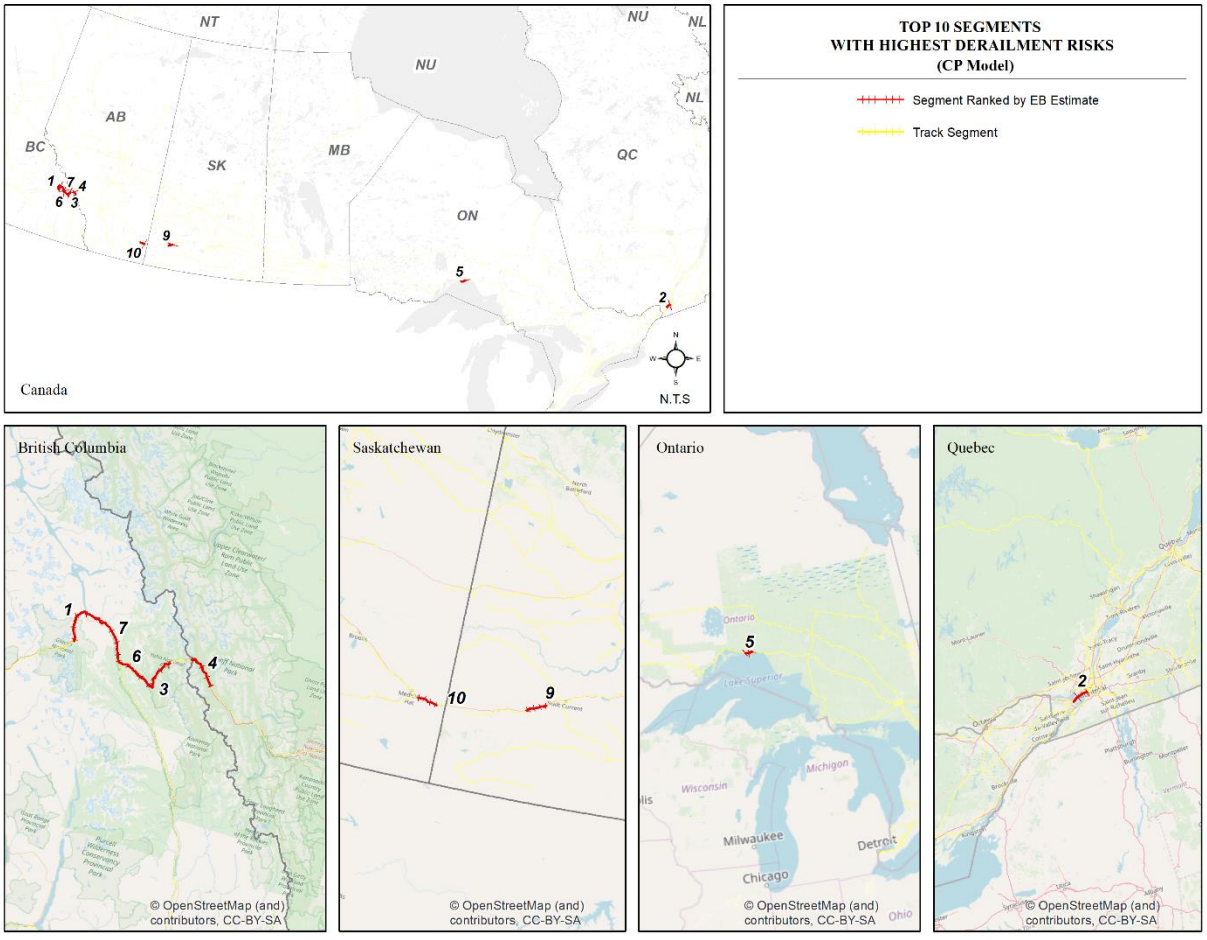


Figure 61: Top 10 Segments with Highest Derailment Risks for CP

6.6. Chapter Summary

This chapter discussed the network screening results for each model. The expected numbers of derailments calculated by the Empirical Bayes (EB) method were used to rank segments. Information of the top 10 segments ("hotspots") in each model were presented. The network screening results allow governing agencies and/or railway companies to prioritize segments for investigation and/or safety improvement.

Table 38 shows the segments that have been identified as the top 10 segments in more than one model. Of the 12 segments, seven of them are located in British Columbia. The other five segments are located in Ontario (2), Saskatchewan (2) and Alberta (1).

Table 38: Segments Identified as Hotspots in Multiple Models

Seg ID	Owner Name	Province	Canada	CN	CP	East CA	West CA
10154	Canadian Pacific	Ontario	●		●	●	
10481	Canadian Pacific	Ontario	●			●	
10829	Canadian Pacific	Saskatchewan			●		●
11173	Canadian Pacific	Saskatchewan			●		●
12258	Canadian Pacific	British Columbia	●		●		
12294	Canadian National	British Columbia	●				●
12316	Canadian National	British Columbia	●		●		●
12344	Canadian Pacific	British Columbia	●		●		
12359	Canadian National	British Columbia	●		●		●
12441	Canadian Pacific	British Columbia		●			●
12520	Canadian Pacific	British Columbia	●		●		●
12836	Canadian Pacific	Alberta	●		●		●

In addition to the ranking of segments, many hotspots are located in mountainous areas. This confirms the observation that topography may have an effect on derailment potential. Higher risk can be expected along segments that have sharp curvatures and/or steep grades.

CHAPTER 7. DEVELOPMENT OF BINOMIAL LOGISTIC MODELS

This chapter discusses the development, analysis and results of the five prediction models using binomial logistic regression modelling method. The outcome variable has two possible values; whether derailment is present or not on a segment. The same independent variables are considered in model development as previously discussed in **ChapterCHAPTER 5**.

7.1. Logit Model Development

Logistic regression modelling was performed using the same datasets for the five models as presented in **Chapter 5.1**. The goal is to obtain the best-fitting model based on the ability of correctly classifying segments with or without derailment for a 10-year period. It is of particular interest to determine the sensitivity of logit models in handling data with excess zeros.

The logit model predicts the probability of derailments. If the probability of derailment is greater than or equal to the chosen cut-off value on the probability scale (default value is presumed to be 0.5), then the segment would be dichotomized as 1, otherwise 0. The outcome variable Y is a binary output:

$$y_i = \begin{cases} 1 \text{ if derailment risk is present} \\ 0 \text{ otherwise} \end{cases}$$

The model predicts the probability (P) of Y=1, P (Y=1 | Y=0), as a function of x, independent variable(s). The logistic regression model estimates the logit-transformed probability as a linear relationship with the independent variables, as given in **Equation 31** below.

$$\text{logit}(p) = \log\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 x_1 + \cdots + \beta_n x_n \quad \text{(Eq. 31)}$$

The above equation can be re-expressed to derive the model form for calculating the probability of outcome Y, P (Y=1), as shown in **Equation 32**.

$$\text{P(x)} = \frac{exp(\beta_0 + \sum_{\text{i=1}}^{\text{n}} \beta_i x_{\text{i}})}{1 + exp(\beta_0 + \sum_{\text{i=1}}^{\text{n}} \beta_i x_{\text{i}})} \quad \text{(Eq. 32)}$$

Where,

P means the probability that the predicted variable will have a value of 1;

$\beta_0, \ldots \beta_n$ are regression coefficients;

$x_0, \ldots x_n$ are independent variables.

The logit model used Maximum Likelihood Estimation method for variable estimation, as given in **Equation 33**.

$$l(\beta) = \prod_{i=1}^{n} P_i^{y_i} (1-p)^{1-y_i} \quad \text{(Eq. 33)}$$

The log likelihood function of logit model is **Equation 34**.

$$\ln l(\beta) = \sum_{i=1}^{n} [y_i x_i \beta - ln(1 + x_i \beta)] \quad \text{(Eq. 34)}$$

Five independent variables of interest were considered: maximum daily train volume (VL_Count), maximum train speed (VL_TrnSpd), segment length (Seg_Length), average daily train volume (Avg_Train) and number of stations on a segment (Stn_Count). A forward selection approach was used in determining the best subset of independent variables that produces the best-fitting model. Candidate models were developed with different combination of independent variables and interaction terms. The candidate models with good predictive capability were shortlisted on the basis of an intuitive interpretation of the signs of the coefficient estimates. Six GOF tests evaluated the relative performance of the shortlisted models, and the best-fitting model was selected.

Appendix D provides the model forms of the candidate models.

7.2. Model Calibration and Validation

The same model calibration and validation method was used, as previously described in **Chapter 4.5**. Based on a sensitivity test, a 70:30 split was selected and applied to all models. A set of goodness-of-fit tests were applied to the calibration dataset for model comparison. The selected models were further assessed by additional performance measures using the validation dataset. The goodness-of-fit tests that were used to assess the performance of the logit models are discussed in the next section.

7.3. Goodness-of-fit Tests for Logit Model

For each model, candidate model forms were developed with different combination of independent variables. Models were then shortlisted if they estimated coefficients with intuitive signs. Six goodness-of-fit (GOF) tests were performed to evaluate and compare the relative performance of the shortlisted models. The GOF tests were: AIC, BIC, deviance statistics, McFadden R-squared and prediction accuracy, classification tables, sensitivity and specificity and receiver operating characteristic curve. The definitions and applications of AIC and BIC are explained in **Chapter 4.6**.

7.3.1. Deviance Statistics

Logistic regression model calculates the null deviance and residual deviance. Null deviance is estimated when the model is only considering the intercept whilst the residual deviance is estimated with the inclusion of independent variables. Deviance can be used to measure model adequacy where smaller deviance value indicates better model fit (Portugués, 2018). **Equations 35** and **36** show the formulas for null and residual deviances.

$$D_0 = -2logL(\hat{\beta}_0) \quad \text{(Eq.35)}$$

Where:

D_0 denotes Null Deviance, and

$logL(\hat{\beta}_0)$ denotes the maximum loglikelihood of a model with intercept only.

$$D = -2logL(\hat{\beta}) + logL(S) \quad \text{(Eq.36)}$$

Where:

D denotes Residual Deviance,

$logL(\hat{\beta})$ denotes the maximum loglikelihood the fitted model with parameters, and

$logL(S)$ denotes the maximum loglikelihood the saturated model.

The change in deviance determines whether or not the inclusion of independent variables improve the fit of the model. A reduction in deviance between null and residual deviances and the low p-value indicate that the inclusion of independent variables has improved the model fit.

This is asymptotically equivalent to a chi-square distribution (**Equation 37**) with the degrees of freedom equal to the number of parameters in the model for determining statistical significance (Cook, et al., 2000). If the fitted model is perfect, then D = 0 and $R^2 = 1$. On the contrary, if the fitted model does not improve the performance at all, then D = D_0 and $R^2 = 0$. Thus, the higher the value, the better the model fit.

$$R^2 = 1 - \frac{D}{D_0} \quad \text{(Eq.37)}$$

Where:

R^2 represents Chi-squared, and

D_0 *and* D are the same as **Equations 35** and **36**.

7.3.2. McFadden R-Squared

McFadden R-Squared ($\boldsymbol{R^2_{McF}}$) is one of the r-squared measures for evaluating model fit in comparison to the null model (McFadden, 1974). **Equation 38** is the model form for calculating McFadden R-Squared.

As shown, $R^2_{MF} = 1$ if $logL(\beta_0)$ = 0. The value lies between 0 and 1 where closer to one means better model fit. Some researchers have noted that this measure can be interpreted as a percentage of how much uncertainty can be explained by the model (Hauer, 1978; Windmeijer, 1995)

$$R^2_{MF} = 1 - \frac{logL(\beta_0)}{Log\ L_0} \quad \text{(Eq.38)}$$

Where:

$Log\ L_0$ denotes the maximum loglikelihood of the null model, with inclusion of only the intercept (all coefficient equal to zero), and

$logL(\beta_0)$ denotes the maximum loglikelihood the fitted model with parameters.

7.3.3. Prediction Accuracy

A prediction accuracy rate measures how often the model can correctly classify the outcome variable (whether derailment is present along a segment). It is applied to the validation dataset by taking into consideration of true and false prediction of positive and negative values. **Equation 39** shows the formula for calculating prediction accuracy:

$$Accuracy = \frac{TP + TN}{TP + FN + FP + TN} \quad \text{(Eq.39)}$$

Where:

TP is true positive,

FN is false negative,

TN is true negative, and

FP is false positive.

7.3.4. Classification Tables

The predicted probabilities from the logistic model can be used to define whether or not a derailment is predicted on a segment. Once the fitted values are calculated based on the cut-off value, a two by two table can be developed to compare the observed values with the fitted values. **Table 39** shows a sample classification table.

Table 39: A Sample Classification Table

		Observed	
		0	**1**
Prediction	**0**	TN	FN
	1	FP	TP

Higher numbers for TN and TP indicate better fit of a model. In this study, classification tables were developed to evaluate the predictive capability of the best-fitting model.

7.3.5. Sensitivity and Specificity

Sensitivity (true positive) and specificity (true negative) rates are metrics that evaluates a model's the classification ability. Sensitivity and specificity define the truth detection ratios of the model. Sensitivity indicates the proportion of segments with derailments are being correctly classified as 1 in the model. Specificity refers to the proportion of segments without derailments are being correctly classified 0 in the model. Higher sensitivity and specificity indicate better model fit. The formulas for calculating these two terms are given in **Equations 40** and **41** below.

$$Sensitivity = \frac{TP}{TP + FN} \quad \text{(Eq.40)}$$

$$Specificity = \frac{TN}{TN + FP} \quad \text{(Eq.41)}$$

Where:

TP, FN, TN and FP are the same as **Equation 39.**

7.3.6. Receiver Operating Characteristic Curve

A receiver operating characteristic (ROC) curve is a plot of all possible pairs of sensitivity against one minus specificity across a full range of cut-off values from 0 to 1. **Figure 62** shows a sample ROC curve.

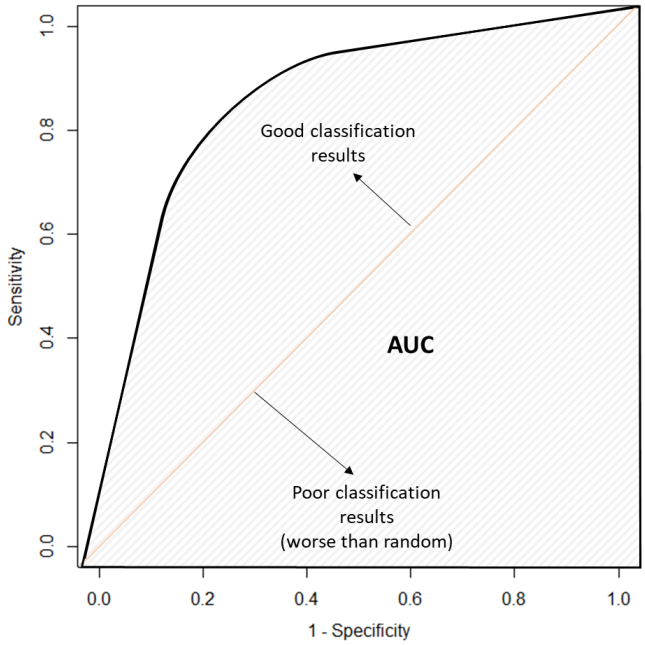


Figure 62: Sample ROC Curve

The closer the area under curve (AUC) to the upper top corner (higher the AUC) the better the model in terms of class classification (the ability to distinguish between 0 and 1). When an AUC is greater than 0.8, it is an indication of a good performing model with high classification capacity (Choi et al., 2019). Model with AUC greater than 0.5 is generally acceptable.

7.4. Logit Model for Canada

In total, 70 candidate models were developed for predicting the probability of derailment on all track segments in Canada.

7.4.1. Comparison of Logit Models for Canada

Table 40 compares the results of the GOF tests applied to the calibration data for the shortlisted Canada models. Model ID#8 (shown in bold) was selected as the best-fitting model. **Chapter 7.4.2** continues the discussion of the results.

Table 40: Comparison of Candidate Logit Mode ls for Canada

Model ID	AIC	BIC	Null Dev	Res Dev	Null Dev – Res Dev (R^2_{Chi} Test)	R^2_{McF}	Model Accuracy
4	1649.7	1677.2	1806.60	1639.71	166.89	0.09	78%
7	1616.71	1644.24	1806.60	1606.71	199.89	0.11	78%
8	**1612.73**	**1640.26**	**1806.60**	**1602.73**	**203.87**	**0.11**	**78%**
9	1643.22	1670.74	1806.60	1633.22	173.38	0.10	78%
35	1618.68	1651.71	1806.60	1606.68	199.92	0.11	78%
36	1614.67	1647.70	1806.60	1602.67	203.93	0.11	78%
51	1625.37	1652.90	1806.60	1615.37	191.23	0.11	77%
67	1626.49	1659.53	1806.60	1614.49	192.11	0.11	78%
68	1628.29	1666.83	1806.60	1614.29	192.31	0.11	78%
69	1649.44	1682.47	1806.60	1637.44	169.16	0.09	79%

7.4.2. Logit Model Results for Canada

This section presents the functional form of the selected model, the coefficient estimates, model statistics and considers the effects of independent variables on the outcome variable.

Equation 42 shows the model form of the best-fitting model for Canada. This equation predicts the probability of derailment to occur on a segment for a 10-year period. The number of years is represented by *N* in the model form.

$$\text{logit}(p) = N \times [(-11.0171) + (0.0239 \times \text{VL\_Count}) + (1.4490 \times \text{logVL\_TrnSpd}) + (0.0458 \times \text{Stn\_Count}) + (0.6794 \times \text{logSeg\_Length})] \quad \text{(Eq.42)}$$

Table 41 presents the coefficient estimates and model statistics.

Table 41: Logit Model Results for Canada

Variable	Estimate	Std. Error	z value	Pr(>|z|)	Model Statistics
(Intercept) ***	-11.0171	0.8301	-13.2718	0.0000	Prediction Accuracy: 0.77 AIC: 1612.73 Null Deviance: 1806.60 Residual Deviance: 1602.73 Observations: 2,553
VL_Count**	0.0239	0.0077	3.1224	0.0018	
log_VL_TrnSpd***	1.4490	0.2223	6.5183	0.0000	
Stn_Count	0.0458	0.0864	0.5303	0.5959	
log_Seg_Length***	0.6794	0.0866	7.8448	0.0000	

***significance level of less than 0.001*

****significance level of less than 0.0001*

Table 41 shows that all independent variables have positive effects on the probability of derailment. In logistic modelling, the coefficient estimate represents the log odds of each variable for affecting the outcome variable (derailment). The log odds were converted to odd ratios by exponentiating the coefficient. For instance, an increase in daily train volume changes the odds of derailment vs. no derailment by a factor of exp (0.0239) = 1.02, when all the other variables remain the same.

Chi-squared test was performed to assess the statistical significance of the coefficient estimates. For this model, a chi-squared test statistic of 203.87 with four degrees of freedom and a p-value of less than 0.001 indicates that the overall effects of daily train volume, maximum train speed, station count and segment length were statistically significant. The reduction in deviance also indicates that the inclusion of independent variables improved the overall model fit.

7.4.3. Logit Model Validation for Canada

The model was further evaluated by performing additional GOF tests on the validation data. These tests included: prediction accuracy, classification tables, sensitivity and specificity rates, and ROC curve.

The prediction accuracy of the selected model was 78%. **Table 42** summarizes the number of true and false positives, as well as true and false negatives predicted by the model. The model statistics show that the selected model had a sensitivity rate of 10% and a specificity rate of 97%. Although the model obtained a high prediction accuracy score, it showed a bias in predicting zero events.

Table 42: Classification Table for Canada Model

		Observed		Model Statistics
		0	**1**	Accuracy: 0.784 95% CI: (0.7524, 0.8132) Sensitivity: 0.1032 Specificity: 0.9656 Detection Rate: 0.7622
Prediction	**0**	561	139	
	1	20	16	

Figure 63 shows the ROC curve for the Canada model. The calculated AUC for the selected model was 0.70 on the validation dataset which can be interpreted as there is a 70% chance that the model can distinguish between positive and negative classes.

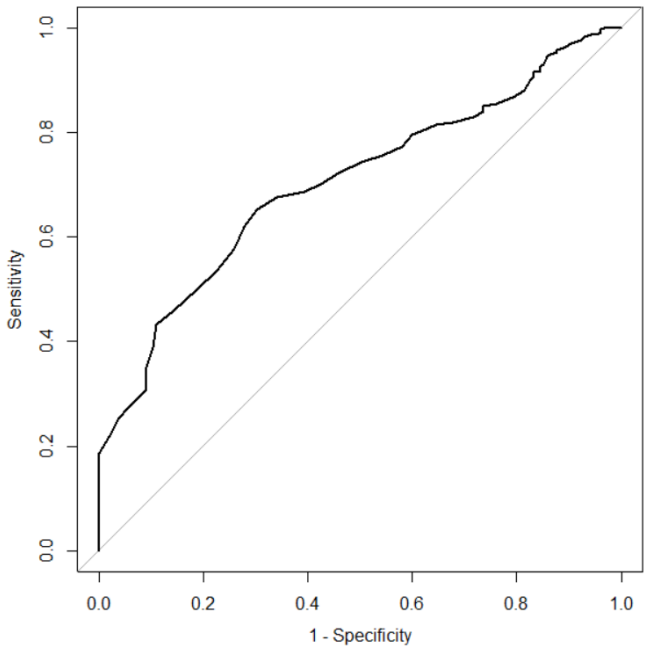


Figure 63: ROC Curve for Canada Model

Figure 64 shows the segments with observed derailments, as denoted in red. The red segments in **Figure 65** represent segments that were classified as 1 by the model (with derailment risk).

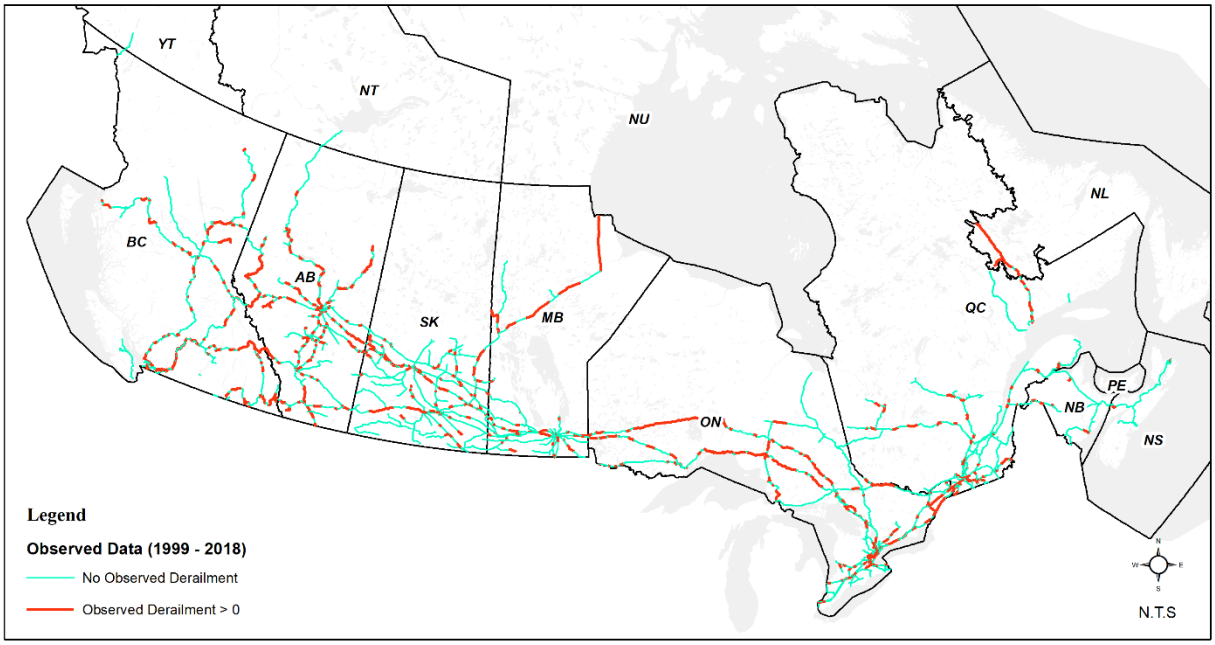


Figure 64: Track Segments with Observed Derailment(s)

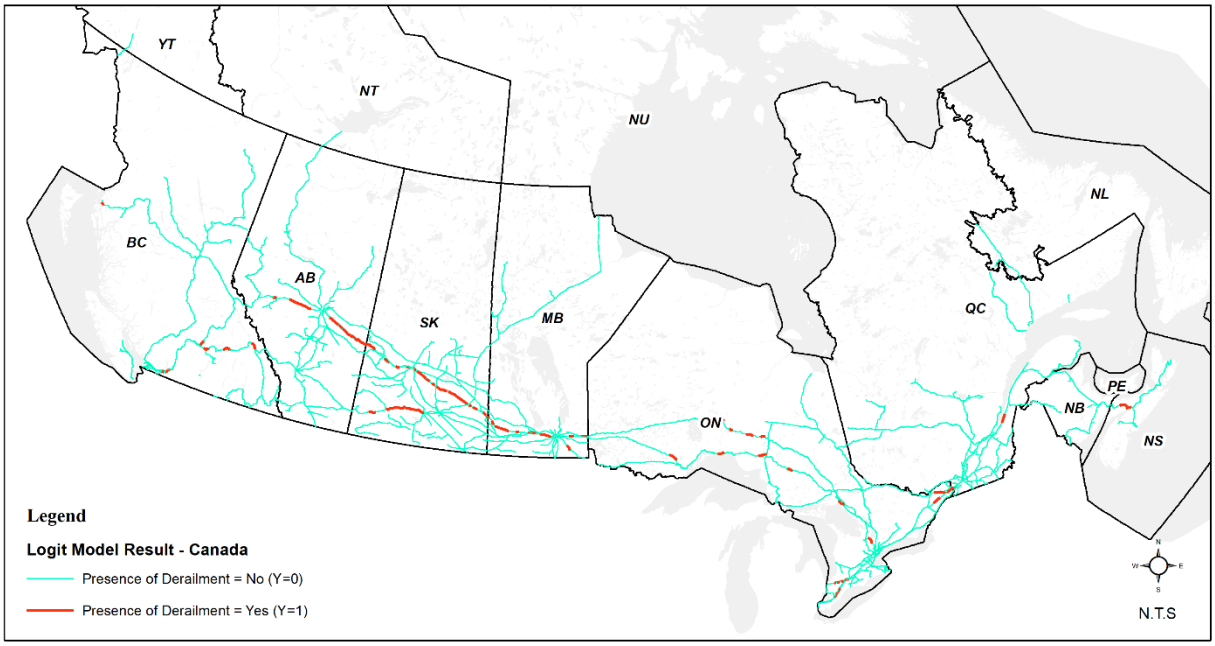


Figure 65: Track Segments Classified with Derailment Risk for Canada

Segments that were classified with derailment risk are mainly found in Western Canada. The continuous segments that traverse across Alberta and Manitoba were identified with positive risk. These are CN-owned tracks that are shared with VIA rail for passenger train services (e.g. Great Western Way route). These segments are associated with higher train speeds and daily train volumes.

7.5. Logit Model for Eastern Canada

In total, 70 candidate models were developed for predicting the probability of derailment on track segments in Eastern Canada.

7.5.1. Comparison of Logit Models for Eastern Canada

Table 43 compares the results of the GOF tests applied to the calibration data for the shortlisted Canada models. Model ID#35 was selected to be the best-fitting model. **Chapter 7.5.2** continues the discussion of the results.

Table 43: Comparison of Shortlisted Logit Models for Eastern Canada

Model ID	AIC	BIC	Null Dev	Res Dev	Null Dev – Res Dev (R^2_{Chi} Test)	R^2_{McF}	Model Accuracy
4	569.4	591.8	608.59	559.43	49.16	0.08	83.5%
7	564.64	587.00	608.59	554.64	53.95	0.09	83.5%
8	570.91	593.27	608.59	560.91	47.68	0.08	83.8%

Model ID	AIC	BIC	Null Dev	Res Dev	Null Dev – Res Dev (R^2_{Chi} Test)	R^2_{McF}	Model Accuracy
9	574.84	597.21	608.59	564.84	43.75	0.07	83.8%
32	568.87	595.70	608.59	556.87	51.72	0.08	83.5%
35	**553.11**	**579.94**	**608.59**	**541.11**	**67.48**	**0.11**	**82.7%**
36	559.88	586.72	608.59	547.88	60.71	0.10	83.1%
37	574.39	601.22	608.59	562.39	46.20	0.08	83.5%
68	571.86	603.16	608.59	557.86	50.73	0.08	83.8%
69	573.96	600.79	608.59	561.96	46.63	0.08	83.1%

7.5.2. Logit Model Results for Eastern Canada

This section presents the functional form of the selected model, the coefficient estimates, model statistics and considers the effects of independent variables on the outcome variable.

Equation 43 shows the model form of the selected model for Eastern Canada. This equation predicts the probability of derailment to occur on a segment for a 10-year period. The number of years is represented by *N* in the model form.

$$\begin{aligned}\text{logit}(p) = N \times [&(-8.6083) + (0.1532 \times \text{VL\_Count}) + (2.3206 \times \text{logVL\_TrnSpd}) \\ &+(-0.1039 \times \text{Stn\_Count}) + (0.0544 \times \text{Seg\_Length}) \\ &+(-1.1086 \times \log(\text{TrnSpd} \times \text{VL\_Count}))]\end{aligned} \quad \text{(Eq.43)}$$

Table 44 shows the coefficient estimates and model statistics.

Table 44: Logit Model Results for Eastern Canada

Variable	Estimate	Std. Error	z value	Pr(>|z|)	Model Statistics
(Intercept) ***	-8.608	1.328	-6.484	0.000	
VL_Count***	0.153	0.034	4.531	0.000	Prediction Accuracy: 0.827
log_VL_TrnSpd***	2.321	0.555	4.185	0.000	AIC: 553.11
Stn_Count	-0.104	0.155	-0.670	0.503	Null Deviance: 608.59
Seg_Length***	0.054	0.011	4.998	0.000	Residual Deviance: 608.59
Log (TrnSpd×VLCount)***	-1.109	0.297	-3.729	0.000	Observations: 925

****significance level of less than 0.0001*

Daily train volume, maximum train speed and segment length variables had positive effects on derailment while station count had a negative association with derailment. To quantify the estimated effects on the outcome variable (derailment), the log odds of each independent variable were converted to odd ratios by exponentiating the coefficient estimates. For instance, a unit increase in segment length changes the odds of derailment by a factor of exp (0.153) = 1.17, when all the other variables remain the same. An increase in number of stations on a segment reduces the odds of derailment by a factor of exp (-0.104) = 0.90, when all the other variables remain the same.

The interaction term between train speed and daily train volumes changed the direction of the effect of train speed from negative to positive allowing for a more intuitively acceptable interpretation of the sign of the model parameters. The introduction of this interaction term enhanced the Eastern Canada model's performance.

Chi-squared test was performed to assess the statistical significance of the coefficient estimates, excluding the interaction term. For this model, a chi-squared test statistic of 67.48, with five degrees of freedom and a p-value of < 0.0010 imply that the overall effects of the independent variables were statistically significant. The reduction in deviance also indicates that the inclusion of independent variables has improved the overall model fit.

7.5.3. Logit Model Validation for Eastern Canada

The selected model was further evaluated by performing additional GOF tests on the validation data. These tests included: prediction accuracy, classification tables, sensitivity and specificity rates, and ROC curve.

The prediction accuracy of the selected model was 82%. **Table 45** summarizes the number of true and false positives, as well as true and false negatives predicted by the model. The model statistics show that the selected model had a sensitivity rate of 13% and a specificity rate of 95%. Like the Canada model, the Eastern Canada model obtained a high prediction accuracy score, but it was biased towards predicting zero events.

Table 45: Classification Table for Eastern Canada Model

		Observed		**Model Statistics**
		0	**1**	Accuracy: 0.8165 95% CI: (0.766, 0.8602) Sensitivity: 0.1333 Specificity: 0.9485 Detection Rate: 0.7950
Prediction	**0**	221	39	
	1	12	6	

Figure 66 shows the ROC curve for the Eastern Canada model. The selected model had an AUC of 0.68 on which is within the acceptable range.

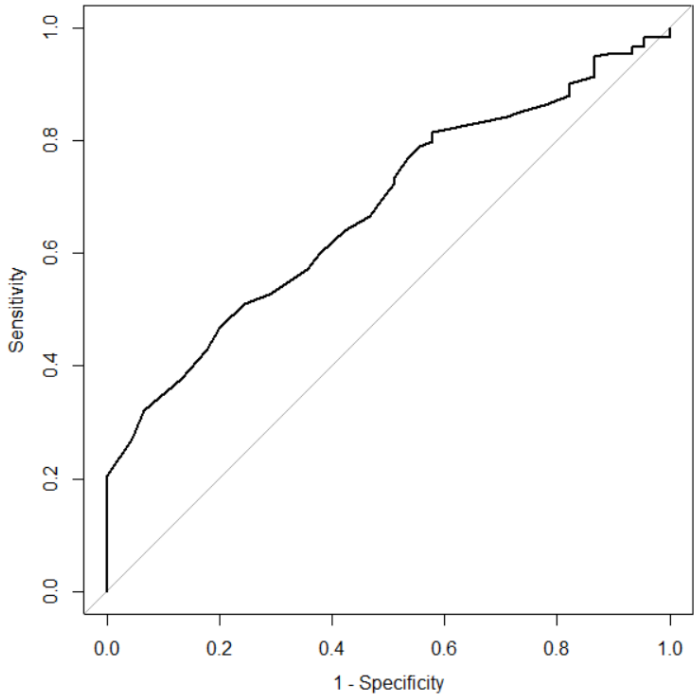


Figure 66: ROC Curve for Eastern Canada Model

Figure 67 shows the segments with observed derailments, as denoted in red. The red segments in **Figure 68** represent segments that were classified as 1 (with derailment risk) by the model. Many of these segments are located in Ontario which is consistent with the observed data.

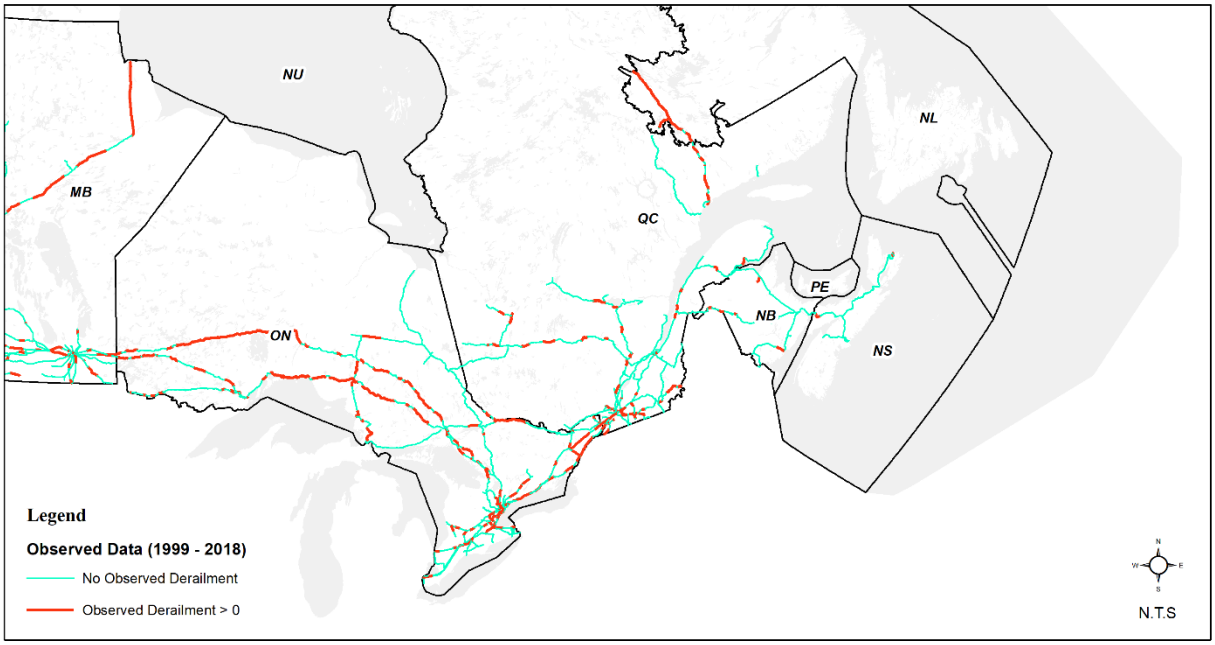


Figure 67: Track Segments with Observed Derailment(s) in Eastern Canada

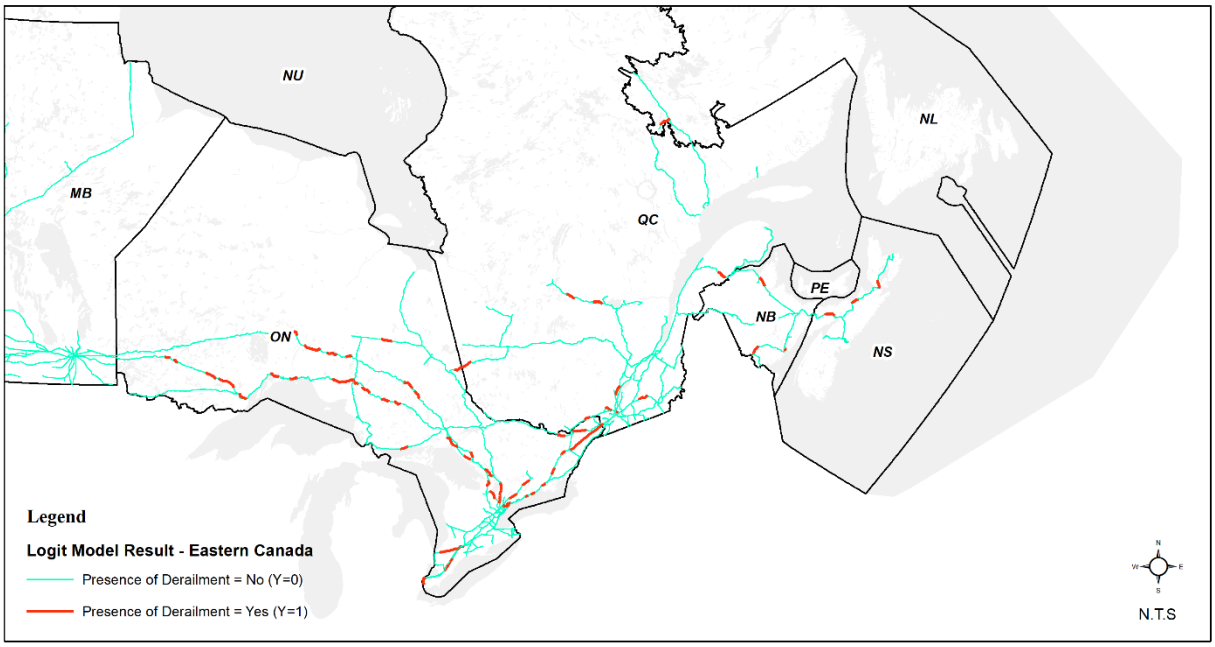


Figure 68: Track Segments Classified with Derailment Risk in Eastern Canada

7.6. Logit Model for Western Canada

In total, 72 candidate models were developed for predicting the probability of derailment on track segments in Western Canada. The candidate models were shortlisted based on intuitive signs of coefficients.

7.6.1. Comparison of Logit Models for Western Canada

Table 46 the results of the GOF tests applied to the calibration data for the shortlisted Western Canada models. Model ID#9 (shown in bold) was selected to be the best-fitting model. **Chapter 7.6.2** continues the discussion of the results.

Table 46: Comparison of Candiate Logit Models for Western Canada

Model ID	AIC	BIC	Null Dev	Res Dev	Null Dev – Res Dev (R^2_{Chi} Test)	R^2_{McF}	Model Accuracy
4	973.1	998.0	1077.00	963.09	113.91	0.12	78%
9	**967.15**	**992.08**	**1077.00**	**957.15**	**119.85**	**0.12**	**78%**
30	974.59	1004.51	1076.00	962.59	113.41	0.12	78%

7.6.2. Logit Model Results for Western Canada

This section presents the functional form of the selected model, the coefficient estimates, model statistics and considers the effects of independent variables on the outcome variable.

Equation 44 shows the model form of the selected model for Western Canada. This equation predicts the probability of derailment to occur on a segment for a 10-year period. The number of years is represented by *N* in the model form.

$$\text{logit}(p) = N \times [(-6.0987) + (0.0544 \times \text{VL\_Count}) + (0.0057 \times \text{VL\_TrnSpd}) + (-0.0604 \times \text{Stn\_Count}) + (0.6151 \times log\ Seg\_Length)] \quad \text{(Eq.44)}$$

Table 47 presents the coefficient estimates and model statistics.

Table 47: Logit Model Results for Western Canada

Variable	Estimate	Std. Error	z value	Pr(>|z|)	Model Statistics
(Intercept) ***	-6.0987	0.4765	-12.7993	0.0000	Prediction Accuracy: 0.78 AIC: 967.15 Null Deviance: 1077.00 Residual Deviance: 957.15 Observations: 1,447
VL_Count	0.0544	0.0162	3.3513	0.0008	
VL_TrnSpd***	0.0057	0.0081	0.7033	0.4818	
Stn_Count	-0.0604	0.1496	-0.4035	0.6866	
log_Seg_Length***	0.6151	0.1361	4.5210	0.0000	

**** significance level of less than 0.0001*

Like the Eastern Canada model, daily train volume, maximum train speed and segment length variables had positive effects on derailment while station count had a negative association with derailment. To better understand the effects of the independent variables on the probability of derailment, odd ratios were calculated by exponentiating the coefficient estimates. For example, one unit increase in daily train volume changes the odds of derailment by a factor of exp (0.0544) = 1.06, when all the other variables remain the same.

Chi-squared test was performed to assess the statistical significance of the coefficient estimates. A Chi-squared test statistic of 119.85, with four degrees of freedom and a p-value of < 0.0010 indicates that the overall effects of the independent variables were statistically significant. The reduction in deviance implies that the inclusion of independent variables has improved the overall model fit.

7.6.3. Logit Model Validation for Western Canada

The selected model was further evaluated by performing additional GOF tests on the validation data. These tests included: prediction accuracy, classification tables, sensitivity and specificity rates, and ROC curve.

The prediction accuracy of the selected model was 79%. **Table 48** summarizes the number of true and false positives, as well as true and false negatives predicted by the model. The model statistics show that the selected model had a sensitivity rate of 11% and a specificity rate of 98%. Similar to the previous models, the model for Western Canada was biased towards predicting zero events.

Table 48: Classification Table for Western Canada Model

		Observed		Model Statistics
		0	**1**	Accuracy: 0.7909 95% CI: (0.7511, 0.8271) Sensitivity: 0.1111 Specificity: 0.9753 Detection Rate: 0.7596
Prediction	**0**	356	88	
	1	9	11	

Figure 69 shows the ROC curve for the Western Canada model. The model had a strong classification ability with an AUC of 0.75.

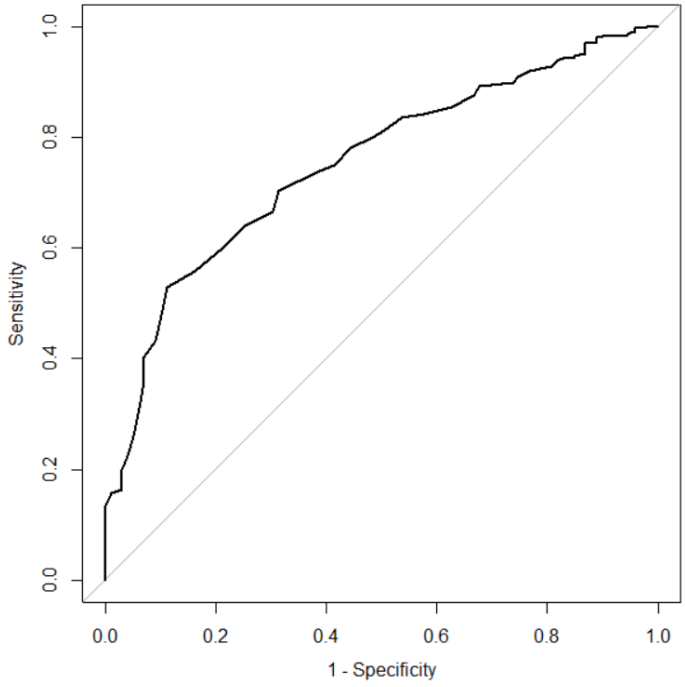


Figure 69: ROC Curve for Western Canada Model

Figure 70 shows the segments with observed derailments, as denoted in red. The red segments in **Figure 71** represent segments that were classified as 1 by the model (with derailment risk). The Western Canada model showed some similar prediction results as the Canada model. A number of continuous segments were classified with derailment risk across Alberta and Manitoba. These segments are associated with high train speeds and daily train volumes, as noted in **Chapter 7.4.3**.

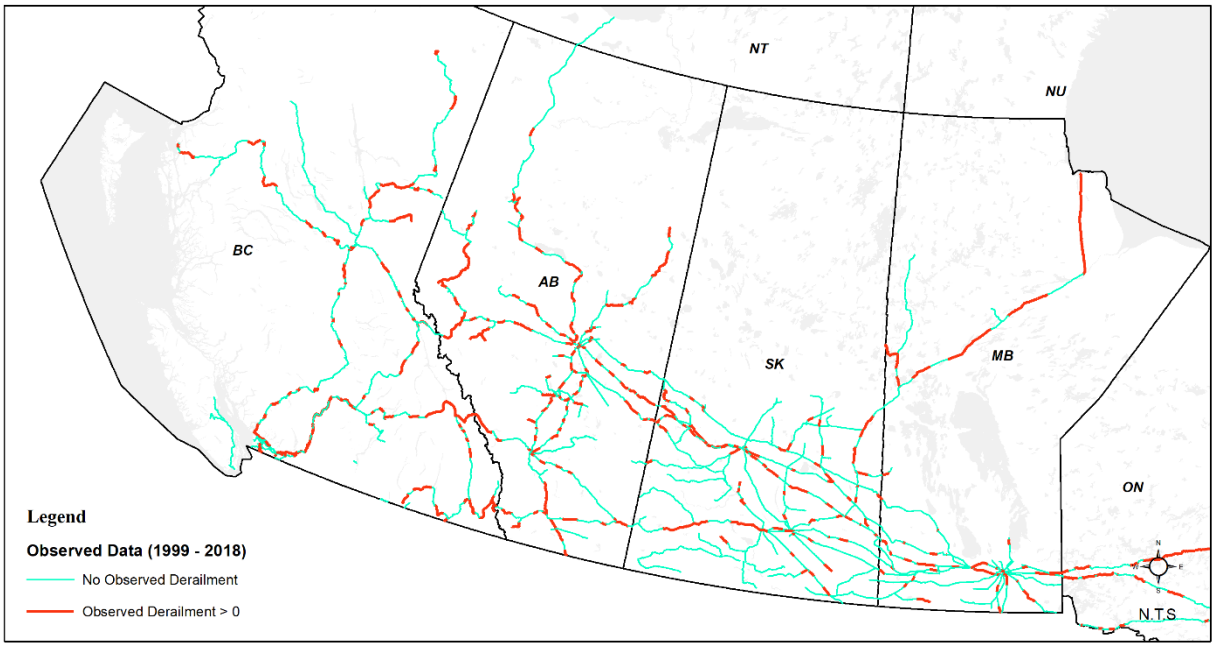


Figure 70: Track Segments with Observed Derailment(s) in Western Canada

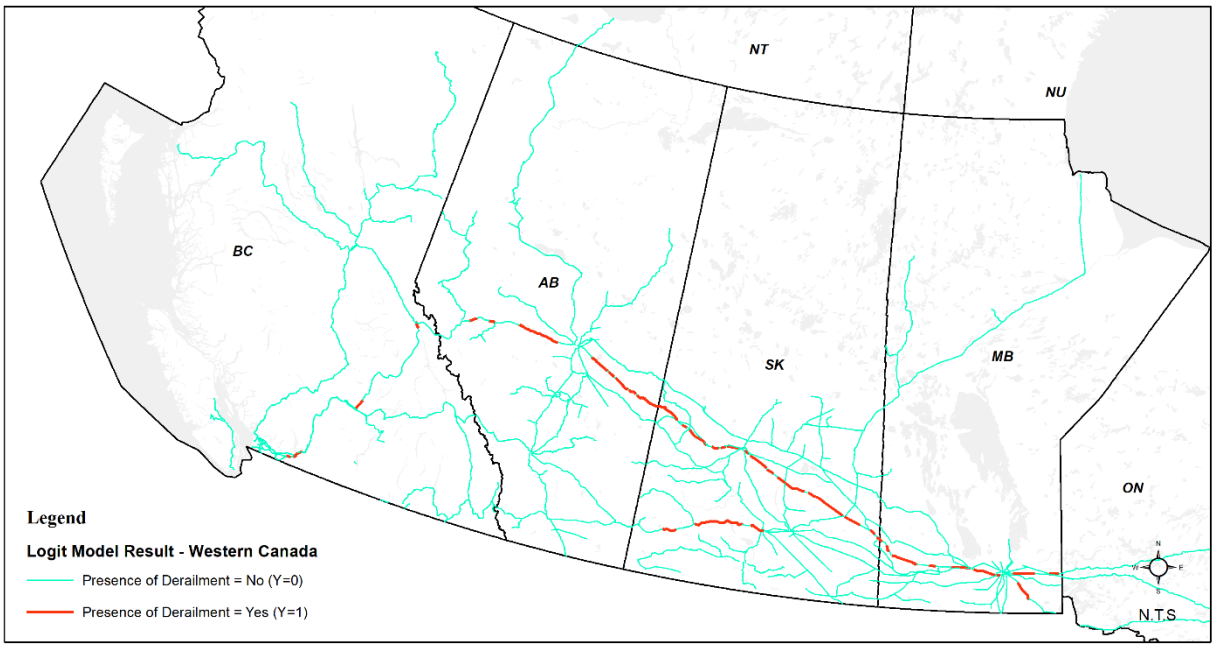


Figure 71: Track Segments Classified with Derailment Risk in Western Canada

7.7. Logit Model for Canadian National Railway

In total, 72 candidate models were developed for predicting the probability of derailment on CN-owned tracks.

7.7.1. Comparison of Logit Models for Canadian National Railway

Table 49 compares the results of the GOF tests applied to the calibration data for the shortlisted CN models. Model ID#36 (shown in bold) was selected as the best-fitting model. **Chapter 7.7.2** continues the discussion of the results.

Table 49: Comparison of Candidate Logit Models for CN

Model ID	AIC	BIC	Null Dev	Res Dev	Null Dev – Res Dev (R^2_{Chi} Test)	R^2_{McF}	Model Accuracy
4	924.10	948.36	940.00	914.10	**25.90**	0.09	80%
9	922.82	947.07	940.00	912.82	27.18	0.09	80%
35	897.11	926.22	939.00	885.11	53.89	0.12	80%
36	**896.50**	**925.61**	939.00	884.50	**54.50**	**0.12**	**80%**
37	924.74	953.84	939.00	912.74	26.26	0.10	80%
49	939.83	964.09	940.00	929.83	10.17	0.08	79%
51	914.45	938.71	940.00	904.45	35.55	0.10	79%
67	920.90	950.01	939.00	908.90	30.10	0.10	80%
68	920.82	954.78	938.00	906.82	31.18	0.10	79%
69	920.90	950.01	939.00	908.90	30.10	0.10	80%
71	925.99	959.94	938.00	911.99	26.01	0.10	80%
72	926.04	955.15	939.00	914.04	24.96	0.09	80%

7.7.2. Logit Model Results for Canadian National Model

This section presents the functional form of the selected model, the coefficient estimates, model statistics and considers the effects of independent variables on the outcome variable. **Equation 45** shows the model form of the best-fitting model for CN-owned tracks. This equation predicts the probability of derailment to occur on a segment for a 10-year period. The number of years is represented by *N* in the model form.

$$\begin{aligned}\text{logit}(p) = N \times [&(-12.1344) + (0.0377 \times \text{VL\_Count}) + (2.6140 \times \text{logVL\_TrnSpd}) \\ &+ (0.3016 \times \text{Stn\_Count}) + (0.4294 \times log\ \text{Seg\_Length}) \\ &+ (-0.5473 \times \log(\text{VL\_Trn\_Spd} \times \text{VL\_Count}))]\end{aligned} \quad \text{(Eq.45)}$$

Table 50 presents the coefficient estimates and model statistics.

Table 50: Logit Model Results for CN

Variable	Estimate	Std. Error	z value	Pr(>|z|)	Model Statistics
(Intercept) ***	-12.1344	1.1440	-10.6068	0.0000	
VL_Count	0.0377	0.0202	1.8658	0.0621	Prediction Accuracy: 0.80
log (VL_TrnSpd) ***	2.6140	0.4536	5.7625	0.0000	AIC: 896.50
Stn_Count***	0.3016	0.0763	3.9535	0.0001	Null Deviance: 939 Residual Deviance: 884.50
log (Seg_Length) ***	0.4294	0.1202	3.5721	0.0004	Observations: 1,332
log (VL_TrnSpd×VLCount) **	-0.5473	0.2109	-2.5956	0.0094	

***significance level of less than 0.001*

****significance level of less than 0.0001*

Table 50 shows that all independent variables have positive effects on the probability of derailment. In terms of effects, one unit increase in daily train volume changes the odds of derailment by a factor of exp (0.0377) = 1.06, when all the other variables remain the same. The interaction term between train speed and train volume was included to obtain the intuitive sign of coefficient estimate for the train speed variable. The introduction of this interaction term enhanced the CN model's performance.

A Chi-squared test was performed to assess the statistical significance of the coefficient estimates. A Chi-squared test statistic of 54.50, with five degrees of freedom and a p-value of < 0.001 indicates that the overall effects of independent variables were statistically significant. The reduction in deviance implies that the inclusion of independent variables has improved the overall model fit.

7.7.3. Logit Model Validation for Canadian National Railway

The selected model was further evaluated by performing additional GOF tests on the validation data. These tests included: prediction accuracy, classification tables, sensitivity and specificity rates, and ROC curve.

The prediction accuracy of the selected model was 81%. **Table 51** summarizes the number of true and false positives, as well as true and false negatives predicted by the model. The model statistics show that the selected model had a sensitivity rate of 18% and a specificity rate of 96%. Like the previous models, the CN model was biased towards predicting zero events.

Table 51: Classification Table for CN model

		Observed		**Model Statistics**
		0	**1**	Accuracy: 0.8062 95% CI: (0.7632, 0.8444) Sensitivity: 0.1842 Specificity: 0.9582 Detection Rate: 0.7700
Prediction	**0**	298	62	
	1	13	14	

Figure 72 shows the ROC curve for CN model. The CN model had an AUC of 0.73 which suggests a strong classification ability.

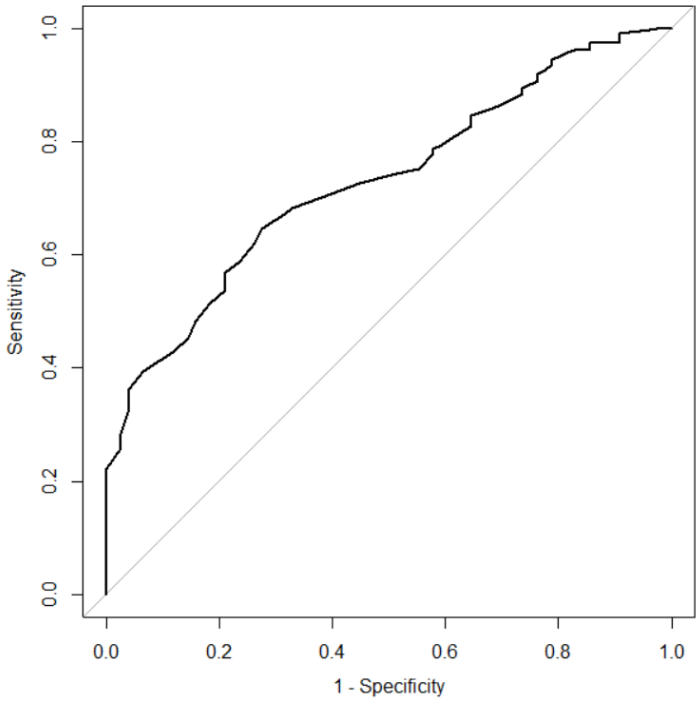


Figure 72: ROC Curve for Eastern Canada Model

Figure 73 shows the segments with observed derailments, as denoted in red. The red segments in **Figure 74** represent segments that were classified as 1 by the model (with derailment risk). Majority of the segments that were classified with derailment risk are located in three provinces: Alberta, Saskatchewan and Ontario.

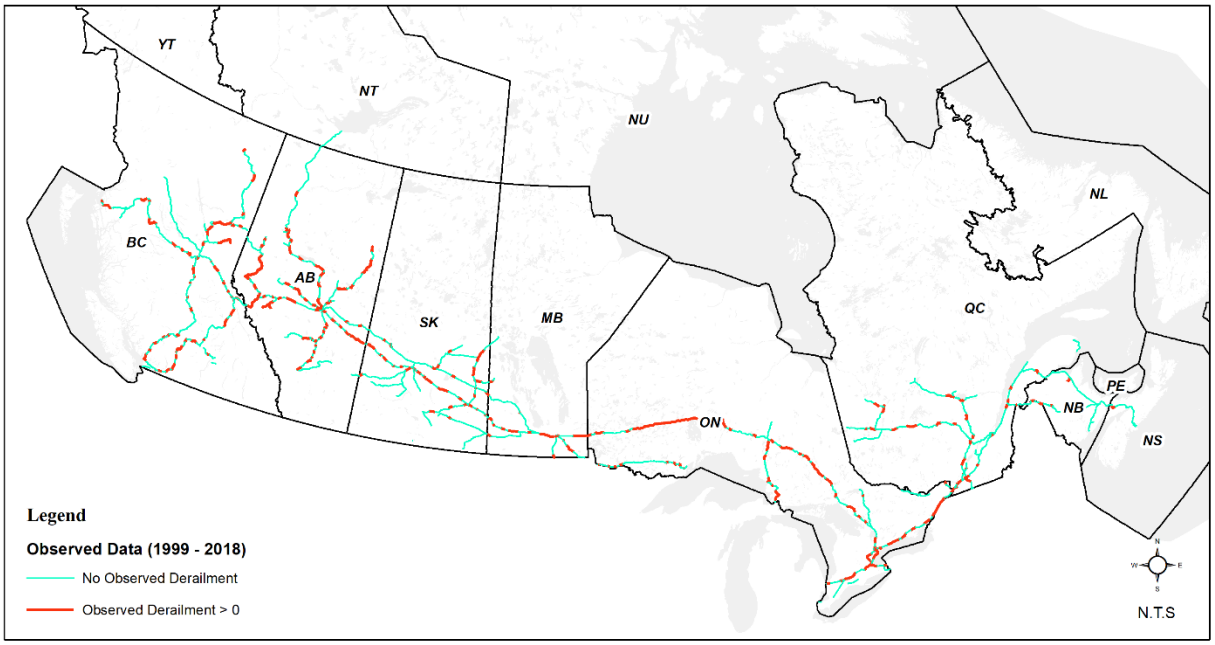


Figure 73: Track Segments with Observed Derailment(s) on CN Railway Network

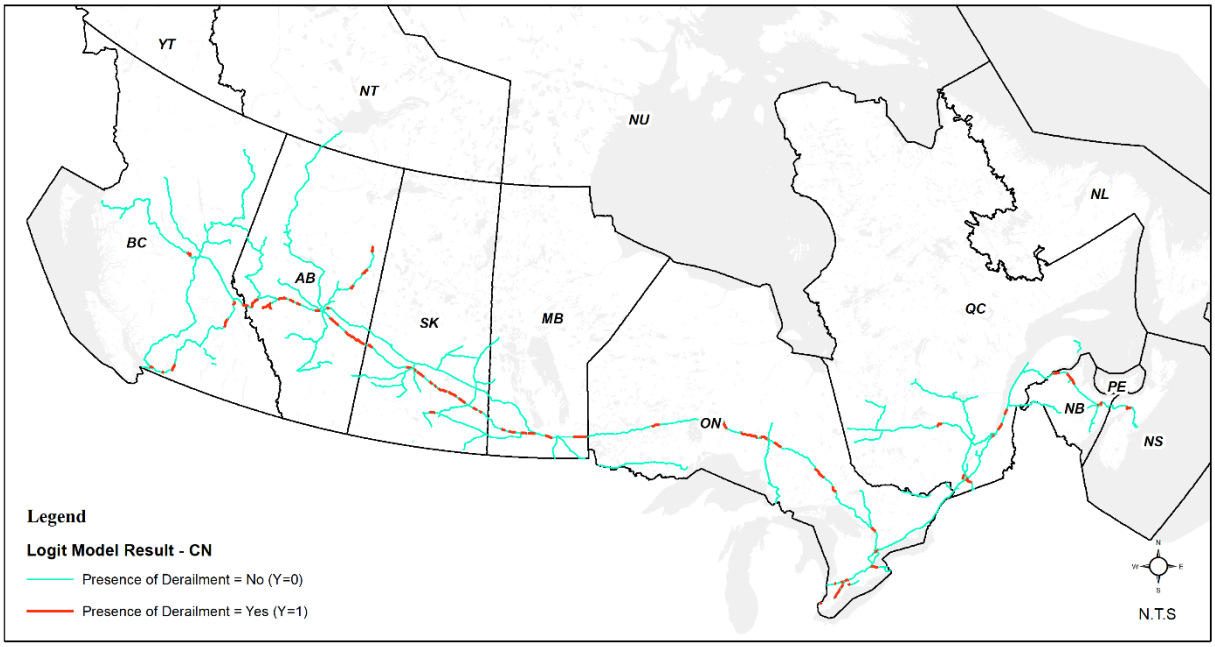


Figure 74: Track Segments Classified with Derailment Risk on CN Railway Network

7.8. Logit Model for Canadian Pacific Railway

In total, 80 candidate models were developed for predicting the probability of derailment on CP-owned tracks.

7.8.1. Comparison of Logit Models for Canadian Pacific Railway

Table 52 compares the results of the GOF tests applied to the calibration data for the shortlisted CN models. Model ID#34 (shown in bold) was selected to be the best-fitting model. **Chapter 7.8.2** continues the discussion of the results.

Table 52: Comparison of Candidate Logit Models for CP

Model ID	AIC	BIC	Null Dev	Res Dev	Null Dev – Res Dev (R^2_{Chi} Test)	R^2_{McF}	Model Accuracy
32	646.11	673.07	654.00	634.11	19.89	0.106767	73%
34	**632.28**	**659.23**	**654.00**	**620.28**	**33.72**	**0.126253**	**73%**
49	649.37	671.83	655.00	639.37	15.63	0.099362	72%
50	637.30	659.76	655.00	627.30	27.70	0.116366	71%
58	644.56	667.02	655.00	634.56	20.44	0.106138	73%
65	650.95	677.90	654.00	638.95	15.05	0.099956	72%
67	645.76	672.71	654.00	633.76	20.24	0.107261	72%
68	619.94	646.89	654.00	607.94	46.06	0.143633	70%

7.8.2. Logit Model Results for Canadian Pacific Railway

This section presents the functional form of the selected model, the coefficient estimates, model statistics and considers the effects of independent variables on the outcome variable.

Equation 46 shows model form of the selected model for CP-owned tracks. This equation predicts the probability of derailment to occur on a segment for a 10-year period. The number of years is represented by *N* in the model form.

$$\begin{aligned}\text{logit}(p) = N \times [&(-11.5166) + (0.0749 \times \text{VL\_Count}) + (2.6696 \times \text{VL\_TrnSpd}) \\ &+(0.0922 \times \text{Stn\_Count}) + (0.0670 \times \text{Seg\_Length}) \\ &+(-0.6502 \times log(\text{VL\_TrnSpd} \times \text{VL\_Count}))]\end{aligned} \quad \text{(Eq.46)}$$

Table 53 presents the coefficient estimates and model statistics.

Table 53: Logit Model Results for CP

Variable	Estimate	Std. Error	z value	Pr(>|z|)	Model Statistics
(Intercept) ***	-11.5166	1.6176	-7.1197	0.0000	
VL_Count*	0.0749	0.0295	2.5417	0.0110	Prediction Accuracy: 0.73
log_VL_TrnSpd***	2.6696	0.7323	3.6455	0.0003	AIC: 632.28
Stn_Count	0.0922	0.1387	0.6651	0.5060	Null Deviance: 654.00
Seg_Length***	0.0670	0.0117	5.7441	0.0000	Residual Deviance: 620.28
log_TrnSpd_VLCount*	-0.6502	0.3176	-2.0473	0.0406	Observations: 940

**significance level of less than 0.01*
****significance level of less than 0.0001*

Table 53 shows that all independent variables have positive effects on the probability of derailment. In terms of effects, a unit increase in segment length changes the odds of derailment vs. no derailment by a factor of exp (0.0670) = 1.07, when all the other variables remain the same. The interaction term between train speed and train volume was included to obtain the intuitive sign of coefficient estimate for the train speed variable. The introduction of this interaction term also enhanced the CP model's performance.

Chi-squared test was performed to assess the statistical significance of the coefficient estimates, excluding the interaction term. A Chi-squared test statistic of 33.72, with five degrees of freedom and a p-value of < 0.0010 indicates that the overall effects of the independent variables were statistically significant. The reduction in deviance suggests that the inclusion of independent variables has improved the model fit.

7.8.3. Logit Model Validation for Canadian Pacific Railway

The selected model was further evaluated by performing additional GOF tests on the validation data. These tests included: prediction accuracy, classification tables, sensitivity and specificity rates, and ROC curve.

The prediction accuracy of the selected model was 75%. **Table 54** summarizes the number of true and false positives, as well as true and false negatives predicted by the model. The model statistics show that the selected model had a sensitivity rate of 14% and a specificity rate of 99%. Similar to previous models, the CP model also showed a bias in predicting zero events.

Table 54: Classification Table for CP model

		Observed		**Model Statistics**
		0	**1**	Accuracy: 0.7536 95% CI: (0.6988, 0.8029) Sensitivity: 0.1429 Specificity: 0.9852 Detection Rate: 0.7143
Prediction	**0**	200	66	
	1	3	11	

Figure 75 presents the ROC curve for the CP model. The selected model had an AUC of 0.73 which implies a strong classification ability.

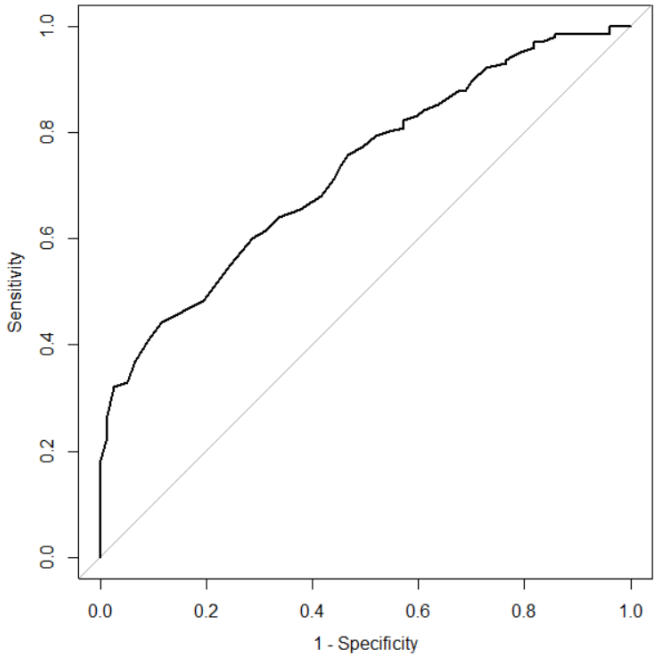


Figure 75: ROC Curve for CP Model

Figure 76 shows the segments with observed derailments, as denoted in red. The red segments in **Figure 77** represent segments that were classified as 1 by the model (with derailment risk). Segments that were classified with derailment risk are associated with high train speeds (> 40 mph) and longer segment lengths (> 20 km).

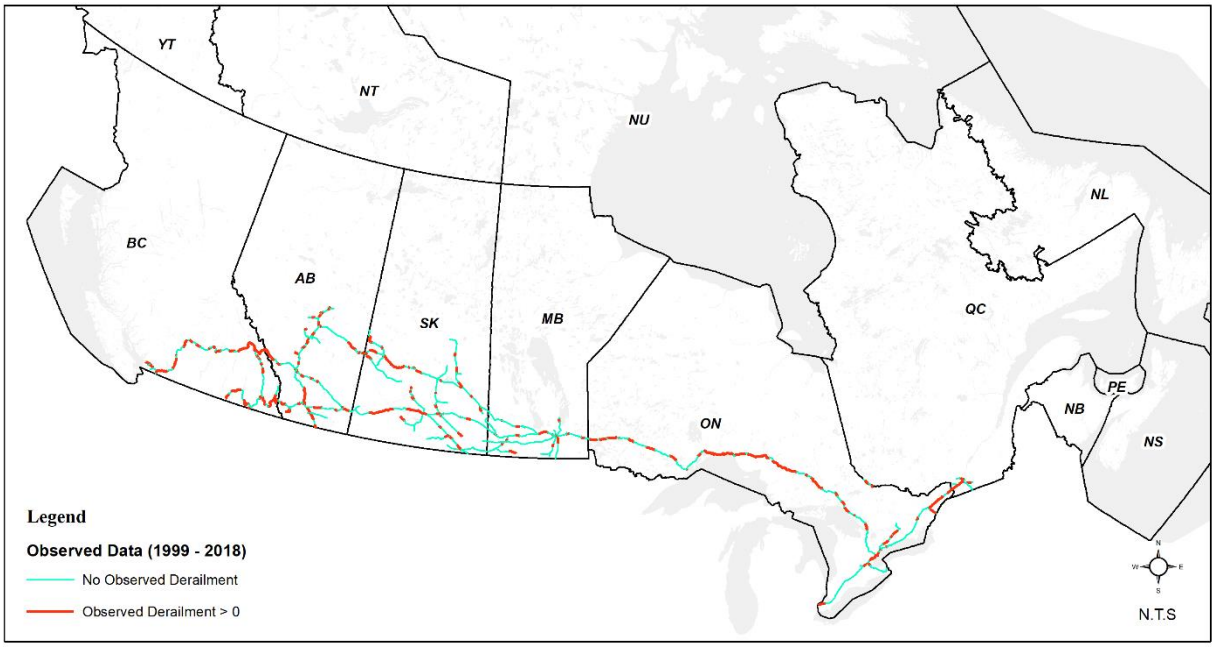


Figure 76: Track Segments with Observed Derailment(s) on CP Railway Network

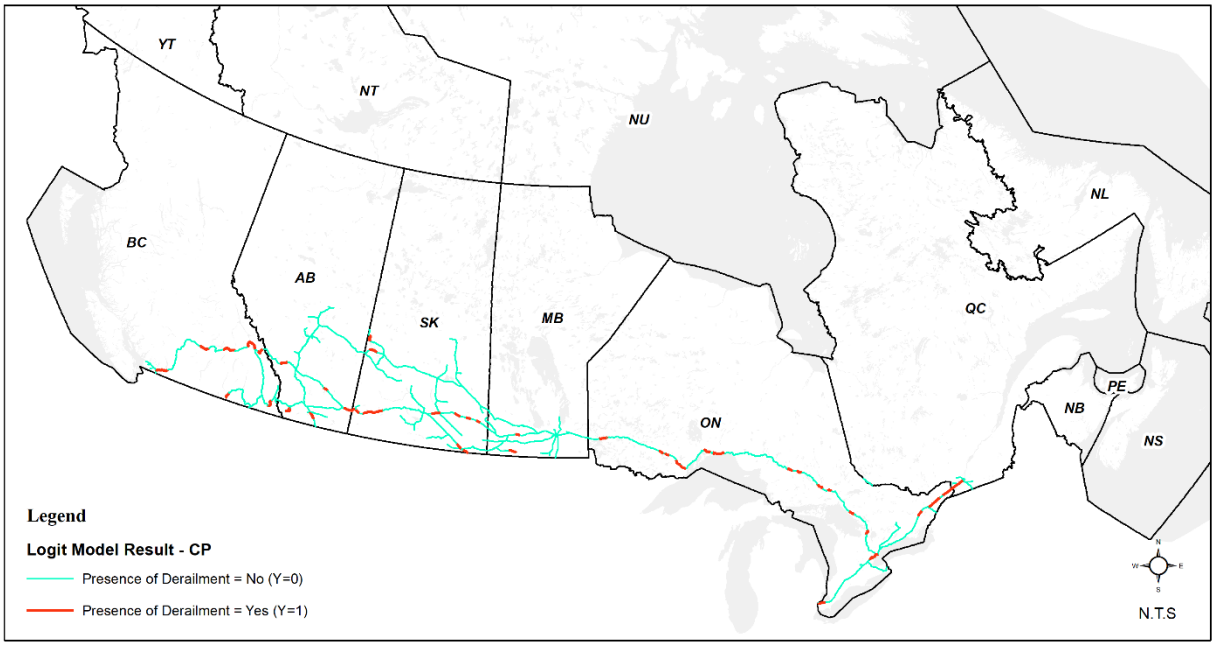


Figure 77: Track Segments Classified with Derailment Risk for CP Model

7.9. Chapter Summary

This Chapter considered the development and results of the five prediction models using logistic regression: Canada, Eastern Canada, Western Canada, Canadian National Railway (CN), and Canadian Pacific Railway (CP). Five variables were used to predict the number of derailments: daily train volume, maximum train

speed, segment length, average train volume and station count. For all models, segments that were classified as 1 (with derailment risk) using a presumed cut-off value of 0.5. These segments were compared with observed data to evaluate the overall predictive capability of the models.

For each of the five models, the shortlisted models were compared using a set of GOF tests to determine the best-fitting models. The effects of independent variables on derailment potential was interpreted as odd ratios. All independent variables showed positive effects on derailments for almost all models except for Canada, CN and CP models. In Eastern Canada and Western Canada models, the station count variable had a negative effect on derailments. The inclusion of the interaction terms between train speed and train volumes helped to produce parameters with intuitively acceptable signs and to improve overall model performance for the Eastern Canada, CN and CP models.

Overall, all selected models produced good prediction accuracy scores in the range of 73% and 83%. The classification tables provided more information on the models' predictive capabilities by summarizing the number of correct and incorrect predictions. All models were biased towards predicting true negatives. As such, the five selected models yielded high specificity rates (95%-99%) and low sensitivity rates (10%-18%). In terms of classification ability, all models demonstrated strong classification abilities with AUCs between 0.68 and 0.75.

Chapter 8 compares the results and overall performance of the logistic regression and negative binomial models in derailment prediction.

CHAPTER 8. COMPARISON OF PREDICTION RESULTS

Chapter 5 discussed the results of conventional NB distribution modelling, and **Chapter 6** discussed the results of logistic regression modelling. The derailment dataset consisted of approximately 79% of zero observations. Treating derailments as discrete counts, the conventional NB distribution method consistently produced ranking results that were as adequate as the rankings obtained from the observed data, but the conventional NB models underestimated the number of derailments for most of the segments. The logistic regression modelling technique was better able to handle over-representation of zero events in the data, but this technique has its limitations. This Chapter compares the derailment prediction performance of the two approaches in greater detail.

8.1. Tetrachoric Correlation Analysis

A tetrachoric correlation analysis was conducted to compare the different modelling approaches. Tetrachoric correlation (Carroll, 1961) is regarded as a useful technique for analyzing the relationship between pairs of dichotomous variables. For the purpose of a correlation analysis, the predicted and expected numbers of derailments in the NB models were converted to binary values where μ_i>1, 1 and $E[y_i] > 1, 1$, otherwise 0.

Table 55 shows the results. Strong correlations (> 0.80) are bolded. The expected number of derailments was highly correlated with the observed number of derailments (0.80-0.87) for all five models. The EB technique produced high correlations with the logit model results for Eastern Canada (0.80) and moderate correlations with the logit model results for Canada, Western Canada, CN and CP models (0.67, 0.55, 0.64 and 0.78 respectively). The number of predicted derailments in the CP model showed a very strong correlation (0.97) with the logit model results.

Table 55: Tetrachoric Correlation Analysis Results

	Models				
Comparison	**CA**	**East CA**	**West CA**	**CN**	**CP**
Obs vs. NB Predicted	0.49	0.57	0.27	0.49	0.47
Obs vs. NB Expected	**0.89**	**0.83**	**0.87**	**0.81**	**0.86**
Obs vs. Logit	0.35	0.46	0.20	0.42	0.47
NB Predicted vs. Logit	**0.85**	**0.86**	**0.86**	0.72	**0.97**
NB Expected vs. Logit	0.67	**0.80**	0.55	0.64	0.78

Another key finding was the logit models' poor to moderate correlations (0.20-0.47) with the observed number of derailments. These weak results are likely attributable to the over-representation of

false negatives (segments with observed derailments falsely classified as 0 in the models). The predicted values obtained from the NB models showed moderate correlations with the observed data (0.47-0.57) for all the models except Western Canada (correlation of only 0.27). The results of the tetrachoric correlation analysis showed that the EB method demonstrated the strongest ability to predict derailment risks.

8.2. Comparison of Network Screening Results

Table 56 presents selected statistics of segments with predicted derailment risk from the five models. The statistics show the proportion of segments (in terms of length) that was identified by the models, as having the greatest derailment risk. These include the top 10 segments in the NB models (EB technique) and segments that were classified with positive derailment risk (classified as 1) in the logit models.

The network screening process conducted using the EB technique found that the top 10 segments with the greatest risk of derailment accounted for 1%-2% of the total track length in the network. For example, in the Canada model, the total length of track with observed derailment risks made up 29% of the Canada's rail network (or 12,976 km of segments). The EB estimates reduced the percentage to 1% (or 334 km). The results of the logit model were similar. In the Canada model, the logit model reduced the percentage of segments with derailment risk to 5% (or 2,075 km).

Table 56: Selected Statistics of Track Segments with Predicted Derailments Risks

	Models				
Segment Statistics by Modelling Methods	**CA**	**East CA**	**West CA**	**CN**	**CP**
a. Total length (km) for top 10 segments from NB models	334	332	319	340	305
b. Total segment length (km) with observed derailments	12,976	7,667	5,309	6,607	4,418
c. Total segment length (km) in the network	45,513	26,477	19,036	22,407	12,486
d. "a" as a percentage of "c"	1%	1%	2%	2%	2%
e. "b" as a percentage of "c"	29%	29%	28%	29%	35%
f. Total length (km) for (+) risk segments from logit models	2,075	2,106	1,480	2,035	1,631
g. Total Segment length (km) with observed derailments	12,976	7,667	5,309	6,607	4,418
h. Total Segment length (km) in the network	45,513	26,477	19,036	22,407	12,486
i. "f" as a percentage of "h"	5%	8%	8%	9%	13%
j. "g" as a percentage of "h"	29%	29%	28%	29%	35%

The network screening results help Transport Canada (TC) and/or track owners by identifying segments with greatest derailment risk and reducing the number of segments needing safety improvement to a manageable number. The EB technique also allows prioritization of the segments at risk which helps TC and/or track owners to identify and prioritize segments by their expected safety performance and perform inspection or maintenance accordingly.

8.3. Segment Identification Methods

The networking screening methods discussed above consisted of using NB or logit model outcomes directly. For NB models, network screening ranked the top 10 segments to identify hotspots that warrant attention. For logit models, segments were classified as 1's when the probability of derailments was greater than a threshold cut-off value (e.g., a cut-off value of 0.5).

A third approach to networking screening involved considering the results from both NB and logit models simultaneously.

To accomplish this, the expected numbers of derailments in the NB models were converted to binary values: if $E[y_i] \geq 1, 1$, otherwise 0. **Appendix E** presents maps of the results. Segments that had a value of 1 from both the NB and the logit models were flagged as 'key segments of concern.' **Appendix F** presents maps of these segments. The total length of these key segments was then compared to the total length of the rail network.

Table 57 shows selected statistics for the segments flagged as at risk for derailments by both the NB and the logit models. Key segments of concern accounted for 0.3% to 6% of the rail network. Western Canada had the lowest percentage of key segments of concern on its network. This result was expected given the weak correlation between NB and logit models discussed in **Chapter 8.1**.

Table 57: Selected Statistics of Segments with Derailments Risks from NB and Logit Models

	Models				
Segment Statistics by Modelling Methods	**CA**	**East CA**	**West CA**	**CN**	**CP**
a. Total length (km) for (+) risk segments from both models	794	344	62	472	694
b. Total Segment Length (km) with observed derailments	12,976	7,667	5,309	6,607	4,418
c. Total Segment length (km) in the network	45,513	26,477	19,036	22,407	12,486
d. "a" as a percentage of "c"	2%	1%	0.3%	2%	6%
e. "b" as a percentage of "c"	29%	29%	28%	29%	35%

Table 58 summarizes the network screening results of the three different identification methods. The percentages of key segments of concern identified by the third method are similar to the percentages identified by NB model.

Table 58: Summary of Selected Statistics for Difference Segment Identification Methods

Segment Identification Methods	Models				
	CA	**East CA**	**West CA**	**CN**	**CP**
a. Total length (km) for top 10 segments from NB models	334	332	319	340	305
b. Total length (km) for (+) risk segments from logit models	2,075	2,106	1,480	2,035	1,631
c. Total length (km) for (+) risk segments from both models	794	344	62	472	694
d. Total Segment length (km) in the network	45,513	26,477	19,036	22,407	12,486
e. "a" as a percentage of "d"	1%	1%	2%	2%	2%
f. "b" as a percentage of "d"	5%	8%	8%	9%	13%
g. "c" as a percentage of "d"	2%	1%	0.3%	2%	6%

8.4. Discussion of Model Performances

Both the NB and the logit models yielded useful network screening results. Both sets of models were able to reduce the rail network to a reasonable and more manageable scale by identifying segments with potential for safety improvement.

The logit regression models were expected to be more sensitive in handling derailment database given the excessive number of zero observations. The logit models developed in this study showed good classification ability and good prediction accuracies, but they did not outperform the NB models when the results are compared to observed counts. This was evident in the tetrachoric correlation analysis discussed in **Chapter 8.1**.

In essence, the results of binomial logit models greatly depended on the choice of cut-off values for classification. To illustrate the effect of cut-off values, a sensitivity test was undertaken on the Canada model. Different cut-off values were applied to the validation dataset to derive the numbers of true positives and negatives predicted from the model.

Table 59 summarizes the model statistics from the sensitivity test. An increase in true positives was associated with lower cut-off values and vice versa. No segments were being classified as 1's (with 0% sensitivity) when a high cut-off value (e.g., 0.8) was chosen. The high accuracy score of 79% could be misleading as the model was biased towards classifying negatives only. A low cut-off value (e.g., 0.2) compromised overall model adequacy as the sensitivity and specificity were close to 50/50 and 50/50 implies random classification. The results demonstrate that model accuracy is not a valid performance measure when the data classification is skewed.

Table 59: Comparison of Model Statistics Based on Cut-off Values

Cut-off Value	True Positives	True Negatives	Accuracy	Sensitivity	Specificity
0.2	105	385	67%	68%	66%
0.3	52	474	71%	34%	82%
0.5	16	561	78%	10%	97%
0.8	0	581	79%	0%	100%

The sensitivity test results in **Table 59** show that logit models are vulnerable to the choice of cut-off value as the cut-off value has significant effects on the predictions. A pre-defined cut-off value of 0.5 was used in this study. This value could, of course, be different on the decision maker's circumstances and objectives.

Optimizing the classifiers could involve difficult choices in a safety analysis. For example, a safety analysis might prefer to lower the cut-off value and prioritize the minimizing of false negatives rather than miss a segment with actual derailment risks. Such decisions have to take into account the fact that the economic costs incurred by over-representation of segments falsely classified as segments of risk can be enormous. In addition, track owners and safety jurisdictions would be spending significant resources to conduct site inspection or perform maintenance that might be unnecessary. Conversely, one can argue that the price for high false negatives is too high if the model fails to detect segments with a potential risk of derailment. It is important to bear in mind the consequences of positive and false negative trade-offs in safety risk prediction.

The results may lead to the conclusion that logit models are more sensitive than NB models in predicting zero events, but this outcome was the result of compromising true positives. The same degree of bias would be present with any 'rare event' datasets. In the context of network screening, the finding might be acceptable as the approach adopted is able to reduce the network to locations likely to be key to safety management. However, it is also possible to argue that many segments with a risk of derailments are then overlooked which is clearly not desirable.

The comparison of the two NB and logit models showed that NB models had an advantage when predicting the estimated number of derailments to prioritize segments for safety analysis. Using NB models together with the EB technique produced more rigorous results with strong correlations between the EB results and the observed data for all five selected models. Given these results, the study concluded that the NB models were the preferred approach to evaluating derailment risks.

8.5. Chapter Summary

This chapter compared the performance of negative binomial (NB) and logistic regression models. The tetrachoric correlation analysis showed that EB method demonstrated the strongest correlation with observed data and the logit models demonstrated the weakest correlation.

The network screening results for the NB and logit models helped identify segments that pose the greatest derailment risks. The top 10 segments with the highest risk of derailments as identified by the NB models' results and the network screening accounted for 1% to 2% of the rail network. The logit models identified 5% to 13% of the rail network as key segments of concern.

A third network screening approach that synthesized the results of the NB and logit models was undertaken. By converting the expected number of derailments to binary outcomes, segments that were denoted as 1's (with derailment risks) in both models were identified as key segments of concern. This approach reduced the segments that required the most attention for safety improvement to 0.3% to 6% of the total network.

The limitations of the logit model were also discussed. Although logit models demonstrated higher specificity rates (true negatives), the poor prediction of true positives meant that many segments with derailment risk were not being detected. This bias could lead to misleading conclusions regarding the safety performance of the rail network. In addition, the results depended on the cut-off values selected and the choice of cut-off is particularly difficult and controversial in the context of safety evaluation.

In conclusion, the advantage of the NB modelling method was its ability to rank segments by expected safety performance. Prioritization could be of value to rail operators as it is unlikely that all site inspection and maintenance can be performed at once. Being able to prioritize segments by the risk of derailments would allow the operators to allocate resources and determine appropriate timelines for conducting the necessary work. The ranking advantage of the NB approach together with consideration of other aspects of the two modelling techniques led to the NB modelling method being considered the more suitable approach in derailment prediction.

CHAPTER 9. CONCLUSIONS AND RECOMMENDATION

There is a wealth of literature on derailment risk analysis in the United States, but similar topics have not been extensively explored in Canada. Researchers have been constrained by restricted access to proprietary information which allowed access only to aggregated that limited a study's ability to evaluating the level of safety in detail. This thesis has presented a methodological framework for developing segment-level derailment prediction models. We believe that this approach can capture variation in track characteristics and changes in traffic exposure on a rail network and effectively determine the level of safety for each segment.

Before model development, the study spent significant effort on data quality control and validation in order to synthesize spatial data with historical derailment information. Some challenges were identified in the data integration process, which included: rail network conversion, varying track classifications, and discrepancies in spatial information. These challenges were resolved using data analysis and geoprocessing tools in GIS. This resulted in an integrated database that contained all segments in the Canada's rail network and key risk attributes.

This study applied two types of regression model for developing derailment prediction models. A series of derailment prediction models were developed to predict the number of derailments. Models were developed for three geographic areas (Canada, Eastern and Western Canada) and two railway companies (Canadian National (CN) and Canadian Pacific Railways (CP)).

The first set of five prediction models was developed using negative binomial (NB) distribution. The maximum daily train traffic, maximum train speed, segment length, average daily train volumes and number of stations on a segment variables were found to correlate with the number of derailments and were retained in the model development process. The introduction of interaction terms enhanced the overall fitting performance of the CN and CP models. Risk factors associated with derailment may include terrain, severity of curvature, train weight (tonnage), and train length.

To account for possible regression-to-the-mean bias, Empirical Bayes (EB) technique was applied to estimate the expected number of derailments for each model. The top 10 segments were identified through a network screening process based on the expected number of derailments. These hotspots accounted for approximately 1% to 2% of Canada's rail network. Twelve segments were identified as hotspots by multiple models.

The second set of prediction models was developed using logistic regression distribution. Binomial logit models were developed for the same geographic areas and companies. The results were assessed on

the basis of various goodness-of-fit tests (such as deviance statistics, McFadden R-Squared, prediction accuracy, classification tables, sensitivity and specificity rates and receiver operating characteristics curve). Segments that were classified with derailment risks by the models accounted for 5% to 13% of the rail network.

The NB and logit model performances were then compared using tetrachoric correlation analysis and the network screening results to evaluate the advantages and disadvantages of the two modelling approaches. The logit models presented several limitations which makes them less suitable than the NB models for derailment prediction. These model limitations largely arose from rail safety problems being statistically 'rare' events and therefore intrinsically difficult to model.

In conclusion, the EB technique provided the greatest accuracy in predicting derailment risks. The EB results were the closest to the observed data. EB estimates can also be used to prioritize track segments by predicted safety performance which is useful for TC and/or track owners as they perform their investigation and implement mitigation measures.

9.1. Research Contribution

This thesis makes a major contribution to advancing research into predicting and evaluating derailment risks in Canada. While recognizing the constraints of having to use aggregated data in past studies, this study provides pioneering research into the development of *segment-level* derailment prediction models. The methodology used in this study adopted a network screening process that resulted in the ranking of segments by safety performance. Network screening reduces a rail network to a manageable scale for governing agencies and/or track owners who wish to investigate segments that have the greatest derailment risk. The screening allows organizations to make more strategic decisions on rail safety improvement and to optimize resource allocation. As such, network screening can result in significant economic savings for governing agencies and railway companies through accident prevention.

The models in this study can be used to determine whether the observed number of derailments is higher or lower than the average safety performance of similar segments. When a network screening process identifies key segments of safety concern, decision-makers can consider current and new mitigation measures. New measures may include emerging technology solutions. For instance, Automatic Data Extraction (ADE) using LiDAR can be implemented along selected segments. ADE is capable of engineering grade mapping of rail infrastructure and also capable of detecting rail track conditions and performing assessments (Sohn et al., 2019).

Another major contribution of this study is the comprehensive review of negative binomial and logistic regression modelling methods in predicting the number of derailments. The results of the comparison of the two approaches provide great insights into the evaluation of derailment risks and the advantages and disadvantages of the two methods.

The analysis and findings in this study have a number of policy implications for rail safety management in Canada. **Figure 78** illustrates how the findings of this research can be implemented in the current rail safety management system. A derailment prediction model can serve as the basis for systematically evaluating rail safety as part of a safety management program. Transport Canada's (2018b) Rail Safety Act review stated that one of the main goals of its safety regime is to develop better capacity in data analytics (including predictive analytics). The advanced statistical analyses and findings of this study provide valuable insights and new knowledge for Canada's rail safety program.

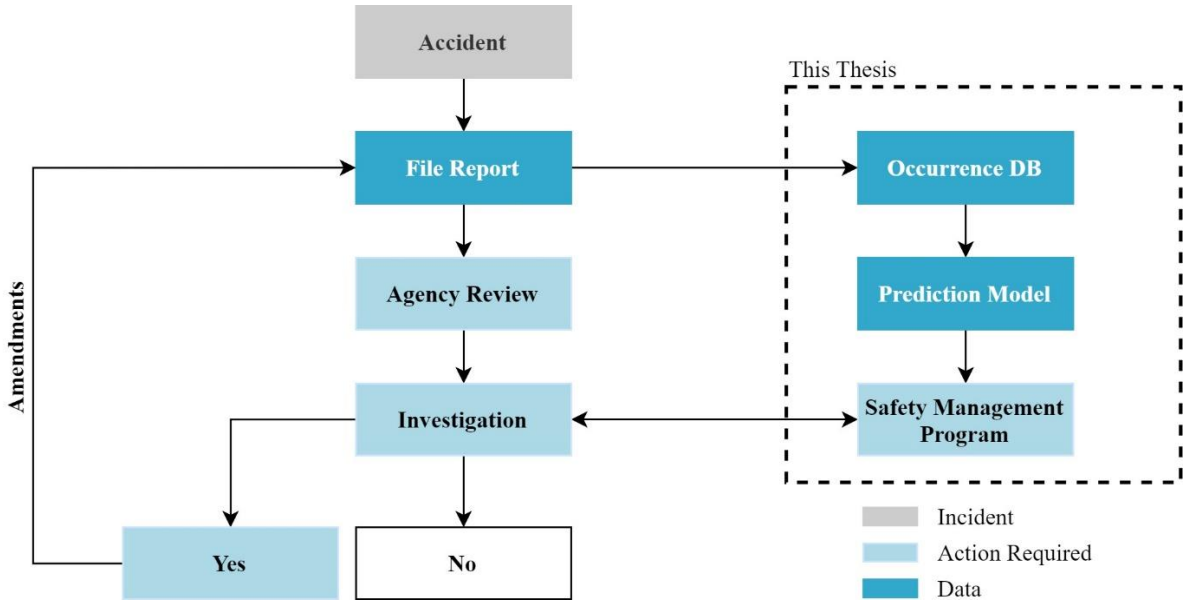


Figure 78: Utilization of Research Findings

9.2. Research Limitations

This study should be considered in light of several limitations. The first and foremost limitation is the lack of explanatory data. The incomplete information in the derailment database limited the number of independent variables that could be included in the modelling process. It is also possible that various confounding factors might contribute to derailments, but such factors were not considered. With these issues in mind, it is recognized that the models developed in this study do not provide a comprehensive list of variables that may be associated with derailments.

The data published by the Transportation Safety Board of Canada (TSB) was originally provided to the TSB by third-party sources. The data in the rail occurrence database system (RODS), for example, was transcribed from accident reports submitted by rail operators. As only a small portion of reportable accidents warrant detailed investigations, much of the information available is not validated.

The last limitation is the lack of prior research on similar topics in Canada. Many of the lessons-learned from the literature review were based on studies conducted in the US. An increase in the number of rail safety research studies in Canada would help address some of the challenges and unknowns identified in this study. This limitation amplifies the need to advance research in the area of rail safety in Canada.

9.3. Future Study

The analytical approach and findings in this study can be used to guide future research and investment of resources to examine and evaluate rail safety, particularly in derailments. Recommendations for future study to advance this topic are discussed below.

Further analysis could be conducted to determine the most appropriate track segmentation method for evaluating rail safety. Different segmentation methods, including those discussed in this research, could be tested and compared to improve understanding of their effects on rail prediction models. The findings could inform practitioners developing best practices for assessing the safety performance of rail infrastructure.

If available, more detailed information should be requested and obtained from governing agencies and/or railway companies to fill data gaps in the derailment database. It is recommended that additional segment-level data should be requested for enhancing the performance of future derailment prediction models. Better data and more complete data can provide opportunities for improving existing models if such data can be made available. Possibly important factors affecting the number of derailments may include terrain (elevation), severity of curvature, track construction year, track materials, etc. and require investigation.

More robust modelling techniques such as zero-inflated models can be explored in future studies. For example, no rail safety studies have explored the use of zero-inflated modelling. Following Poisson or negative binomial distributions, these models assume a dual-state process that estimates the probability of a zero or non-zero accident state of an entity (Chin and Quddus, 2003). This approach corrects the underestimation of probabilities due to non-zero events.

Lastly, the application of machine-learning (ML) techniques to derailment prediction can also be explored in the future. After being trained on a given dataset, ML techniques have the ability to make prediction using inputs outside of the observed datasets (Faria, 2018). The techniques constantly learn from data and make improvements in the prediction models, providing better productivity and efficiency. Since RODS is being continuously updated, a ML environment could be an effective way to process new data and improve the algorithm on an ongoing basis. It should be noted that, as ML algorithms are built under the assumption of balanced class distribution (Choi, et al., 2019), careful consideration should be given to preprocessing an imbalanced dataset before applying a ML approach.

While there are many exciting possibilities for future research, the findings in this study sets the stage for further studies in rail safety. The results of this study and the research topics and studies suggested for the future will provide significant benefits to the rail industry and to the general public.

REFERENCES

Allison, P. D. (2014, March). Measures of fit for logistic regression. In Proceedings of the SAS Global Forum 2014 Conference (pp. 1-13).

Al-Ghamdi, A. S. (2002). Using logistic regression to estimate the influence of accident factors on accident severity. Accident Analysis & Prevention, 34(6), 729-741.

Agresti, A. (2007). *An introduction to categorical data analysis, Second Edition*. New York: Wiley.

Anderson, R. T., and Barkan, C. P. L., Derailment Probability Analyses and Modelling of Mainline Freight Trains, Proceedings of the 8th International Heavy Haul Conference, pp. 491–497, 2005

Anderson, R. T., and Barkan, C. P. L., Railroad Accident Rates for Use in Transportation Risk Analysis. Transportation Research Record: Journal of the Transportation Research Board, 1863(1), pp. 88–98, 2007

Bagheri, M., Saccomanno, F., Chenouri, S., and Fu, L., Reducing the Threat of In-Transit Derailments Involving Dangerous Goods through Effective Placement along the Train Consist, Accident Analysis and Prevention, 43(3), pp. 613–620, 2011

Barkan, C., Tyler Dick, C., and Anderson, R. (2003). Railroad derailment factors affecting hazardous materials transportation risk. Transportation Research Record: Journal of the Transportation Research Board, (1825), 64-74.

Bibel, G., Santos, C. G., Rechkemmer, A., Bahouth, G., Dias Guichot, Y., Garcia-Vera, M. P., and Schulman, C., Disaster Complexity and the Santiago de Compostela Train Derailment, Disaster Health, 3(1), pp. 11–31, 2016

Brimley, B. K., Saito, M., and Schultz, G. G. (2012). Calibration of Highway Safety Manual safety performance function: development of new models for rural two-lane two-way highways. Transportation research record, 2279(1), 82-89.

Britton, M. A., Asnaashari, S., and Read, G. J. M., Analysis of Train Derailment Cause and Outcome in Victoria, Australia, between 2007 and 2013: Implications for Regulation, Journal of Transportation Safety and Security, 9(1), pp. 45–63, 2017

Cafiso, S., and Di Silvestro, G. (2011). Performance of safety indicators in identification of black spots on two-lane rural roads. Transportation research record, 2237(1), 78-87.

Cafiso, S., D'Agostino, C., and Persaud, B. (2018). Investigating the influence of segmentation in estimating safety performance functions for roadway sections. Journal of traffic and transportation engineering (English edition), 5(2), 129-136.

Carroll, J. B. (1961). The nature of the data, or how to choose a correlation coefficient. Psychometrika, 26(4), 347-372.

Chen, Z., and Fan, W. D. (2019). A multinomial logit model of pedestrian-vehicle crash severity in North Carolina. International Journal of Transportation Science and Technology, 8(1), 43-52.

Chin, H. C., and Quddus, M. A. (2003). Modelling count data with excess zeroes: an empirical application to traffic accidents. Sociological methods and research, 32(1), 90-116.

Choi, J., Gu, B., Chin, S., and Lee, J. S. (2019). Machine learning predictive model based on national data for fatal accidents of construction workers. Automation in Construction, 110, 102974.

Cook, D., Dixon P., Duckworth W. M., Kaiser M. S., Koehler, K., Meeker, W.Q. and W. R. Stephenson, R. (2000). *Beyond traditional statistical methods.* Chapter 3: Binary Response and Logistic Regression Analysis.

Crocco, F., De Marco, S., and Mongelli, D. W. E. (2010). An integrated approach for studying the safety of road networks: logistic regression models between traffic accident occurrence and behavioural, environmental and infrastructure parameters. WIT Transactions on Ecology and the Environment, 142, 525-536.

Evans, A. W. (2007). Rail safety and rail privatisation in Britain. Accident Analysis and Prevention, 39(3), 510-523.

Evans, A. W. (2011). Fatal train accidents on Europe's railways: 1980–2009. Accident Analysis & Prevention, 43(1), 391-401.

Faria, J. M. (2018, February). Machine learning safety: An overview. In Proceedings of the 26th Safety-Critical Systems Symposium, York, UK.

Farid, A., Abdel-Aty, M., and Lee, J. (2018). Transferring and calibrating safety performance functions among multiple states. Accident Analysis & Prevention, 117, 276-287.

Federal Highway Administration: Highway Safety Improvement Program Manual (2019). Website: https://safety.fhwa.dot.gov/hsip/resources/fhwasa09029/sec2.cfm

Federal Highway Administration: U.S. Track Classification Quick Reference (2019). Website: http://www.jgmes.com/webstart/library/table_fra_track.htm

Federal Railroad Ad ministration Office of Safety Analysis: Ten Year Accident/Incident Overview (2019). Retrieved from https://safetydata.fra.dot.gov/OfficeofSafety/publicsite/Query/TenYearAccidentIncidentOverview.aspx

Fienberg, S., 1980. The Analysis of Cross-Classified Categorical Data, 2nd ed. The MIT Press, Cambridge, Massachusetts.

Freeman, M. F., and Tukey, J. W. (1950). Transformations related to the angular and the square root. The Annals of Mathematical Statistics, 607-611.

Glickman, T. S., and Rosenfield, D. B. (1984). Risks of catastrophic derailments involving the release of hazardous materials. Management Science, 30(4), 503-511.

Goodman, I., Rowan, B. (2013). Analysis of the Potential Costs of Accidents/Spills Related to Crude by Rail. The Goodman Group, Ltd. Docket No. PHMSA-2012-0082 (HM-251). Retrieved from http://www.thegoodman.com/pdf/TGG20131108_OCIetal_PotentialsCostsCBR.pdf

Hauer, E. (2004). The harm done by tests of significance. Accident Analysis & Prevention, 36(3), 495-500.

Hauer, E. (2015). The art of regression modelling in road safety (Vol. 38). New York: Springer.

Hosmer D.W. and Lemeshow S. (1980) "A goodness-of-fit test for the multiple logistic regression model." Communications in Statistics A10:1043-1069.
Hu, S. R., Li, C. S., and Lee, C. K. (2010). Investigation of key factors for accident severity at railroad grade crossings by using a logit model. Safety Science, 48(2), 186-194.

Hu, S. R., Li, C. S., & Lee, C. K. (2010). Investigation of key factors for accident severity at railroad grade crossings by using a logit model. Safety science, 48(2), 186-194.

Khan, I., Lee, E., and Khan, M. (2018). Developing a highway rail grade crossing accident probability prediction model: a North Dakota case study. Safety, 4(2), 22.

Kish, A., and Clark, D. W., Track Buckling Derailment Prevention through Risk-Based Train Speed Reductions, Proceedings of the AREMA 2009 Annual Conference, 2009

Li, W., Roscoe, G. S., Zhang, Z., Saat, M. R., and Barkan, C. P. (2018). Quantitative Analysis of the Derailment Characteristics of Loaded and Empty Unit Trains. Transportation Research Record, 2672(10), 156-165.

Liu, X., Saat, M. R., and Barkan, C. P. (2012). Analysis of causes of major train derailment and their effect on accident rates. *Transportation Research Record*, *2289*(1), 154-163.

Liu, X., Saat, M. R., and Barkan, C. P. (2013). Integrated risk reduction framework to improve railway hazardous materials transportation safety. Journal of hazardous materials, 260, 131-140.

Lu, J., Haleem, K., Alluri, P., Gan, A., and Liu, K. (2014). Developing local safety performance functions versus calculating calibration factors for SafetyAnalyst applications: A Florida case study. Safety science, 65, 93-105.

Liu, X. (2015). Statistical temporal analysis of freight train derailment rates in the United States: 2000 to 2012. Transportation Research Record: Journal of the Transportation Research Board, (2476), 119-125.

Liu, X. (2016). Collision risk for freight trains in the United States. Transportation Research Record: Journal of the Transportation Research Board, (2546), 121-128.

Liu, X., & Rodriguez, D. F. (2017). Statistical Comparison of Train Accident Rates: Methodology and Decision Support Tool (No. 17-00854).

Liu, X., Saat, M. R., and Barkan, C. P. (2017). Freight-train derailment rates for railroad safety and risk analysis. Accident Analysis & Prevention, 98, 1-9.

Lyon, C., Persaud, B., and Gross, F. (2016). The calibrator: An SPF calibration and assessment tool user guide. Report No. FHWA-SA-17-016, FHWA, US Department of Transportation.

Schafer, D., and Barkan, C. (2008). Relationship between train length and accident causes and rates. Transportation Research Record: Journal of the Transportation Research Board, (2043), 73-82.

Shmueli, G. (2010). To explain or to predict?. Statistical Science, 25(3), 289-310.

Harirforoush, H., and Bellalite, L. (2016). A new integrated GIS-based analysis to detect hotspots: a case study of the city of Sherbrooke. Accident Analysis & Prevention.

Manual, H. S. (2010). American association of state highway and transportation officials (AASHTO). Washington, DC, 10.

Miaou, S. P., and Lum, H. (1993). Modelling vehicle accidents and highway geometric design relationships. Accident Analysis & Prevention, 25(6), 689-709.

Moynihan, T. W., and English, G. W. (2007). Railway safety technologies. Research and Traffic Group.

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

National Research Council (US). Transportation Research Board. Task Force on Development of the Highway Safety Manual, & Transportation Officials. Joint Task Force on the Highway Safety Manual. (2010). Highway safety manual (Vol. 1). AASHTO.

Natural Resources of Canada (2012). National Railway Network Demo Presentation Slides. Retrieved from http://ftp.maps.canada.ca/pub/nrcan_rncan/vector/geobase_nrwn_rfn/doc/Publication/NRWN_demo_20120116.pdf

Natural Resources Canada. (2019). National Railway Network (NRWN) GeoBase Series [Data file]. Retrieved from http://ftp.maps.canada.ca/pub/nrcan_rncan/vector/geobase_nrwn_rfn/

Niveditha, V., Ramesh, A., & Kumar, M. (2015). Development of models for crash prediction and collision estimation-a case study for Hyderabad City. International journal of transportation engineering, 3(2), 143-150.

Wellner, A. (2011). GIS Based Highway Safety Analysis for South Dakota Rural Highways. South Dakota State University.

Ogle, J. H., Alluri, P., and Sarasua, W. A. (2011). Model Minimum Uniform Crash Criteria and Minimum Inventory Roadway Elements: Role of Segmentation in Safety Analysis (No. 11-4156).

Park, P. Y. et al. (2019). Development of a Railway Network Screening Tool to Prevent Train Derailments (Working Paper). Collaborative Research and Development for Improved Rail Safety in Canada project.

Portugués E. G. (2018). *Lab notes for Statistics for Social Sciences II: Multivariate Techniques (v12.3)*. Chapter 4.7 Deviance and Model Fit.

Quin, X., Wellner, A., 2012. Segment Length Impact on Highway Safety Screening Analysis. TRB, Washington DC.

R Core Team., R: A language and environment for statistical computing. R Foundation for Statistical Computing. Vienna, Austria, 2018, Retrieved from URL www.R-project.org

Sayeed, M. A., and Shahin, M. A., Investigation into Impact of Train Speed for Behavior of Ballasted Railway Track Foundations. Procedia Engineering, 143, pp. 1152–1159, 2016

Schafer, D., and Barkan, C., Relationship Between Train Length and Accident Causes and Rates. Transportation Research Record: Journal of the Transportation Research Board, 2043, 2008, pp. 73–82.

Simončič, M. (2001). Road accidents in Slovenia involving a pedestrian, cyclist or motorcyclist and a car. Accident Analysis & Prevention, 33(2), 147-156.

Snow, A., Coleman, N., and Rojik Orest, C. G. (2013), Guidelines for New Development in Proximity to Railway Operations, Retrieved from https://www.proximityissues.ca/wp-content/uploads/2017/09/2013_05_29_Guidelines_NewDevelopment_E.pdf

Sohn, G., Shabazi, M., Park, P., & Asgary, A. (2019). Advanced Rail Infrastructure Mapping Technologies for Train Derailment Mitigation (Poster). In Transportation Association of Canada and ITS Canada 2019 Joint Conference and Exhibition.

Srinivasan.R and Bauer.K (2013), *Safety Performance Function Development Guide: Developing Jurisdiction Specific SPFs.* FHWA-SA-14-005. United States. Federal Highway Administration. Office of Safety.

Stapleton, S. Y., Ingle, A. J., Chakraborty, M., Gates, T. J., and Savolainen, P. T. (2018). Safety Performance Functions for Rural Two-Lane County Road Segments. Transportation Research Record, 2672(52), 226-237.

Statistics Canada. Table 23-10-0045-01 Railway industry operating and income accounts, by mainline companies (x 1,000) Transport Canada, Rail Safety: Oversight and Expertise, Strategic Plan 2010-2015, 2010

Statistics Canada (2019). Table 23-10-0276-01 Weekly rail system performance indicators, by commodities, Transport Canada

Sun, X., & Garber, N. J. (2002). Determining the Safety Effects of Differential Speed Limits on Rural Interstate Highways Using Empirical Bayes Method (No. UVACTS-14-5-36). Center for Transportation Studies at the University of Virginia.

Takyi, E. A., Oluwajana, S. D., & Park, P. Y. (2018). Development of macro-level crime and collision prediction models to support data-driven approach to crime and traffic safety (DDACTS). Transportation research record, 2672(33), 56-66.

Transport Canada (2018a), Canadian Rail Operating Rules

Transport Canada (2018b), The 2018 Railway Safety Act Review

Transport Canada (2018c). Grade Crossing Inventory [Data file]. Retrieved from https://open.canada.ca/data/en/dataset/d0f54727-6c0b-4e5a-aa04-ea1463cf9f4c

Transportation Safety Board of Canada (2013): Railway Investigation Report R13D0054. Accessed December 15, 2019.

Transportation Safety Board of Canada (2019): Rail transportation occurrences in 2018. Retrieved from http://www.bst-tsb.gc.ca/eng/stats/rail/2018/sser-ssro-2018.html

Treichel, T. T., and Barkan, C. P. L. (1993). Working paper on mainline freight train accident Rates. *Unpublished Report to the Association of American Railroads*.

Wang, W., and Li, G. (2012), Development of High-Speed Railway Vehicle Derailment Simulation – Part I: A New Wheel/Rail Contact Method using the Vehicle/Rail Coupled Model, Engineering Failure Analysis, 24, pp. 77–92

Wang, B., Barkan, C. P. L., and Saat, M. R. (2017). Principal factors contributing to heavy haul freight train safety improvements in North America: a quantitative analysis. *Rail TEC Faculty and Student Papers and Presentations IHHA 2017 Cape Town 2-6 September 2017*, 67.

Windmeijer, F. A. (1995). Goodness-of-fit measures in binary choice models. Econometric Reviews, 14(1), 101-116.

Young, J. (2013). Identification of High Collision Locations for the City of Regina Using Gis and Post-network Screening Analysis (Doctoral dissertation, University of Saskatchewan).

Young, J., and Park, P. Y. (2014). Hotzone identification with GIS-based post-network screening analysis. Journal of Transport Geography, 34, 106-120.

Yu, H., Liu, P., Chen, J., & Wang, H. (2014). Comparative analysis of the spatial analysis methods for hotspot identification. Accident Analysis & Prevention, 66, 80-88.

APPENDICES

Appendix A: Negative Binomial Models – Candidate Models

Candidate Models for Canada

```
# Linear combination of variables using years as offset and Max Train Count as exposure
Derail_model1 <- glm.nb(DerailCount ~ log_VL_Count + offset (log(Years)))
Derail_model2 <- glm.nb(DerailCount ~ log_VL_Count + VL_TrnSpd + offset(log(Years)))
Derail_model3 <- glm.nb(DerailCount ~ log_VL_Count + VL_TrnSpd + Stn_Count + offset(log(Years)))
Derail_model4 <- glm.nb(DerailCount ~ log_VL_Count + VL_TrnSpd + Stn_Count + Seg_Length + offset(log(Years)))
Derail_model5 <- glm.nb(DerailCount ~ log_VL_Count + log_VL_TrnSpd + offset(log(Years)))
Derail_model6 <- glm.nb(DerailCount ~ log_VL_Count + log_VL_TrnSpd + Stn_Count + offset(log(Years)))
Derail_model7 <- glm.nb(DerailCount ~ log_VL_Count + log_VL_TrnSpd + Stn_Count + Seg_Length + offset(log(Years)))
Derail_model8 <- glm.nb(DerailCount ~ log_VL_Count + log_VL_TrnSpd + Stn_Count + log_Seg_Length + offset(log(Years)))
Derail_model9 <- glm.nb(DerailCount ~ log_VL_Count + VL_TrnSpd + Stn_Count + log_Seg_Length + offset(log(Years)))

# Linear combination of variables using years as offset and Average Train Count as exposure
Derail_model10 <- glm.nb(DerailCount ~ Avg_Train + offset(log(Years)))
Derail_model11 <- glm.nb(DerailCount ~ Avg_Train + VL_TrnSpd + offset(log(Years)))
Derail_model12 <- glm.nb(DerailCount ~ Avg_Train + VL_TrnSpd + Stn_Count + offset(log(Years)))
Derail_model13 <- glm.nb(DerailCount ~ Avg_Train + VL_TrnSpd + Stn_Count + Seg_Length + offset(log(Years)))
Derail_model14 <- glm.nb(DerailCount ~ Avg_Train + log_VL_TrnSpd + offset(log(Years)))
Derail_model15 <- glm.nb(DerailCount ~ Avg_Train + log_VL_TrnSpd + Stn_Count + offset(log(Years)))
Derail_model16 <- glm.nb(DerailCount ~ Avg_Train + log_VL_TrnSpd + Stn_Count + Seg_Length + offset(log(Years)))
Derail_model17 <- glm.nb(DerailCount ~ Avg_Train + log_VL_TrnSpd + Stn_Count + log_Seg_Length + offset(log(Years)))
Derail_model18 <- glm.nb(DerailCount ~ Avg_Train + VL_TrnSpd + Stn_Count + log_Seg_Length + offset(log(Years)))

# Linear combination of variables using years and segment length as offsets and max Train Count
Derail_model19 <- glm.nb(DerailCount ~ log_VL_Count + offset(log(Years)) + offset(log_Seg_Length))
Derail_model20 <- glm.nb(DerailCount ~ log_VL_Count + VL_TrnSpd + offset(log(Years)) + offset(log_Seg_Length))
Derail_model21 <- glm.nb(DerailCount ~ log_VL_Count + VL_TrnSpd + Stn_Count + offset(log(Years)) + offset(log_Seg_Length))
Derail_model22 <- glm.nb(DerailCount ~ log_VL_Count + log_VL_TrnSpd + offset(log(Years)) + offset(log_Seg_Length))
```

```
Derail_model23 <- glm.nb(DerailCount ~ log_VL_Count + log_VL_TrnSpd + Stn_Count + offset(log(Years)) + offset(log_Seg_Length))

# Linear combination of variables using years and segment length as offsets and Average Train Count
Derail_model24 <- glm.nb (DerailCount ~ Avg_Train + offset(log(Years)) + offset(log_Seg_Length))
Derail_model25 <- glm.nb(DerailCount ~ Avg_Train + VL_TrnSpd + offset(log(Years)) + offset(log_Seg_Length))
Derail_model26 <- glm.nb(DerailCount ~ Avg_Train + VL_TrnSpd + Stn_Count + offset(log(Years)) + offset(log_Seg_Length))
Derail_model27 <- glm.nb(DerailCount ~ Avg_Train + log_VL_TrnSpd + offset(log(Years)) + offset(log_Seg_Length))
Derail_model28 <- glm.nb(DerailCount ~ Avg_Train + log_VL_TrnSpd + Stn_Count + offset(log(Years)) + offset(log_Seg_Length))

# Linear combination of variables using years as offset, Max Train Count as exposure and log_TrnSpd_VLCount as interaction variable
Derail_model29 <- glm.nb(DerailCount ~ log_VL_Count + offset(log(Years)) + log_TrnSpd_VLCount)
Derail_model30 <- glm.nb(DerailCount ~ log_VL_Count + VL_TrnSpd + offset(log(Years)) + log_TrnSpd_VLCount)
Derail_model31 <- glm.nb(DerailCount ~ log_VL_Count + VL_TrnSpd + Stn_Count + offset(log(Years)) + log_TrnSpd_VLCount)
Derail_model32 <- glm.nb(DerailCount ~ log_VL_Count + VL_TrnSpd + Stn_Count + Seg_Length + offset(log(Years)) + log_TrnSpd_VLCount)
Derail_model33 <- glm.nb(DerailCount ~ log_VL_Count + log_VL_TrnSpd + offset(log(Years)) + log_TrnSpd_VLCount)
Derail_model34 <- glm.nb(DerailCount ~ log_VL_Count + log_VL_TrnSpd + Stn_Count + offset(log(Years))+ log_TrnSpd_VLCount)
Derail_model35 <- glm.nb(DerailCount ~ log_VL_Count + log_VL_TrnSpd + Stn_Count + Seg_Length + offset(log(Years)) + log_TrnSpd_VLCount)
Derail_model36 <- glm.nb(DerailCount ~ log_VL_Count + log_VL_TrnSpd + Stn_Count + log_Seg_Length + offset(log(Years)) + log_TrnSpd_VLCount)
Derail_model37 <- glm.nb(DerailCount ~ log_VL_Count + VL_TrnSpd + Stn_Count + log_Seg_Length + offset(log(Years)) + log_TrnSpd_VLCount)

# Linear combination of variables using years as offset, Max Train Count as exposure and log_TrnSpd_AvgTrain as interaction variable
Derail_model38 <- glm.nb(DerailCount ~ log_VL_Count + offset(log(Years)) + log_TrnSpd_AvgTrain)
Derail_model39 <- glm.nb(DerailCount ~ log_VL_Count + VL_TrnSpd + offset(log(Years)) + log_TrnSpd_AvgTrain)
Derail_model40 <- glm.nb(DerailCount ~ log_VL_Count + VL_TrnSpd + Stn_Count + offset(log(Years)) + log_TrnSpd_AvgTrain)
Derail_model41 <- glm.nb(DerailCount ~ log_VL_Count + VL_TrnSpd + Stn_Count + Seg_Length + offset(log(Years)) + log_TrnSpd_AvgTrain)
Derail_model42 <- glm.nb(DerailCount ~ log_VL_Count + log_VL_TrnSpd + offset(log(Years)) + log_TrnSpd_AvgTrain)
Derail_model43 <- glm.nb(DerailCount ~ log_VL_Count + log_VL_TrnSpd + Stn_Count + offset(log(Years)) + log_TrnSpd_AvgTrain)
Derail_model44 <- glm.nb(DerailCount ~ log_VL_Count + log_VL_TrnSpd + Stn_Count + Seg_Length + offset(log(Years)) + log_TrnSpd_AvgTrain)
Derail_model45 <- glm.nb(DerailCount ~ log_VL_Count + log_VL_TrnSpd + Stn_Count + log_Seg_Length + offset(log(Years)) + log_TrnSpd_AvgTrain)
Derail_model46 <- glm.nb(DerailCount ~ log_VL_Count + VL_TrnSpd + Stn_Count + log_Seg_Length + offset(log(Years)) + log_TrnSpd_AvgTrain)
```

```
# Linear combination of variables using years and segment length as offsets, max Train Count as exposure and log_TrnSpd_VLCount as interaction variable
Derail_model47 <- glm.nb(DerailCount ~ log_VL_Count + offset(log(Years)) + offset(log_Seg_Length) + log_TrnSpd_VLCount)
Derail_model48 <- glm.nb(DerailCount ~ log_VL_Count + VL_TrnSpd + offset(log(Years)) + offset(log_Seg_Length) + log_TrnSpd_VLCount)
Derail_model49 <- glm.nb(DerailCount ~ log_VL_Count + VL_TrnSpd + Stn_Count + offset(log(Years)) + offset(log_Seg_Length) + log_TrnSpd_VLCount)
Derail_model50 <- glm.nb(DerailCount ~ log_VL_Count + log_VL_TrnSpd + offset(log(Years)) + offset(log_Seg_Length) + log_TrnSpd_VLCount)
Derail_model51 <- glm.nb(DerailCount ~ log_VL_Count + log_VL_TrnSpd + Stn_Count + offset(log(Years)) + offset(log_Seg_Length) + log_TrnSpd_VLCount)

# Linear combination of variables using years and segment length as offsets, Average Train Count as exposure and log_TrnSpd_VLCount as interaction variable
Derail_model52 <- glm.nb(DerailCount ~ Avg_Train + offset(log(Years)) + offset(log_Seg_Length) + log_TrnSpd_VLCount)
Derail_model53 <- glm.nb(DerailCount ~ Avg_Train + VL_TrnSpd + offset(log(Years)) + offset(log_Seg_Length) + log_TrnSpd_VLCount)
Derail_model54 <- glm.nb(DerailCount ~ Avg_Train + VL_TrnSpd + Stn_Count + offset(log(Years)) + offset(log_Seg_Length) + log_TrnSpd_VLCount)
Derail_model55 <- glm.nb(DerailCount ~ Avg_Train + log_VL_TrnSpd + offset(log(Years)) + offset(log_Seg_Length) + log_TrnSpd_VLCount)
Derail_model56 <- glm.nb(DerailCount ~ Avg_Train + log_VL_TrnSpd + Stn_Count + offset(log(Years)) + offset(log_Seg_Length) + log_TrnSpd_VLCount)

#Square of train speed or train count
Derail_model57 <- glm.nb(DerailCount ~ log_VL_Count + VL_TrnSpd^2 + Stn_Count + Seg_Length + offset(log(Years)))
Derail_model58 <- glm.nb(DerailCount ~ log_VL_Count + VL_Count^2 + VL_TrnSpd + Stn_Count + Seg_Length + offset(log(Years)))
Derail_model59 <- glm.nb(DerailCount ~ Avg_Train + VL_TrnSpd^2 + Stn_Count + Seg_Length + offset(log(Years)))
Derail_model60 <- glm.nb(DerailCount ~ Avg_Train + Avg_Train^2 + VL_TrnSpd + Stn_Count + Seg_Length + offset(log(Years)))

#Square of train speed or train count and log_TrnSpd_VLCount as interaction variable
Derail_model57 <- glm.nb(DerailCount ~ log_VL_Count + VL_TrnSpd^2 + Stn_Count + Seg_Length + offset(log(Years)) + log_TrnSpd_VLCount)
Derail_model58 <- glm.nb(DerailCount ~ log_VL_Count + VL_Count^2 + VL_TrnSpd + Stn_Count + Seg_Length + offset(log(Years)) + log_TrnSpd_VLCount)
Derail_model59 <- glm.nb(DerailCount ~ Avg_Train + VL_TrnSpd^2 + Stn_Count + Seg_Length + offset(log(Years)) + log_TrnSpd_VLCount)
Derail_model60 <- glm.nb(DerailCount ~ Avg_Train + Avg_Train^2 + VL_TrnSpd + Stn_Count + Seg_Length + offset(log(Years)) + log_TrnSpd_VLCount)

#Square of train speed or train count and log_TrnSpd_AvgTrain as interaction variable
Derail_model61 <- glm.nb(DerailCount ~ log_VL_Count + VL_TrnSpd^2 + Stn_Count + Seg_Length + offset(log(Years)) + log_TrnSpd_AvgTrain)
Derail_model62 <- glm.nb(DerailCount ~ log_VL_Count + VL_Count^2 + VL_TrnSpd + Stn_Count + Seg_Length + offset(log(Years)) + log_TrnSpd_AvgTrain)
Derail_model63 <- glm.nb(DerailCount ~ Avg_Train + VL_TrnSpd^2 + Stn_Count + Seg_Length + offset(log(Years)) + log_TrnSpd_AvgTrain)
Derail_model64 <- glm.nb(DerailCount ~ Avg_Train + Avg_Train^2 + VL_TrnSpd + Stn_Count + Seg_Length + offset(log(Years)) + log_TrnSpd_AvgTrain)
```

```
# Refined Models
Derail_model65 <- glm.nb(DerailCount ~ log_VL_Count + log_VL_TrnSpd + Stn_Count + log_Seg_Length + offset(log(Years)) + log_TrnSpd_VLCount + TrnSpd_StnCount)
Derail_model66 <- glm.nb(DerailCount ~ log_VL_Count + log_VL_TrnSpd + Stn_Count + log_Seg_Length + offset(log(Years)) + TrnSpd_VLCount + TrnSpd_StnCount)
Derail_model67 <- glm.nb(DerailCount ~ log_VL_Count + VL_TrnSpd + Stn_Count + Seg_Length + offset(log(Years)) + log_TrnSpd_VLCount)
Derail_model68 <- glm.nb(DerailCount ~ log_VL_Count + VL_TrnSpd + Stn_Count + Seg_Length + offset(log(Years)) + TrnSpd_VLCount + TrnSpd_StnCount)
Derail_model69 <- glm.nb(DerailCount ~ log_VL_Count + VL_TrnSpd + Stn_Count + Seg_Length + offset(log(Years)) + TrnSpd_VLCount)
Derail_model70 <- glm.nb(DerailCount ~ log_VL_Count + VL_TrnSpd + Stn_Count + Seg_Length + offset(log(Years)) + TrnSpd_StnCount)
Derail_model71 <- glm.nb(DerailCount ~ log_VL_Count + VL_TrnSpd + Stn_Count + offset(log(Seg_Length)) + offset(log(Years)) + log_TrnSpd_VLCount)
```

Candidate Models for Eastern Canada

```
# Linear combination of variables using years as offset and Max Train Count as exposure
Derail_model1 <- glm.nb(DerailCount ~ log_VL_Count + offset(log(Years)))
Derail_model2 <- glm.nb(DerailCount ~ log_VL_Count + VL_TrnSpd + offset(log(Years)))
Derail_model3 <- glm.nb(DerailCount ~ log_VL_Count + VL_TrnSpd + Stn_Count + offset(log(Years)))
Derail_model4 <- glm.nb(DerailCount ~ log_VL_Count + VL_TrnSpd + Stn_Count + Seg_Length + offset(log(Years)))
Derail_model5 <- glm.nb(DerailCount ~ log_VL_Count + log_VL_TrnSpd + offset(log(Years)))
Derail_model6 <- glm.nb(DerailCount ~ log_VL_Count + log_VL_TrnSpd + Stn_Count + offset(log(Years)))
Derail_model7 <- glm.nb(DerailCount ~ log_VL_Count + log_VL_TrnSpd + Stn_Count + Seg_Length + offset(log(Years)))
Derail_model8 <- glm.nb(DerailCount ~ log_VL_Count + log_VL_TrnSpd + Stn_Count + log_Seg_Length + offset(log(Years)))
Derail_model9 <- glm.nb(DerailCount ~ log_VL_Count + VL_TrnSpd + Stn_Count + log_Seg_Length + offset(log(Years)))

# Linear combination of variables using years as offset and Average Train Count as exposure
Derail_model10 <- glm.nb(DerailCount ~ Avg_Train + offset(log(Years)))
Derail_model11 <- glm.nb(DerailCount ~ Avg_Train + VL_TrnSpd + offset(log(Years)))
Derail_model12 <- glm.nb(DerailCount ~ Avg_Train + VL_TrnSpd + Stn_Count + offset(log(Years)))
Derail_model13 <- glm.nb(DerailCount ~ Avg_Train + VL_TrnSpd + Stn_Count + Seg_Length + offset(log(Years)))
Derail_model14 <- glm.nb(DerailCount ~ Avg_Train + log_VL_TrnSpd + offset(log(Years)))
Derail_model15 <- glm.nb(DerailCount ~ Avg_Train + log_VL_TrnSpd + Stn_Count + offset(log(Years)))
Derail_model16 <- glm.nb(DerailCount ~ Avg_Train + log_VL_TrnSpd + Stn_Count + Seg_Length + offset(log(Years)))
Derail_model17 <- glm.nb(DerailCount ~ Avg_Train + log_VL_TrnSpd + Stn_Count + log_Seg_Length + offset(log(Years)))
Derail_model18 <- glm.nb(DerailCount ~ Avg_Train + VL_TrnSpd + Stn_Count + log_Seg_Length + offset(log(Years)))

# Linear combination of variables using years and segment length as offsets and max Train Count
Derail_model19 <- glm.nb(DerailCount ~ log_VL_Count + offset(log(Years)) + offset(log_Seg_Length))
Derail_model20 <- glm.nb(DerailCount ~ log_VL_Count + VL_TrnSpd + offset(log(Years)) + offset(log_Seg_Length))
Derail_model21 <- glm.nb(DerailCount ~ log_VL_Count + VL_TrnSpd + Stn_Count + offset(log(Years)) + offset(log_Seg_Length))
Derail_model22 <- glm.nb(DerailCount ~ log_VL_Count + log_VL_TrnSpd + offset(log(Years)) + offset(log_Seg_Length))
Derail_model23 <- glm.nb(DerailCount ~ log_VL_Count + log_VL_TrnSpd + Stn_Count + offset(log(Years)) + offset(log_Seg_Length))
# Linear combination of variables using years and segment length as offsets and Average Train Count
```

```
Derail_model24 <- glm.nb(DerailCount ~ Avg_Train + offset(log(Years)) + offset(log_Seg_Length))
Derail_model25 <- glm.nb(DerailCount ~ Avg_Train + VL_TrnSpd + offset(log(Years)) + offset(log_Seg_Length))
Derail_model26 <- glm.nb(DerailCount ~ Avg_Train + VL_TrnSpd + Stn_Count + offset(log(Years)) + offset(log_Seg_Length))
Derail_model27 <- glm.nb(DerailCount ~ Avg_Train + log_VL_TrnSpd + offset(log(Years)) + offset(log_Seg_Length))
Derail_model28 <- glm.nb(DerailCount ~ Avg_Train + log_VL_TrnSpd + Stn_Count + offset(log(Years)) + offset(log_Seg_Length))

# Linear combination of variables using years as offset, Max Train Count as exposure and log_TrnSpd_VLCount as interaction variable
Derail_model29 <- glm.nb(DerailCount ~ log_VL_Count + offset(log(Years)) + log_TrnSpd_VLCount)
Derail_model30 <- glm.nb(DerailCount ~ log_VL_Count + VL_TrnSpd + offset(log(Years)) + log_TrnSpd_VLCount)
Derail_model31 <- glm.nb(DerailCount ~ log_VL_Count + VL_TrnSpd + Stn_Count + offset(log(Years)) + log_TrnSpd_VLCount)
Derail_model32 <- glm.nb(DerailCount ~ log_VL_Count + VL_TrnSpd + Stn_Count + Seg_Length + offset(log(Years)) + log_TrnSpd_VLCount)
Derail_model33 <- glm.nb(DerailCount ~ log_VL_Count + log_VL_TrnSpd + offset(log(Years)) + log_TrnSpd_VLCount)
Derail_model34 <- glm.nb(DerailCount ~ log_VL_Count + log_VL_TrnSpd + Stn_Count + offset(log(Years))+ log_TrnSpd_VLCount)
Derail_model35 <- glm.nb(DerailCount ~ log_VL_Count + log_VL_TrnSpd + Stn_Count + Seg_Length + offset(log(Years)) + log_TrnSpd_VLCount)
Derail_model36 <- glm.nb(DerailCount ~ log_VL_Count + log_VL_TrnSpd + Stn_Count + log_Seg_Length + offset(log(Years)) + log_TrnSpd_VLCount)
Derail_model37 <- glm.nb(DerailCount ~ log_VL_Count + VL_TrnSpd + Stn_Count + log_Seg_Length + offset(log(Years)) + log_TrnSpd_VLCount)

# Linear combination of variables using years as offset, Max Train Count as exposure and log_TrnSpd_AvgTrain as interaction variable
Derail_model38 <- glm.nb(DerailCount ~ log_VL_Count + offset(log(Years)) + log_TrnSpd_AvgTrain)
Derail_model39 <- glm.nb(DerailCount ~ log_VL_Count + VL_TrnSpd + offset(log(Years)) + log_TrnSpd_AvgTrain)
Derail_model40 <- glm.nb(DerailCount ~ log_VL_Count + VL_TrnSpd + Stn_Count + offset(log(Years)) + log_TrnSpd_AvgTrain)
Derail_model41 <- glm.nb(DerailCount ~ log_VL_Count + VL_TrnSpd + Stn_Count + Seg_Length + offset(log(Years)) + log_TrnSpd_AvgTrain)
Derail_model42 <- glm.nb(DerailCount ~ log_VL_Count + log_VL_TrnSpd + offset(log(Years)) + log_TrnSpd_AvgTrain)
Derail_model43 <- glm.nb(DerailCount ~ log_VL_Count + log_VL_TrnSpd + Stn_Count + offset(log(Years)) + log_TrnSpd_AvgTrain)
Derail_model44 <- glm.nb(DerailCount ~ log_VL_Count + log_VL_TrnSpd + Stn_Count + Seg_Length + offset(log(Years)) + log_TrnSpd_AvgTrain)
Derail_model45 <- glm.nb(DerailCount ~ log_VL_Count + log_VL_TrnSpd + Stn_Count + log_Seg_Length + offset(log(Years)) + log_TrnSpd_AvgTrain)
Derail_model46 <- glm.nb(DerailCount ~ log_VL_Count + VL_TrnSpd + Stn_Count + log_Seg_Length + offset(log(Years)) + log_TrnSpd_AvgTrain)

# Linear combination of variables using years and segment length as offsets, max Train Count as exposure and log_TrnSpd_VLCount as interaction variable
Derail_model47 <- glm.nb(DerailCount ~ log_VL_Count + offset(log(Years)) + offset(log_Seg_Length) + log_TrnSpd_VLCount)
Derail_model48 <- glm.nb(DerailCount ~ log_VL_Count + VL_TrnSpd + offset(log(Years)) + offset(log_Seg_Length) + log_TrnSpd_VLCount)
```

```
Derail_model49 <- glm.nb(DerailCount ~ log_VL_Count + VL_TrnSpd + Stn_Count + offset(log(Years)) + offset(log_Seg_Length) + log_TrnSpd_VLCount)
Derail_model50 <- glm.nb(DerailCount ~ log_VL_Count + log_VL_TrnSpd + offset(log(Years)) + offset(log_Seg_Length) + log_TrnSpd_VLCount)
Derail_model51 <- glm.nb(DerailCount ~ log_VL_Count + log_VL_TrnSpd + Stn_Count + offset(log(Years)) + offset(log_Seg_Length) + log_TrnSpd_VLCount)

# Linear combination of variables using years and segment length as offsets, Average Train Count as exposure and log_TrnSpd_VLCount as interaction variable
Derail_model52 <- glm.nb(DerailCount ~ Avg_Train + offset(log(Years)) + offset(log_Seg_Length) + log_TrnSpd_VLCount)
Derail_model53 <- glm.nb(DerailCount ~ Avg_Train + VL_TrnSpd + offset(log(Years)) + offset(log_Seg_Length) + log_TrnSpd_VLCount)
Derail_model54 <- glm.nb(DerailCount ~ Avg_Train + VL_TrnSpd + Stn_Count + offset(log(Years)) + offset(log_Seg_Length) + log_TrnSpd_VLCount)
Derail_model55 <- glm.nb(DerailCount ~ Avg_Train + log_VL_TrnSpd + offset(log(Years)) + offset(log_Seg_Length) + log_TrnSpd_VLCount)
Derail_model56 <- glm.nb(DerailCount ~ Avg_Train + log_VL_TrnSpd + Stn_Count + offset(log(Years)) + offset(log_Seg_Length) + log_TrnSpd_VLCount)

#Square of train speed or train count
Derail_model57 <- glm.nb(DerailCount ~ log_VL_Count + VL_TrnSpd^2 + Stn_Count + Seg_Length + offset(log(Years)))
Derail_model58 <- glm.nb(DerailCount ~ log_VL_Count + VL_Count^2 + VL_TrnSpd + Stn_Count + Seg_Length + offset(log(Years)))
Derail_model59 <- glm.nb(DerailCount ~ Avg_Train + VL_TrnSpd^2 + Stn_Count + Seg_Length + offset(log(Years)))
Derail_model60 <- glm.nb(DerailCount ~ Avg_Train + Avg_Train^2 + VL_TrnSpd + Stn_Count + Seg_Length + offset(log(Years)))

#Square of train speed or train count and log_TrnSpd_VLCount as interaction variable
Derail_model57 <- glm.nb(DerailCount ~ log_VL_Count + VL_TrnSpd^2 + Stn_Count + Seg_Length + offset(log(Years)) + log_TrnSpd_VLCount)
Derail_model58 <- glm.nb(DerailCount ~ log_VL_Count + VL_Count^2 + VL_TrnSpd + Stn_Count + Seg_Length + offset(log(Years)) + log_TrnSpd_VLCount)
Derail_model59 <- glm.nb(DerailCount ~ Avg_Train + VL_TrnSpd^2 + Stn_Count + Seg_Length + offset(log(Years)) + log_TrnSpd_VLCount)
Derail_model60 <- glm.nb(DerailCount ~ Avg_Train + Avg_Train^2 + VL_TrnSpd + Stn_Count + Seg_Length + offset(log(Years)) + log_TrnSpd_VLCount)

#Square of train speed or train count and log_TrnSpd_AvgTrain as interaction variable
Derail_model61 <- glm.nb(DerailCount ~ log_VL_Count + VL_TrnSpd^2 + Stn_Count + Seg_Length + offset(log(Years)) + log_TrnSpd_AvgTrain)
Derail_model62 <- glm.nb(DerailCount ~ log_VL_Count + VL_Count^2 + VL_TrnSpd + Stn_Count + Seg_Length + offset(log(Years)) + log_TrnSpd_AvgTrain)
Derail_model63 <- glm.nb(DerailCount ~ Avg_Train + VL_TrnSpd^2 + Stn_Count + Seg_Length + offset(log(Years)) + log_TrnSpd_AvgTrain)
Derail_model64 <- glm.nb(DerailCount ~ Avg_Train + Avg_Train^2 + VL_TrnSpd + Stn_Count + Seg_Length + offset(log(Years)) + log_TrnSpd_AvgTrain)
```

```
#Refined Models
Derail_model65 <- glm.nb(DerailCount ~ log_VL_Count + log_VL_TrnSpd + Stn_Count + log_Seg_Length + offset(log(Years)) + log_TrnSpd_VLCount + TrnSpd_StnCount)
Derail_model66 <- glm.nb(DerailCount ~ log_VL_Count + log_VL_TrnSpd + Stn_Count + log_Seg_Length + offset(log(Years)) + TrnSpd_VLCount + TrnSpd_StnCount)
Derail_model67 <- glm.nb(DerailCount ~ log_VL_Count + VL_TrnSpd + Stn_Count + Seg_Length + offset(log(Years)) + TrnSpd_VLCount)
Derail_model68 <- glm.nb(DerailCount ~ log_VL_Count + VL_TrnSpd + Stn_Count + Seg_Length + offset(log(Years)) + TrnSpd_VLCount + TrnSpd_StnCount)
Derail_model69 <- glm.nb(DerailCount ~ log_VL_Count + VL_TrnSpd + Stn_Count + Seg_Length + offset(log(Years)) + TrnSpd_StnCount)
Derail_model70 <- glm.nb(DerailCount ~ log_VL_Count + VL_TrnSpd + Stn_Count + Seg_Length + offset(log(Years)) + log_TrnSpd_VLCount)
Derail_model71 <- glm.nb(DerailCount ~ log_VL_Count + VL_TrnSpd + Stn_Count + Seg_Length + offset(log(Years)) + StnCount_SegLen + log(VLCount_SegLen))
Derail_model72 <- glm.nb(DerailCount ~ log_VL_Count + VL_TrnSpd + Stn_Count + Seg_Length + offset(log(Years)) + log(VLCount_SegLen))
```

Candidate Models for Western Canada

```
# Linear combination of variables using years as offset and Max Train Count as exposure
Derail_model1 <- glm.nb(DerailCount ~ log_VL_Count + offset(log(Years)))
Derail_model2 <- glm.nb(DerailCount ~ log_VL_Count + VL_TrnSpd + offset(log(Years)))
Derail_model3 <- glm.nb(DerailCount ~ log_VL_Count + VL_TrnSpd + Stn_Count + offset(log(Years)))
Derail_model4 <- glm.nb(DerailCount ~ log_VL_Count + VL_TrnSpd + Stn_Count + Seg_Length + offset(log(Years)))
Derail_model5 <- glm.nb(DerailCount ~ log_VL_Count + log_VL_TrnSpd + offset(log(Years)))
Derail_model6 <- glm.nb(DerailCount ~ log_VL_Count + log_VL_TrnSpd + Stn_Count + offset(log(Years)))
Derail_model7 <- glm.nb(DerailCount ~ log_VL_Count + log_VL_TrnSpd + Stn_Count + Seg_Length + offset(log(Years)))
Derail_model8 <- glm.nb(DerailCount ~ log_VL_Count + log_VL_TrnSpd + Stn_Count + log_Seg_Length + offset(log(Years)))
Derail_model9 <- glm.nb(DerailCount ~ log_VL_Count + VL_TrnSpd + Stn_Count + log_Seg_Length + offset(log(Years)))

# Linear combination of variables using years as offset and Average Train Count as exposure
Derail_model10 <- glm.nb(DerailCount ~ log_AvgTrain + offset(log(Years)))
Derail_model11 <- glm.nb(DerailCount ~ log_AvgTrain + VL_TrnSpd + offset(log(Years)))
Derail_model12 <- glm.nb(DerailCount ~ log_AvgTrain + VL_TrnSpd + Stn_Count + offset(log(Years)))
Derail_model13 <- glm.nb(DerailCount ~ log_AvgTrain + VL_TrnSpd + Stn_Count + Seg_Length + offset(log(Years)))
Derail_model14 <- glm.nb(DerailCount ~ log_AvgTrain + log_VL_TrnSpd + offset(log(Years)))
Derail_model15 <- glm.nb(DerailCount ~ log_AvgTrain + log_VL_TrnSpd + Stn_Count + offset(log(Years)))
Derail_model16 <- glm.nb(DerailCount ~ log_AvgTrain + log_VL_TrnSpd + Stn_Count + Seg_Length + offset(log(Years)))
Derail_model17 <- glm.nb(DerailCount ~ log_AvgTrain + log_VL_TrnSpd + Stn_Count + log_Seg_Length + offset(log(Years)))
Derail_model18 <- glm.nb(DerailCount ~ log_AvgTrain + VL_TrnSpd + Stn_Count + log_Seg_Length + offset(log(Years)))

# Linear combination of variables using years and segment length as offsets and max Train Count
Derail_model19 <- glm.nb(DerailCount ~ log_VL_Count + offset(log(Years)) + offset(log_Seg_Length))
Derail_model20 <- glm.nb(DerailCount ~ log_VL_Count + VL_TrnSpd + offset(log(Years)) + offset(log_Seg_Length))
Derail_model21 <- glm.nb(DerailCount ~ log_VL_Count + VL_TrnSpd + Stn_Count + offset(log(Years)) + offset(log_Seg_Length))
Derail_model22 <- glm.nb(DerailCount ~ log_VL_Count + log_VL_TrnSpd + offset(log(Years)) + offset(log_Seg_Length))
Derail_model23 <- glm.nb(DerailCount ~ log_VL_Count + log_VL_TrnSpd + Stn_Count + offset(log(Years)) + offset(log_Seg_Length))
```

```
# Linear combination of variables using years and segment length as offsets and Average Train Count
Derail_model24 <- glm.nb(DerailCount ~ log_AvgTrain + offset(log(Years)) + offset(log_Seg_Length))
Derail_model25 <- glm.nb(DerailCount ~ log_AvgTrain + VL_TrnSpd + offset(log(Years)) + offset(log_Seg_Length))
Derail_model26 <- glm.nb(DerailCount ~ log_AvgTrain + VL_TrnSpd + Stn_Count + offset(log(Years)) + offset(log_Seg_Length))
Derail_model27 <- glm.nb(DerailCount ~ log_AvgTrain + log_VL_TrnSpd + offset(log(Years)) + offset(log_Seg_Length))
Derail_model28 <- glm.nb(DerailCount ~ log_AvgTrain + log_VL_TrnSpd + Stn_Count + offset(log(Years)) + offset(log_Seg_Length))

# Linear combination of variables using years as offset, Max Train Count as exposure and TrnSpd_VLCount as interaction variable
Derail_model29 <- glm.nb(DerailCount ~ log_VL_Count + VL_TrnSpd + Stn_Count + Seg_Length + offset(log(Years))+ log_TrnSpd_VLCount + log_VLCount_SegLen)
Derail_model30 <- glm.nb(DerailCount ~ log_VL_Count + VL_TrnSpd + Stn_Count + Seg_Length + offset(log(Years)) + VLCount_SegLen)
Derail_model31 <- glm.nb(DerailCount ~ log_VL_Count + VL_TrnSpd + Stn_Count + offset(log(Years)) + log_TrnSpd_VLCount)
Derail_model32 <- glm.nb(DerailCount ~ log_VL_Count + VL_TrnSpd + Stn_Count + Seg_Length + offset(log(Years)) + log_TrnSpd_VLCount)
Derail_model33 <- glm.nb(DerailCount ~ log_VL_Count + log_VL_TrnSpd + offset(log(Years)) + log_TrnSpd_VLCount)
Derail_model34 <- glm.nb(DerailCount ~ log_VL_Count + log_VL_TrnSpd + Stn_Count + offset(log(Years))+ log_TrnSpd_VLCount)
Derail_model35 <- glm.nb(DerailCount ~ log_VL_Count + log_VL_TrnSpd + Stn_Count + Seg_Length + offset(log(Years)) + log_TrnSpd_VLCount)
Derail_model36 <- glm.nb(DerailCount ~ log_VL_Count + log_VL_TrnSpd + Stn_Count + log_Seg_Length + offset(log(Years)) + log_TrnSpd_VLCount)
Derail_model37 <- glm.nb(DerailCount ~ log_VL_Count + VL_TrnSpd + Stn_Count + log_Seg_Length + offset(log(Years)) + log_TrnSpd_VLCount)

# Linear combination of variables using years as offset, Max Train Count as exposure and log_TrnSpd_AvgTrain as interaction variable
Derail_model38 <- glm.nb(DerailCount ~ log_VL_Count + offset(log(Years)) + log_TrnSpd_AvgTrain)
Derail_model39 <- glm.nb(DerailCount ~ log_VL_Count + VL_TrnSpd + offset(log(Years)) + log_TrnSpd_AvgTrain)
Derail_model40 <- glm.nb(DerailCount ~ log_VL_Count + VL_TrnSpd + Stn_Count + offset(log(Years)) + log_TrnSpd_AvgTrain)
Derail_model41 <- glm.nb(DerailCount ~ log_VL_Count + VL_TrnSpd + Stn_Count + Seg_Length + offset(log(Years)) + log_TrnSpd_AvgTrain)
Derail_model42 <- glm.nb(DerailCount ~ log_VL_Count + log_VL_TrnSpd + offset(log(Years)) + log_TrnSpd_AvgTrain)
Derail_model43 <- glm.nb(DerailCount ~ log_VL_Count + log_VL_TrnSpd + Stn_Count + offset(log(Years)) + log_TrnSpd_AvgTrain)
Derail_model44 <- glm.nb(DerailCount ~ log_VL_Count + log_VL_TrnSpd + Stn_Count + Seg_Length + offset(log(Years)) + log_TrnSpd_AvgTrain)
Derail_model45 <- glm.nb(DerailCount ~ log_VL_Count + log_VL_TrnSpd + Stn_Count + log_Seg_Length + offset(log(Years)) + log_TrnSpd_AvgTrain)
Derail_model46 <- glm.nb(DerailCount ~ log_VL_Count + VL_TrnSpd + Stn_Count + log_Seg_Length + offset(log(Years)) + log_TrnSpd_AvgTrain)
```

```
# Linear combination of variables using years and segment length as offsets, max Train Count as exposure and log_TrnSpd_VLCount as interaction variable
Derail_model47 <- glm.nb(DerailCount ~ log_VL_Count + offset(log(Years)) + offset(log_Seg_Length) + log_TrnSpd_VLCount)
Derail_model48 <- glm.nb(DerailCount ~ log_VL_Count + VL_TrnSpd + offset(log(Years)) + offset(log_Seg_Length) + log_TrnSpd_VLCount)
Derail_model49 <- glm.nb(DerailCount ~ log_VL_Count + VL_TrnSpd + Stn_Count + offset(log(Years)) + offset(log_Seg_Length) + log_TrnSpd_VLCount)
Derail_model50 <- glm.nb(DerailCount ~ log_VL_Count + log_VL_TrnSpd + offset(log(Years)) + offset(log_Seg_Length) + log_TrnSpd_VLCount)
Derail_model51 <- glm.nb(DerailCount ~ log_VL_Count + log_VL_TrnSpd + Stn_Count + offset(log(Years)) + offset(log_Seg_Length) + log_TrnSpd_VLCount)

# Linear combination of variables using years and segment length as offsets, Average Train Count as exposure and log_TrnSpd_VLCount as interaction variable
Derail_model52 <- glm.nb(DerailCount ~ log_AvgTrain + offset(log(Years)) + offset(log_Seg_Length) + log_TrnSpd_VLCount)
Derail_model53 <- glm.nb(DerailCount ~ log_AvgTrain + VL_TrnSpd + offset(log(Years)) + offset(log_Seg_Length) + log_TrnSpd_VLCount)
Derail_model54 <- glm.nb(DerailCount ~ log_AvgTrain + VL_TrnSpd + Stn_Count + offset(log(Years)) + offset(log_Seg_Length) + log_TrnSpd_VLCount)
Derail_model55 <- glm.nb(DerailCount ~ log_AvgTrain + log_VL_TrnSpd + offset(log(Years)) + offset(log_Seg_Length) + log_TrnSpd_VLCount)
Derail_model56 <- glm.nb(DerailCount ~ log_AvgTrain + log_VL_TrnSpd + Stn_Count + offset(log(Years)) + offset(log_Seg_Length) + log_TrnSpd_VLCount)

#Square of train speed or train count
Derail_model57 <- glm.nb(DerailCount ~ log_VL_Count + VL_TrnSpd^2 + Stn_Count + Seg_Length + offset(log(Years)))
Derail_model58 <- glm.nb(DerailCount ~ log_VL_Count + VL_Count^2 + VL_TrnSpd + Stn_Count + Seg_Length + offset(log(Years)))
Derail_model59 <- glm.nb(DerailCount ~ log_AvgTrain + VL_TrnSpd^2 + Stn_Count + Seg_Length + offset(log(Years)))
Derail_model60 <- glm.nb(DerailCount ~ log_AvgTrain + Avg_Train^2 + VL_TrnSpd + Stn_Count + Seg_Length + offset(log(Years)))

#Square of train speed or train count and log_TrnSpd_VLCount as interaction variable
Derail_model57 <- glm.nb(DerailCount ~ log_VL_Count + VL_TrnSpd^2 + Stn_Count + Seg_Length + offset(log(Years)) + log_TrnSpd_VLCount)
Derail_model58 <- glm.nb(DerailCount ~ log_VL_Count + VL_Count^2 + VL_TrnSpd + Stn_Count + Seg_Length + offset(log(Years)) + log_TrnSpd_VLCount)
Derail_model59 <- glm.nb(DerailCount ~ log_AvgTrain + VL_TrnSpd^2 + Stn_Count + Seg_Length + offset(log(Years)) + log_TrnSpd_AvgTrain)
Derail_model60 <- glm.nb(DerailCount ~ log_AvgTrain + Avg_Train^2 + VL_TrnSpd + Stn_Count + Seg_Length + offset(log(Years)) + log_TrnSpd_AvgTrain)

#Square of train speed or train count and log_TrnSpd_AvgTrain as interaction variable
Derail_model61 <- glm.nb(DerailCount ~ log_VL_Count + VL_TrnSpd^2 + Stn_Count + Seg_Length + offset(log(Years)) + log_TrnSpd_AvgTrain)
Derail_model62 <- glm.nb(DerailCount ~ log_VL_Count + VL_Count^2 + VL_TrnSpd + Stn_Count + Seg_Length + offset(log(Years)) + log_TrnSpd_AvgTrain)
Derail_model63 <- glm.nb(DerailCount ~ log_AvgTrain + VL_TrnSpd^2 + Stn_Count + Seg_Length + offset(log(Years)) + log_TrnSpd_AvgTrain)
```

```
Derail_model64 <- glm.nb(DerailCount ~ log_AvgTrain + Avg_Train^2 + VL_TrnSpd + Stn_Count + Seg_Length + offset(log(Years)) + log_TrnSpd_AvgTrain)

#Refined Models
Derail_model65 <- glm.nb(DerailCount ~ Avg_Train + VL_TrnSpd + Stn_Count + Seg_Length + offset(log(Years)))
Derail_model66 <- glm.nb(DerailCount ~ Avg_Train + VL_TrnSpd + Stn_Count + Seg_Length + offset(log(Years)) + log_TrnSpd_AvgTrain)
Derail_model67 <- glm.nb(DerailCount ~ log_AvgTrain + VL_TrnSpd + Stn_Count + Seg_Length + offset(log(Years)))
Derail_model68 <- glm.nb(DerailCount ~ log_AvgTrain + VL_TrnSpd + Stn_Count + Seg_Length + offset(log(Years)) + log_TrnSpd_AvgTrain)
Derail_model69 <- glm.nb(DerailCount ~ Avg_Train + VL_TrnSpd + Stn_Count + Seg_Length + offset(log(Years)) + AvgTrain_SegLen)
Derail_model70 <- glm.nb(DerailCount ~ Avg_Train + VL_TrnSpd + Stn_Count + Seg_Length + offset(log(Years)) + AvgTrain_StnCount)
Derail_model71 <- glm.nb(DerailCount ~ log_VL_Count + VL_TrnSpd + Stn_Count + Seg_Length + offset(log(Years)) + StnCount_SegLen + log(VLCount_SegLen))
Derail_model72 <- glm.nb(DerailCount ~ Avg_Train + VL_TrnSpd + Stn_Count + Seg_Length + offset(log(Years)) + AvgTrain_SegLen + log_TrnSpd_AvgTrain)
```

Candidate Models for Canadian National Railway

```
# Linear combination of variables using years as offset and Max Train Count as exposure
Derail_model1 <- glm.nb(DerailCount ~ log_VL_Count + offset(log(Years)))
Derail_model2 <- glm.nb(DerailCount ~ log_VL_Count + VL_TrnSpd + offset(log(Years)))
Derail_model3 <- glm.nb(DerailCount ~ log_VL_Count + VL_TrnSpd + Stn_Count + offset(log(Years)))
Derail_model4 <- glm.nb(DerailCount ~ log_VL_Count + VL_TrnSpd + Stn_Count + Seg_Length + offset(log(Years)))
Derail_model5 <- glm.nb(DerailCount ~ log_VL_Count + log_VL_TrnSpd + offset(log(Years)))
Derail_model6 <- glm.nb(DerailCount ~ log_VL_Count + log_VL_TrnSpd + Stn_Count + offset(log(Years)))
Derail_model7 <- glm.nb(DerailCount ~ log_VL_Count + log_VL_TrnSpd + Stn_Count + Seg_Length + offset(log(Years)))
Derail_model8 <- glm.nb(DerailCount ~ log_VL_Count + log_VL_TrnSpd + Stn_Count + log_Seg_Length + offset(log(Years)))
Derail_model9 <- glm.nb(DerailCount ~ log_VL_Count + VL_TrnSpd + Stn_Count + log_Seg_Length + offset(log(Years)))

# Linear combination of variables using years as offset and Average Train Count as exposure
Derail_model10 <- glm.nb(DerailCount ~ Avg_Train + offset(log(Years)))
Derail_model11 <- glm.nb(DerailCount ~ Avg_Train + VL_TrnSpd + offset(log(Years)))
Derail_model12 <- glm.nb(DerailCount ~ Avg_Train + VL_TrnSpd + Stn_Count + offset(log(Years)))
Derail_model13 <- glm.nb(DerailCount ~ Avg_Train + VL_TrnSpd + Stn_Count + Seg_Length + offset(log(Years)))
Derail_model14 <- glm.nb(DerailCount ~ Avg_Train + log_VL_TrnSpd + offset(log(Years)))
Derail_model15 <- glm.nb(DerailCount ~ Avg_Train + log_VL_TrnSpd + Stn_Count + offset(log(Years)))
Derail_model16 <- glm.nb(DerailCount ~ Avg_Train + log_VL_TrnSpd + Stn_Count + Seg_Length + offset(log(Years)))
Derail_model17 <- glm.nb(DerailCount ~ Avg_Train + log_VL_TrnSpd + Stn_Count + log_Seg_Length + offset(log(Years)))
Derail_model18 <- glm.nb(DerailCount ~ Avg_Train + VL_TrnSpd + Stn_Count + log_Seg_Length + offset(log(Years)))

# Linear combination of variables using years and segment length as offsets and max Train Count
Derail_model19 <- glm.nb(DerailCount ~ log_VL_Count + offset(log(Years)) + offset(log_Seg_Length))
Derail_model20 <- glm.nb(DerailCount ~ log_VL_Count + VL_TrnSpd + offset(log(Years)) + offset(log_Seg_Length))
Derail_model21 <- glm.nb(DerailCount ~ log_VL_Count + VL_TrnSpd + Stn_Count + offset(log(Years)) + offset(log_Seg_Length))
Derail_model22 <- glm.nb(DerailCount ~ log_VL_Count + log_VL_TrnSpd + offset(log(Years)) + offset(log_Seg_Length))
Derail_model23 <- glm.nb(DerailCount ~ log_VL_Count + log_VL_TrnSpd + Stn_Count + offset(log(Years)) + offset(log_Seg_Length))
```

```
# Linear combination of variables using years and segment length as offsets and Average Train Count
Derail_model24 <- glm.nb(DerailCount ~ Avg_Train + offset(log(Years)) + offset(log_Seg_Length))
Derail_model25 <- glm.nb(DerailCount ~ Avg_Train + VL_TrnSpd + offset(log(Years)) + offset(log_Seg_Length))
Derail_model26 <- glm.nb(DerailCount ~ Avg_Train + VL_TrnSpd + Stn_Count + offset(log(Years)) + offset(log_Seg_Length))
Derail_model27 <- glm.nb(DerailCount ~ Avg_Train + log_VL_TrnSpd + offset(log(Years)) + offset(log_Seg_Length))
Derail_model28 <- glm.nb(DerailCount ~ Avg_Train + log_VL_TrnSpd + Stn_Count + offset(log(Years)) + offset(log_Seg_Length))

# Linear combination of variables using years as offset, Max Train Count as exposure and log_TrnSpd_VLCount as interaction variable
Derail_model29 <- glm.nb(DerailCount ~ log_VL_Count + offset(log(Years)) + log_TrnSpd_VLCount)
Derail_model30 <- glm.nb(DerailCount ~ log_VL_Count + VL_TrnSpd + offset(log(Years)) + log_TrnSpd_VLCount)
Derail_model31 <- glm.nb(DerailCount ~ log_VL_Count + VL_TrnSpd + Stn_Count + offset(log(Years)) + log_TrnSpd_VLCount)
Derail_model32 <- glm.nb(DerailCount ~ log_VL_Count + VL_TrnSpd + Stn_Count + Seg_Length + offset(log(Years)) + log_TrnSpd_VLCount)
Derail_model33 <- glm.nb(DerailCount ~ log_VL_Count + log_VL_TrnSpd + offset(log(Years)) + log_TrnSpd_VLCount)
Derail_model34 <- glm.nb(DerailCount ~ log_VL_Count + log_VL_TrnSpd + Stn_Count + offset(log(Years))+ log_TrnSpd_VLCount)
Derail_model35 <- glm.nb(DerailCount ~ log_VL_Count + log_VL_TrnSpd + Stn_Count + Seg_Length + offset(log(Years)) + log_TrnSpd_VLCount)
Derail_model36 <- glm.nb(DerailCount ~ log_VL_Count + log_VL_TrnSpd + Stn_Count + log_Seg_Length + offset(log(Years)) + log_TrnSpd_VLCount)
Derail_model37 <- glm.nb(DerailCount ~ log_VL_Count + VL_TrnSpd + Stn_Count + log_Seg_Length + offset(log(Years)) + log_TrnSpd_VLCount)

# Linear combination of variables using years as offset, Max Train Count as exposure and log_TrnSpd_AvgTrain as interaction variable
Derail_model38 <- glm.nb(DerailCount ~ log_VL_Count + offset(log(Years)) + log_TrnSpd_AvgTrain)
Derail_model39 <- glm.nb(DerailCount ~ log_VL_Count + VL_TrnSpd + offset(log(Years)) + log_TrnSpd_AvgTrain)
Derail_model40 <- glm.nb(DerailCount ~ log_VL_Count + VL_TrnSpd + Stn_Count + offset(log(Years)) + log_TrnSpd_AvgTrain)
Derail_model41 <- glm.nb(DerailCount ~ log_VL_Count + VL_TrnSpd + Stn_Count + Seg_Length + offset(log(Years)) + log_TrnSpd_AvgTrain)
Derail_model42 <- glm.nb(DerailCount ~ log_VL_Count + log_VL_TrnSpd + offset(log(Years)) + log_TrnSpd_AvgTrain)
Derail_model43 <- glm.nb(DerailCount ~ log_VL_Count + log_VL_TrnSpd + Stn_Count + offset(log(Years)) + log_TrnSpd_AvgTrain)
Derail_model44 <- glm.nb(DerailCount ~ log_VL_Count + log_VL_TrnSpd + Stn_Count + Seg_Length + offset(log(Years)) + log_TrnSpd_AvgTrain)
Derail_model45 <- glm.nb(DerailCount ~ log_VL_Count + log_VL_TrnSpd + Stn_Count + log_Seg_Length + offset(log(Years)) + log_TrnSpd_AvgTrain)
Derail_model46 <- glm.nb(DerailCount ~ log_VL_Count + VL_TrnSpd + Stn_Count + log_Seg_Length + offset(log(Years)) + log_TrnSpd_AvgTrain)
```

```
# Linear combination of variables using years and segment length as offsets, max Train Count as exposure and log_TrnSpd_VLCount as interaction variable
Derail_model47 <- glm.nb(DerailCount ~ log_VL_Count + offset(log(Years)) + offset(log_Seg_Length) + log_TrnSpd_VLCount)
Derail_model48 <- glm.nb(DerailCount ~ log_VL_Count + VL_TrnSpd + offset(log(Years)) + offset(log_Seg_Length) + log_TrnSpd_VLCount)
Derail_model49 <- glm.nb(DerailCount ~ log_VL_Count + VL_TrnSpd + Stn_Count + offset(log(Years)) + offset(log_Seg_Length) + log_TrnSpd_VLCount)
Derail_model50 <- glm.nb(DerailCount ~ log_VL_Count + log_VL_TrnSpd + offset(log(Years)) + offset(log_Seg_Length) + log_TrnSpd_VLCount)
Derail_model51 <- glm.nb(DerailCount ~ log_VL_Count + log_VL_TrnSpd + Stn_Count + offset(log(Years)) + offset(log_Seg_Length) + log_TrnSpd_VLCount)

# Linear combination of variables using years and segment length as offsets, Average Train Count as exposure and log_TrnSpd_VLCount as interaction variable
Derail_model52 <- glm.nb(DerailCount ~ Avg_Train + offset(log(Years)) + offset(log_Seg_Length) + log_TrnSpd_VLCount)
Derail_model53 <- glm.nb(DerailCount ~ Avg_Train + VL_TrnSpd + offset(log(Years)) + offset(log_Seg_Length) + log_TrnSpd_VLCount)
Derail_model54 <- glm.nb(DerailCount ~ Avg_Train + VL_TrnSpd + Stn_Count + offset(log(Years)) + offset(log_Seg_Length) + log_TrnSpd_VLCount)
Derail_model55 <- glm.nb(DerailCount ~ Avg_Train + log_VL_TrnSpd + offset(log(Years)) + offset(log_Seg_Length) + log_TrnSpd_VLCount)
Derail_model56 <- glm.nb(DerailCount ~ Avg_Train + log_VL_TrnSpd + Stn_Count + offset(log(Years)) + offset(log_Seg_Length) + log_TrnSpd_VLCount)

#Square of train speed or train count and log_TrnSpd_VLCount as interaction variable
Derail_model57 <- glm.nb(DerailCount ~ log_VL_Count + log_VL_TrnSpd + Stn_Count + Seg_Length + offset(log(Years)) + VLCount_StnCount)
Derail_model58 <- glm.nb(DerailCount ~ log_VL_Count + log_VL_TrnSpd + Stn_Count + offset(log_Seg_Length) + offset(log(Years)) + VLCount_StnCount)
Derail_model59 <- glm.nb(DerailCount ~ log_VL_Count + log_VL_TrnSpd + Stn_Count + offset(log_Seg_Length) + offset(log(Years)) + StnCount_SegLen)
Derail_model60 <- glm.nb(DerailCount ~ log_VL_Count + log_VL_TrnSpd + Stn_Count + Seg_Length + offset(log(Years)) + TrnSpd_StnCount)

#Square of train speed or train count and log_TrnSpd_AvgTrain as interaction variable
Derail_model61 <- glm.nb(DerailCount ~ log_VL_Count + VL_TrnSpd + Stn_Count + Seg_Length + offset(log(Years)) + VLCount_StnCount)
Derail_model62 <- glm.nb(DerailCount ~ log_VL_Count + VL_TrnSpd + Stn_Count + offset(log_Seg_Length) + offset(log(Years)) + VLCount_StnCount)
Derail_model63 <- glm.nb(DerailCount ~ log_VL_Count + VL_TrnSpd + Stn_Count + offset(log_Seg_Length) + offset(log(Years)) + StnCount_SegLen)
Derail_model64 <- glm.nb(DerailCount ~ log_VL_Count + VL_TrnSpd + Stn_Count + Seg_Length + offset(log(Years)) + TrnSpd_StnCount)

#Refined Models
Derail_model65 <- glm.nb(DerailCount ~ log_VL_Count + log_VL_TrnSpd + Stn_Count + log_Seg_Length + offset(log(Years)) + log_TrnSpd_VLCount + TrnSpd_StnCount)
Derail_model66 <- glm.nb(DerailCount ~ log_VL_Count + log_VL_TrnSpd + Stn_Count + log_Seg_Length + offset(log(Years)) + TrnSpd_VLCount + TrnSpd_StnCount)
Derail_model67 <- glm.nb(DerailCount ~ log_VL_Count + VL_TrnSpd + Stn_Count + Seg_Length + offset(log(Years)) + TrnSpd_VLCount)
Derail_model68 <- glm.nb(DerailCount ~ log_VL_Count + VL_TrnSpd + Stn_Count + Seg_Length + offset(log(Years)) + TrnSpd_VLCount + TrnSpd_StnCount)
```

```
Derail_model69 <- glm.nb(DerailCount ~ log_VL_Count + VL_TrnSpd + Stn_Count + Seg_Length + offset(log(Years)) + TrnSpd_VLCount)
Derail_model70 <- glm.nb(DerailCount ~ log_VL_Count + VL_TrnSpd + Stn_Count + Seg_Length + offset(log(Years)) + log_TrnSpd_VLCount)
Derail_model71 <- glm.nb(DerailCount ~ log_VL_Count + VL_TrnSpd + Stn_Count + Seg_Length + offset(log(Years)) + StnCount_SegLen + log(VLCount_SegLen))
Derail_model72 <- glm.nb(DerailCount ~ log_VL_Count + VL_TrnSpd + Stn_Count + Seg_Length + offset(log(Years)) + log(VLCount_SegLen))
```

Candidate Models for Canadian Pacific Railway

```
# Linear combination of variables using years as offset and Max Train Count as exposure
Derail_model1 <- glm.nb(DerailCount ~ log_VL_Count + offset(log(Years)))
Derail_model2 <- glm.nb(DerailCount ~ log_VL_Count + VL_TrnSpd + offset(log(Years)))
Derail_model3 <- glm.nb(DerailCount ~ log_VL_Count + VL_TrnSpd + Stn_Count + offset(log(Years)))
Derail_model4 <- glm.nb(DerailCount ~ log_VL_Count + VL_TrnSpd + Stn_Count + Seg_Length + offset(log(Years)))
Derail_model5 <- glm.nb(DerailCount ~ log_VL_Count + log_VL_TrnSpd + Stn_Count + offset(log(Years)))
Derail_model6 <- glm.nb(DerailCount ~ log_VL_Count + log_VL_TrnSpd + Stn_Count + Seg_Length + offset(log(Years)))
Derail_model7 <- glm.nb(DerailCount ~ log_VL_Count + log_VL_TrnSpd + Seg_Length + offset(log(Years)))
Derail_model8 <- glm.nb(DerailCount ~ log_VL_Count + log_VL_TrnSpd + log_Seg_Length + offset(log(Years)))
Derail_model9 <- glm.nb(DerailCount ~ log_VL_Count + VL_TrnSpd + log_Seg_Length + offset(log(Years)))

# Linear combination of variables using years as offset and Average Train Count as exposure
Derail_model10 <- glm.nb(DerailCount ~ Avg_Train + offset(log(Years)))
Derail_model11 <- glm.nb(DerailCount ~ Avg_Train + VL_TrnSpd + offset(log(Years)))
Derail_model12 <- glm.nb(DerailCount ~ Avg_Train + VL_TrnSpd + Stn_Count + offset(log(Years)))
Derail_model13 <- glm.nb(DerailCount ~ Avg_Train + VL_TrnSpd + Stn_Count + Seg_Length + offset(log(Years)))
Derail_model14 <- glm.nb(DerailCount ~ Avg_Train + log_VL_TrnSpd + Stn_Count + offset(log(Years)))
Derail_model15 <- glm.nb(DerailCount ~ Avg_Train + log_VL_TrnSpd + Stn_Count + offset(log(Years)))
Derail_model16 <- glm.nb(DerailCount ~ Avg_Train + log_VL_TrnSpd + Seg_Length + offset(log(Years)))
Derail_model17 <- glm.nb(DerailCount ~ Avg_Train + log_VL_TrnSpd + log_Seg_Length + offset(log(Years)))
Derail_model18 <- glm.nb(DerailCount ~ Avg_Train + VL_TrnSpd + log_Seg_Length + offset(log(Years)))

# Linear combination of variables using years and segment length as offsets and max Train Count
Derail_model19 <- glm.nb(DerailCount ~ log_VL_Count + offset(log(Years)) + offset(log_Seg_Length))
Derail_model20 <- glm.nb(DerailCount ~ log_VL_Count + VL_TrnSpd + offset(log(Years)) + offset(log_Seg_Length))
Derail_model21 <- glm.nb(DerailCount ~ log_VL_Count + VL_TrnSpd + Stn_Count + offset(log(Years)) + offset(log_Seg_Length))
Derail_model22 <- glm.nb(DerailCount ~ log_VL_Count + log_VL_TrnSpd + Stn_Count + offset(log(Years)) + offset(log_Seg_Length))
Derail_model23 <- glm.nb(DerailCount ~ log_VL_Count + log_VL_TrnSpd + Stn_Count + Seg_Length + offset(log(Years)) + offset(log_Seg_Length))
```

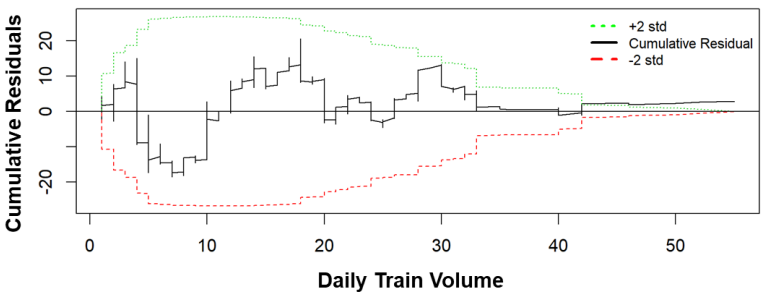
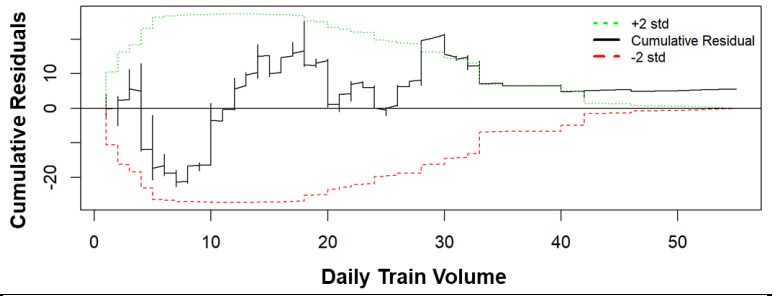
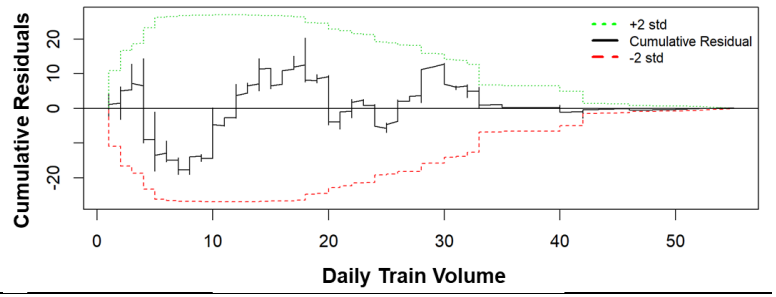
```
# Linear combination of variables using years and segment length as offsets and Average Train Count
Derail_model24 <- glm.nb(DerailCount ~ Avg_Train + offset(log(Years)) + offset(log_Seg_Length))
Derail_model25 <- glm.nb(DerailCount ~ Avg_Train + VL_TrnSpd + offset(log(Years)) + offset(log_Seg_Length))
Derail_model26 <- glm.nb(DerailCount ~ Avg_Train + VL_TrnSpd + Stn_Count + offset(log(Years)) + offset(log_Seg_Length))
Derail_model27 <- glm.nb(DerailCount ~ Avg_Train + log_VL_TrnSpd + Stn_Count + offset(log(Years)) + offset(log_Seg_Length))
Derail_model28 <- glm.nb(DerailCount ~ Avg_Train + log_VL_TrnSpd + Stn_Count + offset(log(Years)) + offset(log_Seg_Length))

# Linear combination of variables using years as offset, Max Train Count as exposure and log_TrnSpd_VLCount as interaction variable
Derail_model29 <- glm.nb(DerailCount ~ log_VL_Count + offset(log(Years)) + log_TrnSpd_VLCount)
Derail_model30 <- glm.nb(DerailCount ~ log_VL_Count + VL_TrnSpd + offset(log(Years)) + log_TrnSpd_VLCount)
Derail_model31 <- glm.nb(DerailCount ~ log_VL_Count + VL_TrnSpd + Stn_Count + offset(log(Years)) + log_TrnSpd_VLCount)
Derail_model32 <- glm.nb(DerailCount ~ log_VL_Count + VL_TrnSpd + Stn_Count + Seg_Length + offset(log(Years)) + log_TrnSpd_VLCount)
Derail_model33 <- glm.nb(DerailCount ~ log_VL_Count + log_VL_TrnSpd + Stn_Count + offset(log(Years)) + log_TrnSpd_VLCount)
Derail_model34 <- glm.nb(DerailCount ~ log_VL_Count + log_VL_TrnSpd + Stn_Count + Seg_Length + offset(log(Years)) + log_TrnSpd_VLCount)
Derail_model35 <- glm.nb(DerailCount ~ log_VL_Count + log_VL_TrnSpd + Seg_Length + offset(log(Years)) + log_TrnSpd_VLCount)
Derail_model36 <- glm.nb(DerailCount ~ log_VL_Count + log_VL_TrnSpd + log_Seg_Length + offset(log(Years)) + log_TrnSpd_VLCount)
Derail_model37 <- glm.nb(DerailCount ~ log_VL_Count + VL_TrnSpd + log_Seg_Length + offset(log(Years)) + log_TrnSpd_VLCount)

# Linear combination of variables using years as offset, Max Train Count as exposure and log_TrnSpd_AvgTrain as interaction variable
Derail_model38 <- glm.nb(DerailCount ~ log_VL_Count + offset(log(Years)) + log_TrnSpd_AvgTrain)
Derail_model39 <- glm.nb(DerailCount ~ log_VL_Count + VL_TrnSpd + offset(log(Years)) + log_TrnSpd_AvgTrain)
Derail_model40 <- glm.nb(DerailCount ~ log_VL_Count + VL_TrnSpd + Stn_Count + offset(log(Years)) + log_TrnSpd_AvgTrain)
Derail_model41 <- glm.nb(DerailCount ~ log_VL_Count + VL_TrnSpd + Stn_Count + Seg_Length + offset(log(Years)) + log_TrnSpd_AvgTrain)
Derail_model42 <- glm.nb(DerailCount ~ log_VL_Count + log_VL_TrnSpd + Stn_Count + Seg_Length + offset(log(Years)) + log_TrnSpd_AvgTrain)
Derail_model43 <- glm.nb(DerailCount ~ log_VL_Count + log_VL_TrnSpd + Stn_Count + offset(log(Years)) + log_TrnSpd_AvgTrain)
Derail_model44 <- glm.nb(DerailCount ~ log_VL_Count + log_VL_TrnSpd + Seg_Length + offset(log(Years)) + log_TrnSpd_AvgTrain)
Derail_model45 <- glm.nb(DerailCount ~ log_VL_Count + log_VL_TrnSpd + log_Seg_Length + offset(log(Years)) + log_TrnSpd_AvgTrain)
Derail_model46 <- glm.nb(DerailCount ~ log_VL_Count + VL_TrnSpd + log_Seg_Length + offset(log(Years)) + log_TrnSpd_AvgTrain)
# Linear combination of variables using years and segment length as offsets, max Train Count as exposure and log_TrnSpd_VLCount as interaction variable
Derail_model47 <- glm.nb(DerailCount ~ log_VL_Count + offset(log(Years)) + offset(log_Seg_Length) + log_TrnSpd_VLCount)
```

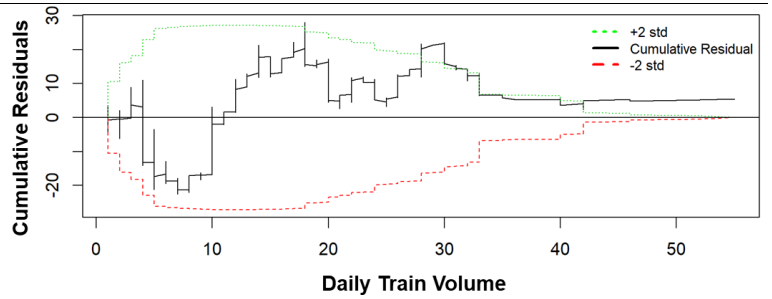
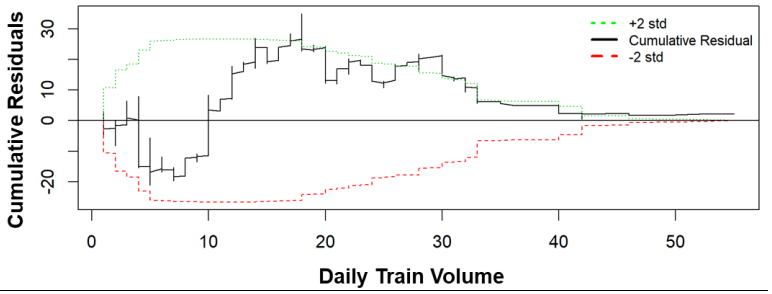
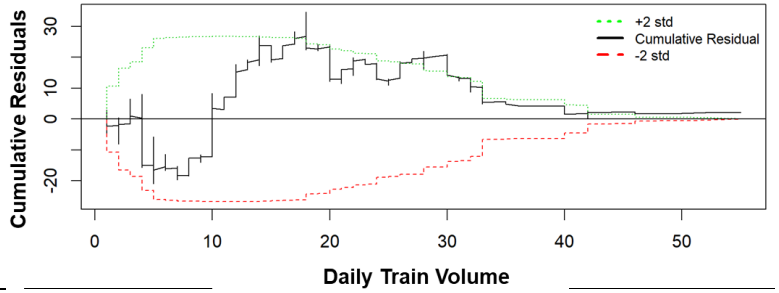
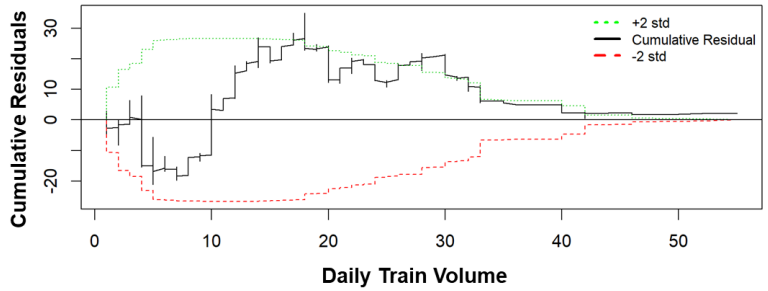
```
Derail_model48 <- glm.nb(DerailCount ~ log_VL_Count + VL_TrnSpd + offset(log(Years)) + offset(log_Seg_Length) + log_TrnSpd_VLCount)
Derail_model49 <- glm.nb(DerailCount ~ log_VL_Count + VL_TrnSpd + Stn_Count + offset(log(Years)) + offset(log_Seg_Length) + log_TrnSpd_VLCount)
Derail_model50 <- glm.nb(DerailCount ~ log_VL_Count + log_VL_TrnSpd + Stn_Count + offset(log(Years)) + offset(log_Seg_Length) + log_TrnSpd_VLCount)
Derail_model51 <- glm.nb(DerailCount ~ log_VL_Count + log_VL_TrnSpd + Stn_Count + Seg_Length + offset(log(Years)) + offset(log_Seg_Length) + log_TrnSpd_VLCount)

# Linear combination of variables using years and segment length as offsets, Average Train Count as exposure and log_TrnSpd_VLCount as interaction variable
Derail_model52 <- glm.nb(DerailCount ~ Avg_Train + offset(log(Years)) + offset(log_Seg_Length) + log_TrnSpd_VLCount)
Derail_model53 <- glm.nb(DerailCount ~ Avg_Train + VL_TrnSpd + offset(log(Years)) + offset(log_Seg_Length) + log_TrnSpd_VLCount)
Derail_model54 <- glm.nb(DerailCount ~ Avg_Train + VL_TrnSpd + Stn_Count + offset(log(Years)) + offset(log_Seg_Length) + log_TrnSpd_VLCount)
Derail_model55 <- glm.nb(DerailCount ~ Avg_Train + log_VL_TrnSpd + Stn_Count + offset(log(Years)) + offset(log_Seg_Length) + log_TrnSpd_VLCount)
Derail_model56 <- glm.nb(DerailCount ~ Avg_Train + log_VL_TrnSpd + Stn_Count + Seg_Length + offset(log(Years)) + offset(log_Seg_Length) + log_TrnSpd_VLCount)

#Square of train speed or train count
Derail_model57 <- glm.nb(DerailCount ~ log_VL_Count + VL_TrnSpd^2 + Seg_Length + offset(log(Years)))
Derail_model58 <- glm.nb(DerailCount ~ log_VL_Count + VL_Count^2 + VL_TrnSpd + Stn_Count + Seg_Length + offset(log(Years)))
Derail_model59 <- glm.nb(DerailCount ~ Avg_Train + VL_TrnSpd^2 + Seg_Length + offset(log(Years)))
Derail_model60 <- glm.nb(DerailCount ~ Avg_Train + Avg_Train^2 + VL_TrnSpd + Stn_Count + Seg_Length + offset(log(Years)))

#Square of train speed or train count and log_TrnSpd_VLCount as interaction variable
Derail_model57 <- glm.nb(DerailCount ~ log_VL_Count + VL_TrnSpd^2 + Seg_Length + offset(log(Years)) + log_TrnSpd_VLCount)
Derail_model58 <- glm.nb(DerailCount ~ log_VL_Count + VL_Count^2 + VL_TrnSpd + Stn_Count + Seg_Length + offset(log(Years)) + log_TrnSpd_VLCount)
Derail_model59 <- glm.nb(DerailCount ~ Avg_Train + VL_TrnSpd^2 + Seg_Length + offset(log(Years)) + log_TrnSpd_VLCount)
Derail_model60 <- glm.nb(DerailCount ~ Avg_Train + Avg_Train^2 + VL_TrnSpd + Stn_Count + Seg_Length + offset(log(Years)) + log_TrnSpd_VLCount)

#Square of train speed or train count and log_TrnSpd_AvgTrain as interaction variable
Derail_model61 <- glm.nb(DerailCount ~ log_VL_Count + VL_TrnSpd + Stn_Count + Seg_Length + offset(log(Years)) + VLCount_StnCount)
Derail_model62 <- glm.nb(DerailCount ~ log_VL_Count + VL_TrnSpd + Stn_Count + Seg_Length + offset(log(Years)) + VLCount_StnCount + StnCount_SegLen)
Derail_model63 <- glm.nb(DerailCount ~ log_VL_Count + VL_TrnSpd + Stn_Count + Seg_Length + offset(log(Years)) + TrnSpd_StnCount)
Derail_model64 <- glm.nb(DerailCount ~ log_VL_Count + VL_TrnSpd + Stn_Count + Seg_Length + offset(log(Years)) + StnCount_SegLen)
#Refined Models
```

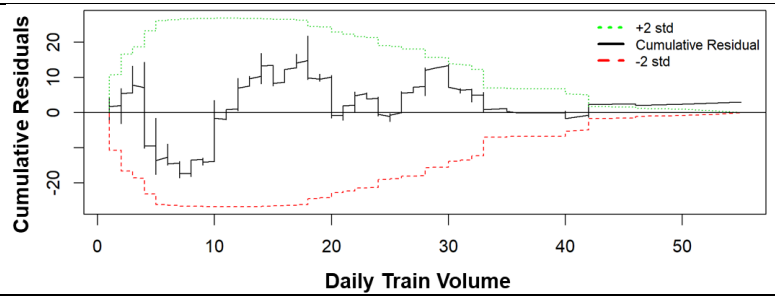
```
Derail_model65 <- glm.nb(DerailCount ~ log_VL_Count + VL_TrnSpd + Stn_Count + offset(log_Seg_Length) + offset(log(Years)) + VLCount_SegLen + log_TrnSpd_SegLen)
Derail_model66 <- glm.nb(DerailCount ~ log_VL_Count + VL_TrnSpd + Stn_Count + offset(log_Seg_Length) + offset(log(Years)) + log_VLCount_SegLen + log_TrnSpd_SegLen)
Derail_model67 <- glm.nb(DerailCount ~ log_VL_Count + VL_TrnSpd + Stn_Count + offset(log_Seg_Length) + offset(log(Years)) + log_VLCount_SegLen + TrnSpd_SegLen)
Derail_model68 <- glm.nb(DerailCount ~ log_VL_Count + VL_TrnSpd + Stn_Count + offset(log_Seg_Length) + offset(log(Years)) + TrnSpd_VLCount + TrnSpd_SegLen)
Derail_model69 <- glm.nb(DerailCount ~ log_VL_Count + VL_TrnSpd + Stn_Count + offset(log_Seg_Length) + offset(log(Years)) + StnCount_SegLen + TrnSpd_VLCount +
TrnSpd_SegLen)
Derail_model70 <- glm.nb(DerailCount ~ log_VL_Count + log_VL_TrnSpd + Stn_Count + offset(log_Seg_Length) + offset(log(Years)) + VLCount_StnCount)
Derail_model71 <- glm.nb(DerailCount ~ log_VL_Count + log_VL_TrnSpd + Stn_Count + offset(log_Seg_Length) + offset(log(Years)) + VLCount_StnCount + StnCount_SegLen)
Derail_model72 <- glm.nb(DerailCount ~ log_VL_Count + VL_TrnSpd + Stn_Count + offset(log_Seg_Length) + offset(log(Years)) + VLCount_StnCount)
Derail_model73 <- glm.nb(DerailCount ~ log_VL_Count + VL_TrnSpd + Stn_Count + offset(log_Seg_Length) + offset(log(Years)) + VLCount_StnCount + StnCount_SegLen)
```

Appendix B: Model Forms and Cure Plots for Shortlisted Models

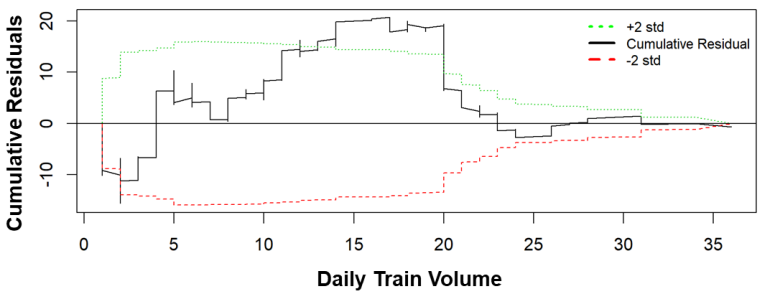
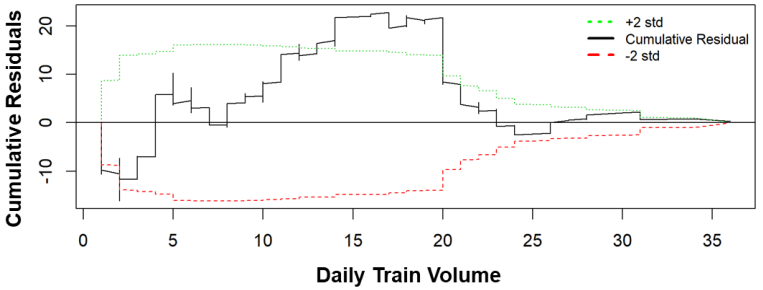
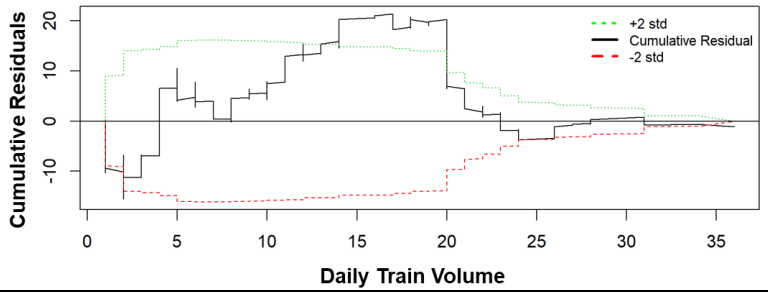
Shortlisted Models for Canada

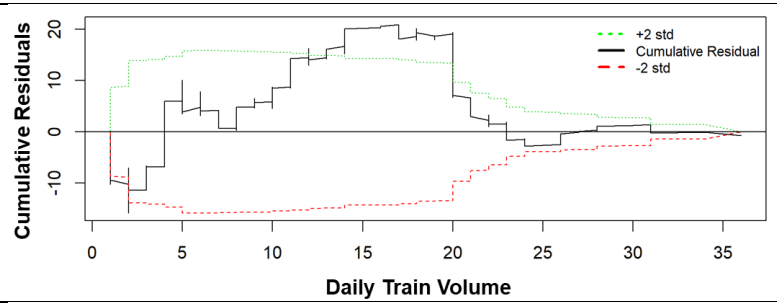
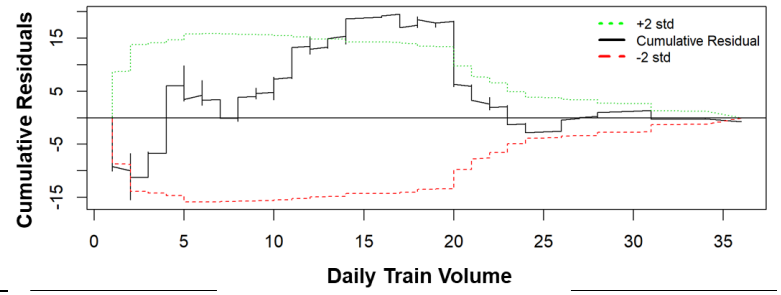
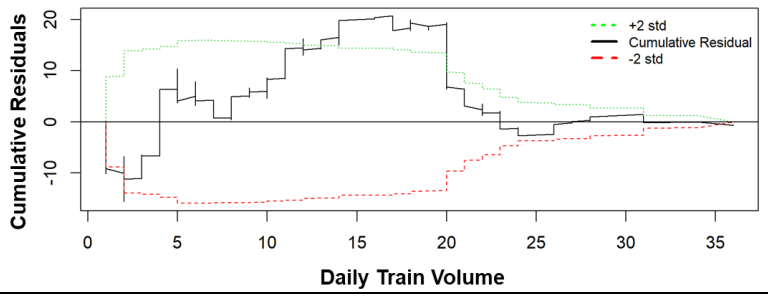
Model	Model Form	CURE Plot
4	$\mu_i = N \times exp(-5.7595) \times exp\begin{bmatrix}(0.3965 \times \text{logVL\_Count}) + (0.0104 \times \text{VL\_TrnSpd}) \\ +(-0.0013 \times \text{Stn\_Count}) + (0.0589 \times \text{Seg\_Length})\end{bmatrix}$	
8	$\mu_i = N \times exp(-9.7583) \times exp\begin{bmatrix}(0.2700 \times \text{logVL\_Count}) + (1.0039 \times \text{logVL\_TrnSpd}) \\ +(0.0039 \times \text{Stn\_Count}) + (0.7444 \times \text{logSeg\_Length})\end{bmatrix}$	
9	$\mu_i = N \times exp(-6.7009) \times exp\begin{bmatrix}(0.4119 \times \text{logVL\_Count}) + (0.0080 \times \text{VL\_TrnSpd}) \\ +(0.0187 \times \text{Stn\_Count}) + (0.7688 \times \text{logSeg\_Length})\end{bmatrix}$	

65	$\mu_i = N \times exp(-11.5383) \times exp\begin{bmatrix}(0.2715 \times logVL\_Count) + (1.4761 \times logVL\_TrnSpd) \\ +(0.7625 \times Stn\_Count) + (0.7389 \times logSeg\_Length) \\ +(-0.0159 \times (TrnSpd \times Stn\_Count))\end{bmatrix}$	
67	$\mu_i = N \times exp(-6.1591) \times exp\begin{bmatrix}(0.5651 \times logVL\_Count) + (0.0169 \times VL\_TrnSpd) \\ +(0.0054 \times Stn\_Count) + (0.0597 \times Seg\_Length) \\ +(-0.0004 \times (TrnSpd \times VL\_Count))\end{bmatrix}$	
68	$= N \times exp(-6.3235) \times exp\begin{bmatrix}(0.5468 \times logVL\_Count) + (0.0209 \times VL\_TrnSpd) \\ +(0.3162 \times Stn\_Count) + (0.0593 \times Seg\_Length) \\ +(-0.0004 \times (TrnSpd \times VL\_Count)) \\ +(-0.0066 \times (TrnSpd \times Stn\_Count))\end{bmatrix}$	
69	$= N \times exp(-6.1591) \times exp\begin{bmatrix}(0.5651 \times logVL\_Count) + (0.0169 \times VL\_TrnSpd) \\ +(0.0054 \times Stn\_Count) + (0.0597 \times Seg\_Length) \\ +(-0.0004 \times (TrnSpd \times VL\_Count))\end{bmatrix}$	

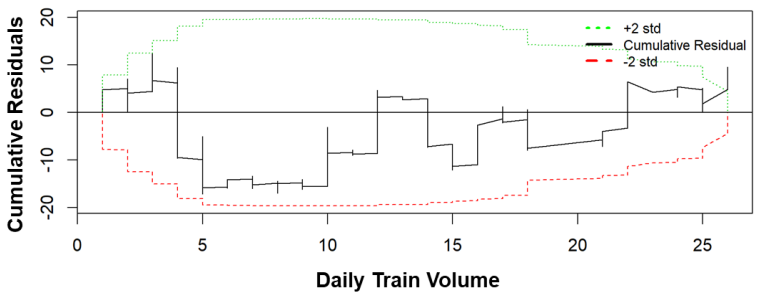
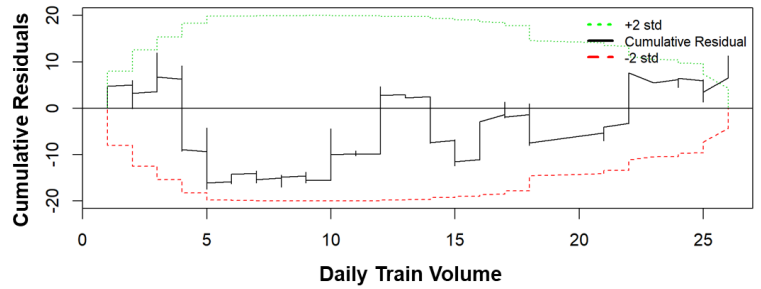
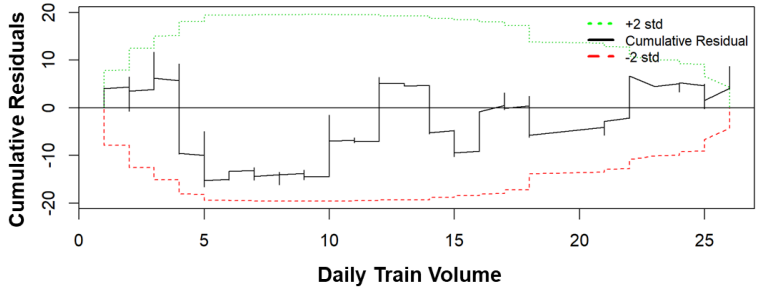
70	$= \mathrm{N} \times exp(-6.0010) \times exp \begin{bmatrix} (0.3929 \times \mathrm{logVL\_Count}) + (0.0160 \times \mathrm{VL\_TrnSpd}) \\ +(0.3725 \times \mathrm{Stn\_Count}) + (0.0585 \times \mathrm{Seg\_Length}) \\ +(-0.0080 \times (\mathrm{TrnSpd} \times \mathrm{Stn\_Count})) \end{bmatrix}$	

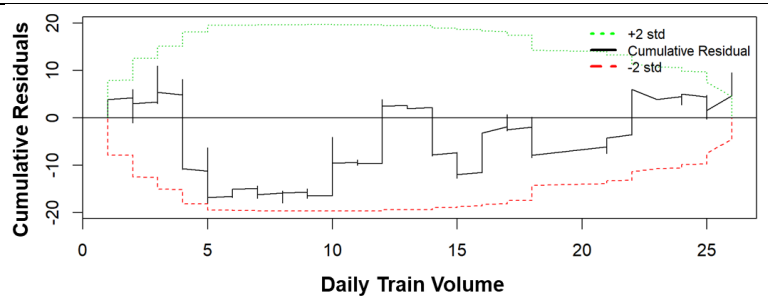
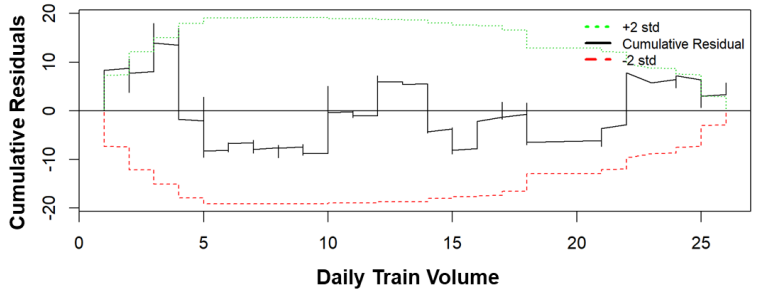
Shortlisted Models for Eastern Canada

Model	Model Form	CURE Plot
4	$\mu_i = N \times exp(-5.3219) \times exp\left[\begin{matrix}(0.0858 \times \text{logVL\_Count}) + (0.0139 \times \text{VL\_TrnSpd}) \\ +(-0.0989 \times \text{Stn\_Count}) + (0.0540 \times \text{Seg\_Length})\end{matrix}\right]$	
8	$\mu_i = N \times exp(-8.9046) \times exp\left[\begin{matrix}(0.0121 \times \text{logVL\_Count}) + (0.9718 \times \text{logVL\_TrnSpd}) \\ +(-0.0504 \times \text{Stn\_Count}) + (0.6154 \times \text{logSeg\_Length})\end{matrix}\right]$	
9	$\mu_i = N \times exp(-5.9773) \times exp\left[\begin{matrix}(0.1121 \times \text{logVL\_Count}) + (0.0108 \times \text{VL\_TrnSpd}) \\ +(-0.0550 \times \text{Stn\_Count}) + (0.6433 \times \text{logSeg\_Length})\end{matrix}\right]$	

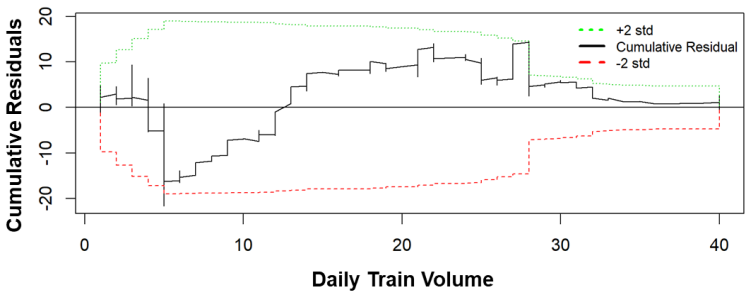
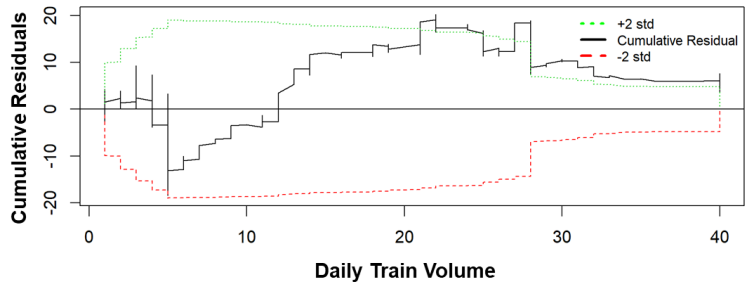
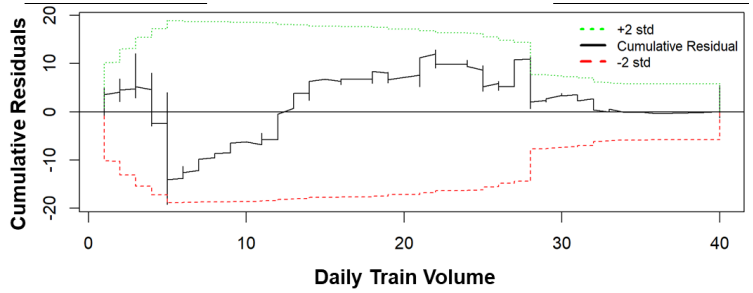
69	$\mu_i = N \times exp(-5.5095) \times exp\begin{bmatrix}(0.0769 \times \text{logVL\_Count}) + (0.0183 \times \text{VL\_TrnSpd}) \\ +(0.2365 \times \text{Stn\_Count}) + (0.0537 \times \text{Seg\_Length}) \\ +(-0.0068 \times (\text{Trn\_Spd} \times \text{Stn\_Count})\end{bmatrix}$	
71	$\mu_i = N \times exp(-5.2658) \times exp\begin{bmatrix}(0.0307 \times \text{logVL\_Count}) + (0.0141 \times \text{VL\_TrnSpd}) \\ +(-0.3504 \times \text{Stn\_Count}) + (0.0420 \times \text{Seg\_Length}) \\ +(0.0119 \times (\text{Stn\_Count} \times \text{Seg\_Length}) \\ +(0.0571 \times \log(\text{VL\_Count} \times \text{Seg\_Length})\end{bmatrix}$	
72	$\mu_i = N \times exp(-5.3171) \times exp\begin{bmatrix}(0.0896 \times \text{logVL\_Count}) + (0.0140 \times \text{VL\_TrnSpd}) \\ +(-0.0990 \times \text{Stn\_Count}) + (0.0543 \times \text{Seg\_Length}) \\ +(0.0041 \times \log(\text{VL\_Count} \times \text{Seg\_Length})\end{bmatrix}$	

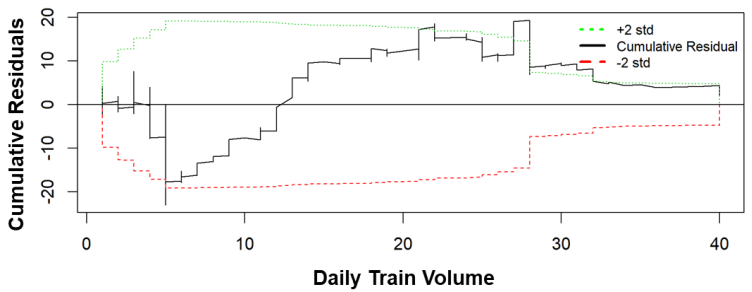
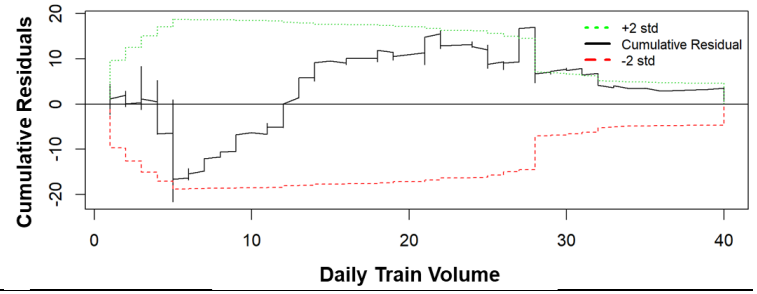
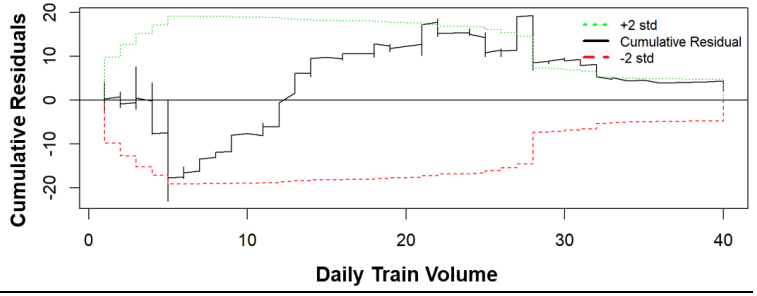
Shortlisted Models for Western Canada

Model	Model Form	CURE Plot
4	$\mu_i = N \times exp(-7.0619) \times exp\begin{bmatrix}(0.0200 \times \text{logVL\_Count}) + (0.0622 \times \text{VL\_TrnSpd}) \\ +(0.2256 \times \text{Stn\_Count}) + (0.0518 \times \text{Seg\_Length})\end{bmatrix}$	
9	$\mu_i = N \times exp(-8.0016) \times exp\begin{bmatrix}(0.0037 \times \text{logVL\_Count}) + (0.0628 \times \text{VL\_TrnSpd}) \\ +(0.2401 \times \text{Stn\_Count}) + (0.6921 \times \text{logSeg\_Length})\end{bmatrix}$	
30	$\mu_i = N \times exp(-7.1692) \times exp\begin{bmatrix}(0.0857 \times \text{logVL\_Count}) + (0.0621 \times \text{VL\_TrnSpd}) \\ +(0.2180 \times \text{Stn\_Count}) + (0.0563 \times \text{Seg\_Length}) \\ +(-0.0006 \times (\text{VL\_Count} \times \text{Seg\_Length}))\end{bmatrix}$	

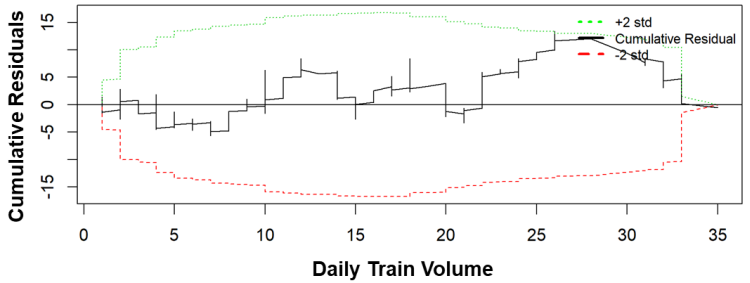
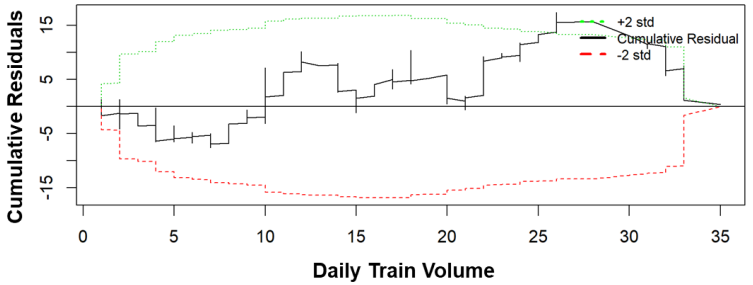
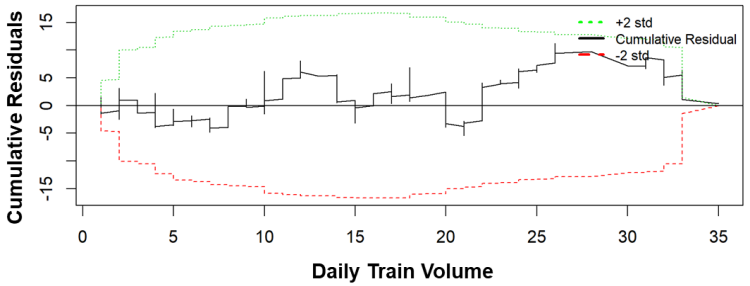
67	$\mu_i = N \times exp(-7.0338) \times exp \begin{bmatrix} (0.0439 \times \text{logAvg\_Train}) + (0.0606 \times \text{VL\_TrnSpd}) \\ +(0.2266 \times \text{Stn\_Count}) + (0.0521 \times \text{Seg\_Length}) \end{bmatrix}$	
70	$\mu_i = N \times exp(-7.2390) \times exp \begin{bmatrix} (0.0110 \times \text{logAvg\_Train}) + (0.0659 \times \text{VL\_TrnSpd}) \\ +(0.5587 \times \text{Stn\_Count}) + (0.0501 \times \text{Seg\_Length}) \\ +(-0.0590 \times (\text{Avg\_Train} \times \text{Stn\_Count})) \end{bmatrix}$	

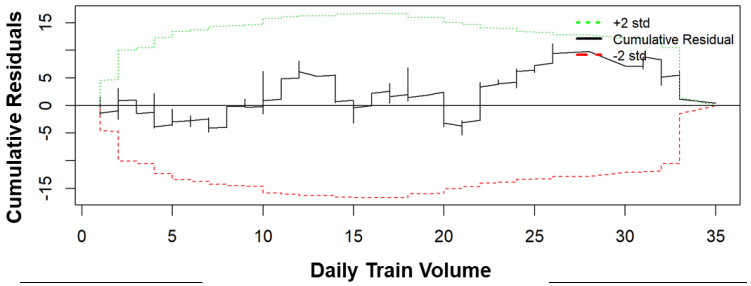
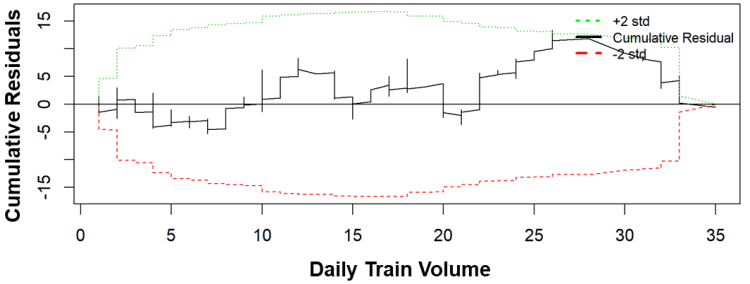
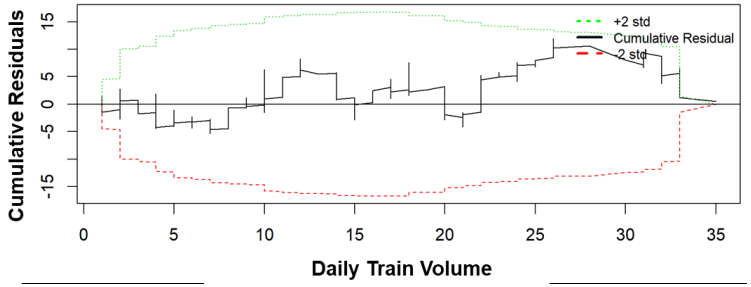
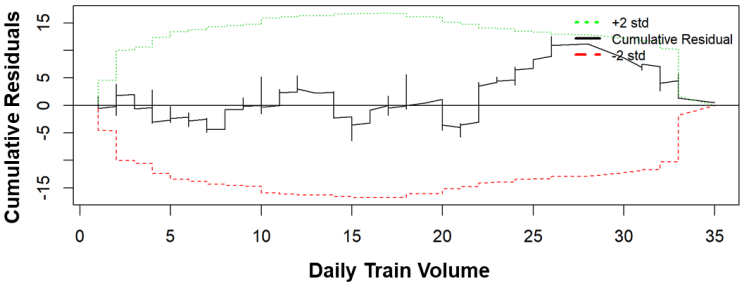
Shortlisted Models for Canadian National Railway

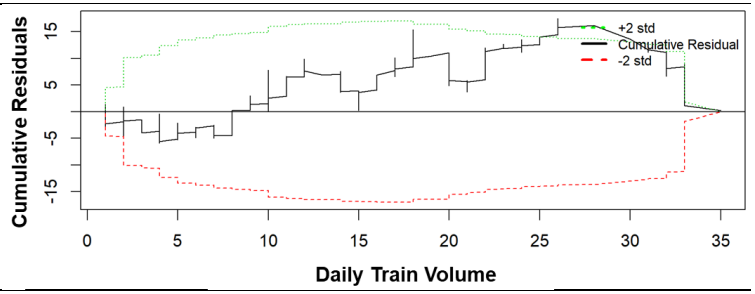
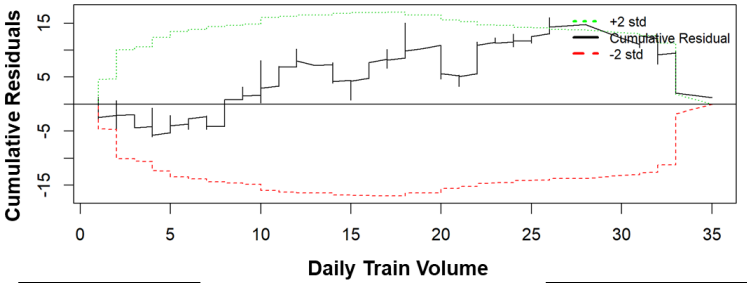
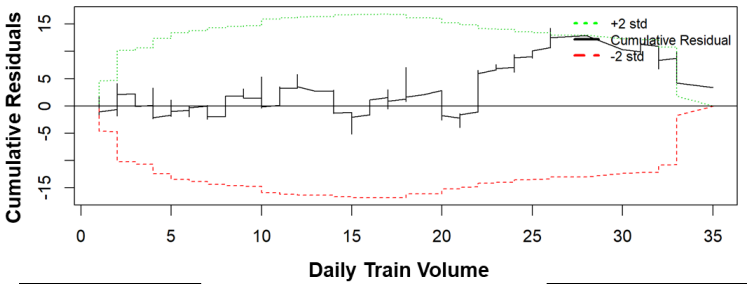
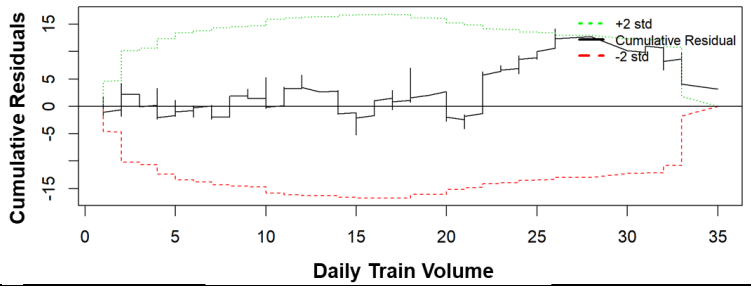
Model	Model Form	CURE Plot
61	$\mu_i = \mathrm{N} \times exp(-5.4929) \times \exp\begin{bmatrix}(0.0774 \times \mathrm{logVL\_Count}) + (0.0181 \times \mathrm{VL\_TrnSpd}) \\ +(0.3403 \times \mathrm{Stn\_Count}) + (0.0370 \times \mathrm{Seg\_Length}) \\ +(-0.0082 \times (\mathrm{VL\_Count} \times \mathrm{Stn\_Count}))\end{bmatrix}$	
62	$\mu_i = \mathrm{N} \times exp(-1.1020) \times \exp\begin{bmatrix}(0.0116 \times \mathrm{VL\_Count}) + (0.0161 \times \mathrm{VL\_TrnSpd}) \\ +(-0.1094 \times \mathrm{Stn\_Count}) + (0.0444 \times \mathrm{logSeg\_Length}) \\ +(-0.4152 \times \log(\mathrm{VL\_TrnSpd} \times \mathrm{VL\_Count}))\end{bmatrix}$	
63	$\mu_i = \mathrm{N} \times \mathrm{Seg\_Length} \times exp(-7.4553) \times \exp\begin{bmatrix}(0.1485 \times \mathrm{logVL\_Count}) + (0.0154 \times \mathrm{VL\_TrnSpd}) \\ +(0.2775 \times \mathrm{Stn\_Count}) \\ +(-0.0089 \times (\mathrm{VL\_Count} \times \mathrm{Stn\_Count}))\end{bmatrix}$	

Model	Model Form	CURE Plot
67	$\mu_i = \mathrm{N} \times exp(-5.6602) \times \exp\begin{bmatrix}(0.0870 \times \mathrm{logVL\_Count}) + (0.0274 \times \mathrm{VL\_TrnSpd}) \\ +(0.2193 \times \mathrm{Stn\_Count}) + (0.0364 \times \mathrm{Seg\_Length}) \\ +(-0.0004 \times (\mathrm{VL\_TrnSpd} \times \mathrm{VL\_Count}))\end{bmatrix}$	
68	$\mu_i = \mathrm{N} \times exp(-6.0451) \times \exp\begin{bmatrix}(0.0570 \times \mathrm{logVL\_Count}) + (0.0344 \times \mathrm{VL\_TrnSpd}) \\ +(0.5316 \times \mathrm{Stn\_Count}) + (0.0376 \times \mathrm{Seg\_Length}) \\ +\left(-0.0003 \times (\mathrm{VL\_TrnSpd} \times \mathrm{VL\_Count})\right) \\ +(-0.0060 \times (\mathrm{VL\_TrnSpd} \times \mathrm{Stn\_Count}))\end{bmatrix}$	
69	$\mu_i = \mathrm{N} \times exp(-5.6602) \times \exp\begin{bmatrix}(0.0870 \times \mathrm{logVL\_Count}) + (0.0274 \times \mathrm{VL\_TrnSpd}) \\ +(0.2193 \times \mathrm{Stn\_Count}) + (0.0364 \times \mathrm{Seg\_Length}) \\ +(-0.0004 \times (\mathrm{VL\_TrnSpd} \times \mathrm{VL\_Count}))\end{bmatrix}$	

Shortlisted Models for Canadian Pacific Railway

Model	Model Form	
4	$\mu_i = N \times exp(-5.4810) \times exp\begin{bmatrix}(0.4090 \times \text{logVL\_Count}) + (0.0061 \times \text{VL\_TrnSpd}) \\ +(0.0374 \times \text{Stn\_Count}) + (0.0607 \times \text{Seg\_Length})\end{bmatrix}$	
6	$\mu_i = N \times exp(-7.9624) \times exp\begin{bmatrix}(0.3034 \times \text{logVL\_Count}) + (0.7973 \times \text{logVL\_TrnSpd}) \\ +(0.0324 \times \text{Stn\_Count}) + (0.0598 \times \text{Seg\_Length})\end{bmatrix}$	
61	$\mu_i = N \times exp(-5.4520) \times \exp\begin{bmatrix}(0.4033 \times \text{logVL\_Count}) + (0.0069 \times \text{VL\_TrnSpd}) \\ +(-0.5734 \times \text{Stn\_Count}) + (0.0589 \times \text{Seg\_Length}) \\ +(0.0210 \times (\text{VL\_Count} \times \text{Stn\_Count}))\end{bmatrix}$	

Model	Model Form	
62	$\mu_i = \mathrm{N} \times exp(-5.4480) \times \exp\begin{bmatrix}(0.4053 \times \mathrm{logVL\_Count}) + (0.0068 \times \mathrm{VL\_TrnSpd}) \\ +(-0.5733 \times \mathrm{Stn\_Count}) + (0.0587 \times \mathrm{Seg\_Length}) \\ +(0.0193 \times (\mathrm{VL\_TrnSpd} \times \mathrm{Stn\_Count})) \\ +(0.0019 \times (\mathrm{Stn\_Count} \times \mathrm{Seg\_Length}))\end{bmatrix}$	
63	$\mu_i = \mathrm{N} \times exp(-5.4281) \times \exp\begin{bmatrix}(0.4324 \times \mathrm{logVL\_Count}) + (0.0040 \times \mathrm{VL\_TrnSpd}) \\ +(-0.2001 \times \mathrm{Stn\_Count}) + (0.0607 \times \mathrm{Seg\_Length}) \\ +(0.0035 \times (\mathrm{VL\_TrnSpd} \times \mathrm{Stn\_Count}))\end{bmatrix}$	
64	$\mu_i = \mathrm{N} \times exp(-5.4274) \times \exp\begin{bmatrix}(0.4214 \times \mathrm{logVL\_Count}) + (0.0057 \times \mathrm{VL\_TrnSpd}) \\ +(-0.3535 \times \mathrm{Stn\_Count}) + (0.0580 \times \mathrm{Seg\_Length}) \\ +(0.0152 \times (\mathrm{Stn\_Count} \times \mathrm{Seg\_Length}))\end{bmatrix}$	
65	$\mu_i = \mathrm{N} \times \mathrm{Seg\_Length} \times exp(-5.2232) \times \exp\begin{bmatrix}(0.2736 \times \mathrm{logVL\_Count}) + (0.0151 \times \mathrm{VL\_TrnSpd}) \\ +(0.0166 \times \mathrm{Stn\_Count}) \\ +(0.0008 \times (\mathrm{VL\_Count} \times \mathrm{Seg\_Length}) \\ +(-0.3347 \times \log(\mathrm{VL\_TrnSpd} \times \mathrm{Seg\_Length}))\end{bmatrix}$	

Model	Model Form	
68	$\mu_i = N \times Seg\_Length \times exp(-7.3212) \times \exp\left[\begin{matrix}(0.6288 \times logVL\_Count) + (0.0105 \times VL\_TrnSpd) \\ +(0.0723 \times Stn\_Count) \\ +(-0.0005 \times (VL\_TrnSpd \times VL\_Count)) \\ +(-0.0002 \times (VL\_TrnSpd \times Seg\_Length))\end{matrix}\right]$	
69	$\mu_i = N \times Seg\_Length \times exp(-7.3448) \times \exp\left[\begin{matrix}(0.6775 \times logVL\_Count) + (0.0113 \times VL\_TrnSpd) \\ +(-0.3727 \times Stn\_Count) \\ +(0.0176 \times (Stn\_Count \times Seg\_Length))) \\ +(-0.0006 \times (VL\_TrnSpd \times Seg\_Length))\end{matrix}\right]$	
72	$\mu_i = N \times Seg\_Length \times exp(-7.0022) \times \exp\left[\begin{matrix}(0.4216 \times logVL\_Count) + (0.0044 \times VL\_TrnSpd) \\ +(-0.5442 \times Stn\_Count) \\ +(0.0198 \times (VL\_Count \times Stn\_Count))\end{matrix}\right]$	
73	$\mu_i = N \times Seg\_Length \times exp(-7.0057) \times \exp\left[\begin{matrix}(0.4217 \times logVL\_Count) + (0.0047 \times VL\_TrnSpd) \\ +(-0.5414 \times Stn\_Count) \\ +(0.0232 \times (VL\_Count \times Stn\_Count)) \\ +(0.0239 \times (Stn\_Count \times Seg\_Length))\end{matrix}\right]$	

Appendix C: Variance Inflation Factors (VIFs) for Selected Model

Canada Model:

log_VL_Count	log_VL_TrnSpd	Stn_Count	Seg_Length
1.8345	1.8101	1.0316	1.0492

Eastern Canada Model:

log_VL_Count	VL_TrnSpd	Stn_Count	Seg_Length	TrnSpd_StnCount
1.5247	2.3094	11.8340	1.0246	12.9190

Western Canada Model:

log_VL_Count	VL_TrnSpd	Stn_Count	Seg_Length
2.1326	2.0859	1.0456	1.0856

CN Model:

log_VL_Count	VL_TrnSpd	Stn_Count	Seg_Length	TrnSpd_VLCount
3.2079	4.0531	1.2657	1.2326	5.5079

CP Model:

log_VL_Count	VL_TrnSpd	Stn_Count	Seg_Length	VLCount_StnCount	VLCount_StnCount
1.6886	1.7089	4.4762	1.0353	44.7730	4.4773

Note: Interaction terms can expect high VIFs since the variables are already part of the regression model functions.

Appendix D: Logit Models – Candidate Models

Candidate Models for Canada

```
# Linear combination of variables using years as offset and Max Train Count as exposure
logit_model1 <- glm(DERAIL ~ VL_Count + offset(log(Years)), family = binomial(link = "logit"))
logit_model2 <- glm(DERAIL ~ VL_Count + VL_TrnSpd + offset(log(Years)), family = binomial(link = "logit"))
logit_model3 <- glm(DERAIL ~ VL_Count + VL_TrnSpd + Stn_Count + offset(log(Years)), family = binomial(link = "logit"))
logit_model4 <- glm(DERAIL ~ VL_Count + VL_TrnSpd + Stn_Count + Seg_Length + offset(log(Years)), family = binomial(link = "logit"))
logit_model5 <- glm(DERAIL ~ VL_Count + log_VL_TrnSpd + offset(log(Years)), family = binomial(link = "logit"))
logit_model6 <- glm(DERAIL ~ VL_Count + log_VL_TrnSpd + Stn_Count + offset(log(Years)), family = binomial(link = "logit"))
logit_model7 <- glm(DERAIL ~ VL_Count + log_VL_TrnSpd + Stn_Count + Seg_Length + offset(log(Years)), family = binomial(link = "logit"))
logit_model8 <- glm(DERAIL ~ VL_Count + log_VL_TrnSpd + Stn_Count + log_Seg_Length + offset(log(Years)), family = binomial(link = "logit"))
logit_model9 <- glm(DERAIL ~ VL_Count + VL_TrnSpd + Stn_Count + log_Seg_Length + offset(log(Years)), family = binomial(link = "logit"))

# Linear combination of variables using years as offset and Average Train Count as exposure
logit_model10 <- glm(DERAIL ~ Avg_Train + offset(log(Years)), family = binomial(link = "logit"))
logit_model11 <- glm(DERAIL ~ Avg_Train + VL_TrnSpd + offset(log(Years)), family = binomial(link = "logit"))
logit_model12 <- glm(DERAIL ~ Avg_Train + VL_TrnSpd + Stn_Count + offset(log(Years)), family = binomial(link = "logit"))
logit_model13 <- glm(DERAIL ~ Avg_Train + VL_TrnSpd + Stn_Count + Seg_Length + offset(log(Years)), family = binomial(link = "logit"))
logit_model14 <- glm(DERAIL ~ Avg_Train + log_VL_TrnSpd + offset(log(Years)), family = binomial(link = "logit"))
logit_model15 <- glm(DERAIL ~ Avg_Train + log_VL_TrnSpd + Stn_Count + offset(log(Years)), family = binomial(link = "logit"))
logit_model16 <- glm(DERAIL ~ Avg_Train + log_VL_TrnSpd + Stn_Count + Seg_Length + offset(log(Years)), family = binomial(link = "logit"))
logit_model17 <- glm(DERAIL ~ Avg_Train + log_VL_TrnSpd + Stn_Count + log_Seg_Length + offset(log(Years)), family = binomial(link = "logit"))
logit_model18 <- glm(DERAIL ~ Avg_Train + VL_TrnSpd + Stn_Count + log_Seg_Length + offset(log(Years)), family = binomial(link = "logit"))

# Linear combination of variables using years and segment length as offsets and max Train Count
logit_model19 <- glm(DERAIL ~ VL_Count + offset(log(Years)) + offset(log_Seg_Length), family = binomial(link = "logit"))
logit_model20 <- glm(DERAIL ~ VL_Count + VL_TrnSpd + offset(log(Years)) + offset(log_Seg_Length), family = binomial(link = "logit"))
logit_model21 <- glm(DERAIL ~ VL_Count + VL_TrnSpd + Stn_Count + offset(log(Years)) + offset(log_Seg_Length), family = binomial(link = "logit"))
```

```
logit_model22 <- glm(DERAIL ~ VL_Count + log_VL_TrnSpd + offset(log(Years)) + offset(log_Seg_Length), family = binomial(link = "logit"))
logit_model23 <- glm(DERAIL ~ VL_Count + log_VL_TrnSpd + Stn_Count + offset(log(Years)) + offset(log_Seg_Length), family = binomial(link = "logit"))

# Linear combination of variables using years and segment length as offsets and Average Train Count
logit_model24 <- glm.nb (DERAIL ~ Avg_Train + offset(log(Years)) + offset(log_Seg_Length), family = binomial(link = "logit"))
logit_model25 <- glm(DERAIL ~ Avg_Train + VL_TrnSpd + offset(log(Years)) + offset(log_Seg_Length), family = binomial(link = "logit"))
logit_model26 <- glm(DERAIL ~ Avg_Train + VL_TrnSpd + Stn_Count + offset(log(Years)) + offset(log_Seg_Length), family = binomial(link = "logit"))
logit_model27 <- glm(DERAIL ~ Avg_Train + log_VL_TrnSpd + offset(log(Years)) + offset(log_Seg_Length), family = binomial(link = "logit"))
logit_model28 <- glm(DERAIL ~ Avg_Train + log_VL_TrnSpd + Stn_Count + offset(log(Years)) + offset(log_Seg_Length), family = binomial(link = "logit"))

# Linear combination of variables using years as offset, Max Train Count as exposure and log_TrnSpd_VLCount as interaction variable
logit_model29 <- glm(DERAIL ~ VL_Count + offset(log(Years)) + log_TrnSpd_VLCount, family = binomial(link = "logit"))
logit_model30 <- glm(DERAIL ~ VL_Count + VL_TrnSpd + offset(log(Years)) + log_TrnSpd_VLCount, family = binomial(link = "logit"))
logit_model31 <- glm(DERAIL ~ VL_Count + VL_TrnSpd + Stn_Count + offset(log(Years)) + log_TrnSpd_VLCount, family = binomial(link = "logit"))
logit_model32 <- glm(DERAIL ~ VL_Count + VL_TrnSpd + Stn_Count + Seg_Length + offset(log(Years)) + log_TrnSpd_VLCount, family = binomial(link = "logit"))
logit_model33 <- glm(DERAIL ~ VL_Count + log_VL_TrnSpd + offset(log(Years)) + log_TrnSpd_VLCount, family = binomial(link = "logit"))
logit_model34 <- glm(DERAIL ~ VL_Count + log_VL_TrnSpd + Stn_Count + offset(log(Years))+ log_TrnSpd_VLCount, family = binomial(link = "logit"))
logit_model35 <- glm(DERAIL ~ VL_Count + log_VL_TrnSpd + Stn_Count + Seg_Length + offset(log(Years)) + log_TrnSpd_VLCount, family = binomial(link = "logit"))
logit_model36 <- glm(DERAIL ~ VL_Count + log_VL_TrnSpd + Stn_Count + log_Seg_Length + offset(log(Years)) + log_TrnSpd_VLCount, family = binomial(link = "logit"))
logit_model37 <- glm(DERAIL ~ VL_Count + VL_TrnSpd + Stn_Count + log_Seg_Length + offset(log(Years)) + log_TrnSpd_VLCount, family = binomial(link = "logit"))

# Linear combination of variables using years as offset, Max Train Count as exposure and log_TrnSpd_AvgTrain as interaction variable
logit_model38 <- glm(DERAIL ~ VL_Count + offset(log(Years)) + log_TrnSpd_AvgTrain, family = binomial(link = "logit"))
logit_model39 <- glm(DERAIL ~ VL_Count + VL_TrnSpd + offset(log(Years)) + log_TrnSpd_AvgTrain, family = binomial(link = "logit"))
logit_model40 <- glm(DERAIL ~ VL_Count + VL_TrnSpd + Stn_Count + offset(log(Years)) + log_TrnSpd_AvgTrain, family = binomial(link = "logit"))
logit_model41 <- glm(DERAIL ~ VL_Count + VL_TrnSpd + Stn_Count + Seg_Length + offset(log(Years)) + log_TrnSpd_AvgTrain, family = binomial(link = "logit"))
logit_model42 <- glm(DERAIL ~ VL_Count + log_VL_TrnSpd + offset(log(Years)) + log_TrnSpd_AvgTrain, family = binomial(link = "logit"))
logit_model43 <- glm(DERAIL ~ VL_Count + log_VL_TrnSpd + Stn_Count + offset(log(Years)) + log_TrnSpd_AvgTrain, family = binomial(link = "logit"))
logit_model44 <- glm(DERAIL ~ VL_Count + log_VL_TrnSpd + Stn_Count + Seg_Length + offset(log(Years)) + log_TrnSpd_AvgTrain, family = binomial(link = "logit"))
logit_model45 <- glm(DERAIL ~ VL_Count + log_VL_TrnSpd + Stn_Count + log_Seg_Length + offset(log(Years)) + log_TrnSpd_AvgTrain, family = binomial(link = "logit"))
logit_model46 <- glm(DERAIL ~ VL_Count + VL_TrnSpd + Stn_Count + log_Seg_Length + offset(log(Years)) + log_TrnSpd_AvgTrain, family = binomial(link = "logit"))
```

```
# Linear combination of variables using years and segment length as offsets, max Train Count as exposure and log_TrnSpd_VLCount as interaction variable
logit_model47 <- glm(DERAIL ~ VL_Count + offset(log(Years)) + offset(log_Seg_Length) + log_TrnSpd_VLCount, family = binomial(link = "logit"))
logit_model48 <- glm(DERAIL ~ VL_Count + VL_TrnSpd + offset(log(Years)) + offset(log_Seg_Length) + log_TrnSpd_VLCount, family = binomial(link = "logit"))
logit_model49 <- glm(DERAIL ~ VL_Count + VL_TrnSpd + Stn_Count + offset(log(Years)) + offset(log_Seg_Length) + log_TrnSpd_VLCount, family = binomial(link = "logit"))
logit_model50 <- glm(DERAIL ~ VL_Count + log_VL_TrnSpd + offset(log(Years)) + offset(log_Seg_Length) + log_TrnSpd_VLCount, family = binomial(link = "logit"))
logit_model51 <- glm(DERAIL ~ VL_Count + log_VL_TrnSpd + Stn_Count + offset(log(Years)) + offset(log_Seg_Length) + log_TrnSpd_VLCount, family = binomial(link = "logit"))

# Linear combination of variables using years and segment length as offsets, Average Train Count as exposure and log_TrnSpd_VLCount as interaction variable
logit_model52 <- glm(DERAIL ~ Avg_Train + offset(log(Years)) + offset(log_Seg_Length) + log_TrnSpd_VLCount, family = binomial(link = "logit"))
logit_model53 <- glm(DERAIL ~ Avg_Train + VL_TrnSpd + offset(log(Years)) + offset(log_Seg_Length) + log_TrnSpd_VLCount, family = binomial(link = "logit"))
logit_model54 <- glm(DERAIL ~ Avg_Train + VL_TrnSpd + Stn_Count + offset(log(Years)) + offset(log_Seg_Length) + log_TrnSpd_VLCount, family = binomial(link = "logit"))
logit_model55 <- glm(DERAIL ~ Avg_Train + log_VL_TrnSpd + offset(log(Years)) + offset(log_Seg_Length) + log_TrnSpd_VLCount, family = binomial(link = "logit"))
logit_model56 <- glm(DERAIL ~ Avg_Train + log_VL_TrnSpd + Stn_Count + offset(log(Years)) + offset(log_Seg_Length) + log_TrnSpd_VLCount, family = binomial(link = "logit"))

#Square of train speed or train count
logit_model57 <- glm(DERAIL ~ VL_Count + VL_TrnSpd^2 + Stn_Count + Seg_Length + offset(log(Years)), family = binomial(link = "logit"))
logit_model58 <- glm(DERAIL ~ VL_Count + VL_Count^2 + VL_TrnSpd + Stn_Count + Seg_Length + offset(log(Years)), family = binomial(link = "logit"))
logit_model59 <- glm(DERAIL ~ Avg_Train + VL_TrnSpd^2 + Stn_Count + Seg_Length + offset(log(Years)), family = binomial(link = "logit"))
logit_model60 <- glm(DERAIL ~ Avg_Train + Avg_Train^2 + VL_TrnSpd + Stn_Count + Seg_Length + offset(log(Years)), family = binomial(link = "logit"))

#Square of train speed or train count and log_TrnSpd_VLCount as interaction variable
logit_model57 <- glm(DERAIL ~ VL_Count + VL_TrnSpd^2 + Stn_Count + Seg_Length + offset(log(Years)) + log_TrnSpd_VLCount, family = binomial(link = "logit"))
logit_model58 <- glm(DERAIL ~ VL_Count + VL_Count^2 + VL_TrnSpd + Stn_Count + Seg_Length + offset(log(Years)) + log_TrnSpd_VLCount, family = binomial(link = "logit"))
logit_model59 <- glm(DERAIL ~ Avg_Train + VL_TrnSpd^2 + Stn_Count + Seg_Length + offset(log(Years)) + log_TrnSpd_VLCount, family = binomial(link = "logit"))
logit_model60 <- glm(DERAIL ~ Avg_Train + Avg_Train^2 + VL_TrnSpd + Stn_Count + Seg_Length + offset(log(Years)) + log_TrnSpd_VLCount, family = binomial(link = "logit"))
vif(glm(DERAIL ~ Avg_Train + Avg_Train^2 + VL_TrnSpd + Stn_Count + Seg_Length + offset(log(Years)) + log_TrnSpd_VLCount), family = binomial(link = "logit"))
```

```
#Square of train speed or train count and log_TrnSpd_AvgTrain as interaction variable
logit_model61 <- glm(DERAIL ~ VL_Count + VL_TrnSpd^2 + Stn_Count + Seg_Length + offset(log(Years)) + log_TrnSpd_AvgTrain, family = binomial(link = "logit"))
logit_model62 <- glm(DERAIL ~ VL_Count + VL_Count^2 + VL_TrnSpd + Stn_Count + Seg_Length + offset(log(Years)) + log_TrnSpd_AvgTrain, family = binomial(link = "logit"))
logit_model63 <- glm(DERAIL ~ Avg_Train + VL_TrnSpd^2 + Stn_Count + Seg_Length + offset(log(Years)) + log_TrnSpd_AvgTrain, family = binomial(link = "logit"))
logit_model64 <- glm(DERAIL ~ Avg_Train + Avg_Train^2 + VL_TrnSpd + Stn_Count + Seg_Length + offset(log(Years)) + log_TrnSpd_AvgTrain, family = binomial(link = "logit"))

#Refined Models
logit_model65 <- glm(DERAIL ~ VL_Count + log_VL_TrnSpd + Stn_Count + log_Seg_Length + offset(log(Years)) + log_TrnSpd_VLCount + TrnSpd_StnCount, family = binomial(link = "logit"))
logit_model66 <- glm(DERAIL ~ VL_Count + log_VL_TrnSpd + Stn_Count + log_Seg_Length + offset(log(Years)) + TrnSpd_VLCount + TrnSpd_StnCount, family = binomial(link = "logit"))
logit_model67 <- glm(DERAIL ~ VL_Count + VL_TrnSpd + Stn_Count + Seg_Length + offset(log(Years)) + TrnSpd_VLCount, family = binomial(link = "logit"))
logit_model68 <- glm(DERAIL ~ VL_Count + VL_TrnSpd + Stn_Count + Seg_Length + offset(log(Years)) + TrnSpd_VLCount + TrnSpd_StnCount, family = binomial(link = "logit"))
logit_model69 <- glm(DERAIL ~ log_VL_Count + VL_TrnSpd + Stn_Count + Seg_Length + offset(log(Years)) + TrnSpd_VLCount, family = binomial(link = "logit"))
logit_model70 <- glm(DERAIL ~ log_VL_Count + VL_TrnSpd + Stn_Count + Seg_Length + offset(log(Years)) + log_TrnSpd_VLCount, family = binomial(link = "logit"))
logit_model71 <- glm(DERAIL ~ VL_Count + VL_TrnSpd + Stn_Count + offset(log(Seg_Length)) + offset(log(Years)) + log_TrnSpd_VLCount, family = binomial(link = "logit"))
```

Candidate Models for Eastern Canada

```
# Linear combination of variables using years as offset and Max Train Count as exposure
logit_model1 <- glm(DERAIL ~ VL_Count + offset(log(Years)), family = binomial(link = "logit"))
logit_model2 <- glm(DERAIL ~ VL_Count + VL_TrnSpd + offset(log(Years)), family = binomial(link = "logit"))
logit_model3 <- glm(DERAIL ~ VL_Count + VL_TrnSpd + Stn_Count + offset(log(Years)), family = binomial(link = "logit"))
logit_model4 <- glm(DERAIL ~ VL_Count + VL_TrnSpd + Stn_Count + Seg_Length + offset(log(Years)), family = binomial(link = "logit"))
logit_model5 <- glm(DERAIL ~ VL_Count + log_VL_TrnSpd + offset(log(Years)), family = binomial(link = "logit"))
logit_model6 <- glm(DERAIL ~ VL_Count + log_VL_TrnSpd + Stn_Count + offset(log(Years)), family = binomial(link = "logit"))
logit_model7 <- glm(DERAIL ~ VL_Count + log_VL_TrnSpd + Stn_Count + Seg_Length + offset(log(Years)), family = binomial(link = "logit"))
logit_model8 <- glm(DERAIL ~ VL_Count + log_VL_TrnSpd + Stn_Count + log_Seg_Length + offset(log(Years)), family = binomial(link = "logit"))
logit_model9 <- glm(DERAIL ~ VL_Count + VL_TrnSpd + Stn_Count + log_Seg_Length + offset(log(Years)), family = binomial(link = "logit"))

# Linear combination of variables using years as offset and Average Train Count as exposure
logit_model10 <- glm(DERAIL ~ Avg_Train + offset(log(Years)), family = binomial(link = "logit"))
logit_model11 <- glm(DERAIL ~ Avg_Train + VL_TrnSpd + offset(log(Years)), family = binomial(link = "logit"))
logit_model12 <- glm(DERAIL ~ Avg_Train + VL_TrnSpd + Stn_Count + offset(log(Years)), family = binomial(link = "logit"))
logit_model13 <- glm(DERAIL ~ Avg_Train + VL_TrnSpd + Stn_Count + Seg_Length + offset(log(Years)), family = binomial(link = "logit"))
logit_model14 <- glm(DERAIL ~ Avg_Train + log_VL_TrnSpd + offset(log(Years)), family = binomial(link = "logit"))
logit_model15 <- glm(DERAIL ~ Avg_Train + log_VL_TrnSpd + Stn_Count + offset(log(Years)), family = binomial(link = "logit"))
logit_model16 <- glm(DERAIL ~ Avg_Train + log_VL_TrnSpd + Stn_Count + Seg_Length + offset(log(Years)), family = binomial(link = "logit"))
logit_model17 <- glm(DERAIL ~ Avg_Train + log_VL_TrnSpd + Stn_Count + log_Seg_Length + offset(log(Years)), family = binomial(link = "logit"))
logit_model18 <- glm(DERAIL ~ Avg_Train + VL_TrnSpd + Stn_Count + log_Seg_Length + offset(log(Years)), family = binomial(link = "logit"))

# Linear combination of variables using years and segment length as offsets and max Train Count
logit_model19 <- glm(DERAIL ~ VL_Count + offset(log(Years)) + offset(log_Seg_Length), family = binomial(link = "logit"))
logit_model20 <- glm(DERAIL ~ VL_Count + VL_TrnSpd + offset(log(Years)) + offset(log_Seg_Length), family = binomial(link = "logit"))
logit_model21 <- glm(DERAIL ~ VL_Count + VL_TrnSpd + Stn_Count + offset(log(Years)) + offset(log_Seg_Length), family = binomial(link = "logit"))
logit_model22 <- glm(DERAIL ~ VL_Count + log_VL_TrnSpd + offset(log(Years)) + offset(log_Seg_Length), family = binomial(link = "logit"))
logit_model23 <- glm(DERAIL ~ VL_Count + log_VL_TrnSpd + Stn_Count + offset(log(Years)) + offset(log_Seg_Length), family = binomial(link = "logit"))
```

```
# Linear combination of variables using years and segment length as offsets and Average Train Count
logit_model24 <- glm(DERAIL ~ Avg_Train + offset(log(Years)) + offset(log_Seg_Length), family = binomial(link = "logit"))
logit_model25 <- glm(DERAIL ~ Avg_Train + VL_TrnSpd + offset(log(Years)) + offset(log_Seg_Length), family = binomial(link = "logit"))
logit_model26 <- glm(DERAIL ~ Avg_Train + VL_TrnSpd + Stn_Count + offset(log(Years)) + offset(log_Seg_Length), family = binomial(link = "logit"))
logit_model27 <- glm(DERAIL ~ Avg_Train + log_VL_TrnSpd + offset(log(Years)) + offset(log_Seg_Length), family = binomial(link = "logit"))
logit_model28 <- glm(DERAIL ~ Avg_Train + log_VL_TrnSpd + Stn_Count + offset(log(Years)) + offset(log_Seg_Length), family = binomial(link = "logit"))

# Linear combination of variables using years as offset, Max Train Count as exposure and log_TrnSpd_VLCount as interaction variable
logit_model29 <- glm(DERAIL ~ VL_Count + offset(log(Years)) + log_TrnSpd_VLCount, family = binomial(link = "logit"))
logit_model30 <- glm(DERAIL ~ VL_Count + VL_TrnSpd + offset(log(Years)) + log_TrnSpd_VLCount, family = binomial(link = "logit"))
logit_model31 <- glm(DERAIL ~ VL_Count + VL_TrnSpd + Stn_Count + offset(log(Years)) + log_TrnSpd_VLCount, family = binomial(link = "logit"))
logit_model32 <- glm(DERAIL ~ VL_Count + VL_TrnSpd + Stn_Count + Seg_Length + offset(log(Years)) + log_TrnSpd_VLCount, family = binomial(link = "logit"))
logit_model33 <- glm(DERAIL ~ VL_Count + log_VL_TrnSpd + offset(log(Years)) + log_TrnSpd_VLCount, family = binomial(link = "logit"))
logit_model34 <- glm(DERAIL ~ VL_Count + log_VL_TrnSpd + Stn_Count + offset(log(Years))+ log_TrnSpd_VLCount, family = binomial(link = "logit"))
logit_model35 <- glm(DERAIL ~ VL_Count + log_VL_TrnSpd + Stn_Count + Seg_Length + offset(log(Years)) + log_TrnSpd_VLCount, family = binomial(link = "logit"))
logit_model36 <- glm(DERAIL ~ VL_Count + log_VL_TrnSpd + Stn_Count + log_Seg_Length + offset(log(Years)) + log_TrnSpd_VLCount, family = binomial(link = "logit"))
logit_model37 <- glm(DERAIL ~ VL_Count + VL_TrnSpd + Stn_Count + log_Seg_Length + offset(log(Years)) + log_TrnSpd_VLCount, family = binomial(link = "logit"))

# Linear combination of variables using years as offset, Max Train Count as exposure and log_TrnSpd_AvgTrain as interaction variable
logit_model38 <- glm(DERAIL ~ VL_Count + offset(log(Years)) + log_TrnSpd_AvgTrain, family = binomial(link = "logit"))
logit_model39 <- glm(DERAIL ~ VL_Count + VL_TrnSpd + offset(log(Years)) + log_TrnSpd_AvgTrain, family = binomial(link = "logit"))
logit_model40 <- glm(DERAIL ~ VL_Count + VL_TrnSpd + Stn_Count + offset(log(Years)) + log_TrnSpd_AvgTrain, family = binomial(link = "logit"))
logit_model41 <- glm(DERAIL ~ VL_Count + VL_TrnSpd + Stn_Count + Seg_Length + offset(log(Years)) + log_TrnSpd_AvgTrain, family = binomial(link = "logit"))
logit_model42 <- glm(DERAIL ~ VL_Count + log_VL_TrnSpd + offset(log(Years)) + log_TrnSpd_AvgTrain, family = binomial(link = "logit"))
logit_model43 <- glm(DERAIL ~ VL_Count + log_VL_TrnSpd + Stn_Count + offset(log(Years)) + log_TrnSpd_AvgTrain, family = binomial(link = "logit"))
logit_model44 <- glm(DERAIL ~ VL_Count + log_VL_TrnSpd + Stn_Count + Seg_Length + offset(log(Years)) + log_TrnSpd_AvgTrain, family = binomial(link = "logit"))
logit_model45 <- glm(DERAIL ~ VL_Count + log_VL_TrnSpd + Stn_Count + log_Seg_Length + offset(log(Years)) + log_TrnSpd_AvgTrain, family = binomial(link = "logit"))
logit_model46 <- glm(DERAIL ~ VL_Count + VL_TrnSpd + Stn_Count + log_Seg_Length + offset(log(Years)) + log_TrnSpd_AvgTrain, family = binomial(link = "logit"))

# Linear combination of variables using years and segment length as offsets, max Train Count as exposure and log_TrnSpd_VLCount as interaction variable
```

```
logit_model47 <- glm(DERAIL ~ VL_Count + offset(log(Years)) + offset(log_Seg_Length) + log_TrnSpd_VLCount, family = binomial(link = "logit"))
logit_model48 <- glm(DERAIL ~ VL_Count + VL_TrnSpd + offset(log(Years)) + offset(log_Seg_Length) + log_TrnSpd_VLCount, family = binomial(link = "logit"))
logit_model49 <- glm(DERAIL ~ VL_Count + VL_TrnSpd + Stn_Count + offset(log(Years)) + offset(log_Seg_Length) + log_TrnSpd_VLCount, family = binomial(link = "logit"))
logit_model50 <- glm(DERAIL ~ VL_Count + log_VL_TrnSpd + offset(log(Years)) + offset(log_Seg_Length) + log_TrnSpd_VLCount, family = binomial(link = "logit"))
logit_model51 <- glm(DERAIL ~ VL_Count + log_VL_TrnSpd + Stn_Count + offset(log(Years)) + offset(log_Seg_Length) + log_TrnSpd_VLCount, family = binomial(link = "logit"))


# Linear combination of variables using years and segment length as offsets, Average Train Count as exposure and log_TrnSpd_VLCount as interaction variable
logit_model52 <- glm(DERAIL ~ Avg_Train + offset(log(Years)) + offset(log_Seg_Length) + log_TrnSpd_VLCount, family = binomial(link = "logit"))
logit_model53 <- glm(DERAIL ~ Avg_Train + VL_TrnSpd + offset(log(Years)) + offset(log_Seg_Length) + log_TrnSpd_VLCount, family = binomial(link = "logit"))
logit_model54 <- glm(DERAIL ~ Avg_Train + VL_TrnSpd + Stn_Count + offset(log(Years)) + offset(log_Seg_Length) + log_TrnSpd_VLCount, family = binomial(link = "logit"))
logit_model55 <- glm(DERAIL ~ Avg_Train + log_VL_TrnSpd + offset(log(Years)) + offset(log_Seg_Length) + log_TrnSpd_VLCount, family = binomial(link = "logit"))
logit_model56 <- glm(DERAIL ~ Avg_Train + log_VL_TrnSpd + Stn_Count + offset(log(Years)) + offset(log_Seg_Length) + log_TrnSpd_VLCount, family = binomial(link = "logit"))

#Square of train speed or train count
logit_model57 <- glm(DERAIL ~ VL_Count + VL_TrnSpd^2 + Stn_Count + Seg_Length + offset(log(Years)), family = binomial(link = "logit"))
logit_model58 <- glm(DERAIL ~ VL_Count + VL_Count^2 + VL_TrnSpd + Stn_Count + Seg_Length + offset(log(Years)), family = binomial(link = "logit"))
logit_model59 <- glm(DERAIL ~ Avg_Train + VL_TrnSpd^2 + Stn_Count + Seg_Length + offset(log(Years)), family = binomial(link = "logit"))
logit_model60 <- glm(DERAIL ~ Avg_Train + Avg_Train^2 + VL_TrnSpd + Stn_Count + Seg_Length + offset(log(Years)), family = binomial(link = "logit"))

#Square of train speed or train count and log_TrnSpd_VLCount as interaction variable
logit_model57 <- glm(DERAIL ~ VL_Count + VL_TrnSpd^2 + Stn_Count + Seg_Length + offset(log(Years)) + log_TrnSpd_VLCount, family = binomial(link = "logit"))
logit_model58 <- glm(DERAIL ~ VL_Count + VL_Count^2 + VL_TrnSpd + Stn_Count + Seg_Length + offset(log(Years)) + log_TrnSpd_VLCount, family = binomial(link = "logit"))
logit_model59 <- glm(DERAIL ~ Avg_Train + VL_TrnSpd^2 + Stn_Count + Seg_Length + offset(log(Years)) + log_TrnSpd_VLCount, family = binomial(link = "logit"))
logit_model60 <- glm(DERAIL ~ Avg_Train + Avg_Train^2 + VL_TrnSpd + Stn_Count + Seg_Length + offset(log(Years)) + log_TrnSpd_VLCount, family = binomial(link = "logit"))
```

```
#Square of train speed or train count and log_TrnSpd_AvgTrain as interaction variable
logit_model61 <- glm(DERAIL ~ VL_Count + VL_TrnSpd^2 + Stn_Count + Seg_Length + offset(log(Years)) + log_TrnSpd_AvgTrain, family = binomial(link = "logit"))
logit_model62 <- glm(DERAIL ~ VL_Count + VL_Count^2 + VL_TrnSpd + Stn_Count + Seg_Length + offset(log(Years)) + log_TrnSpd_AvgTrain, family = binomial(link = "logit"))
logit_model63 <- glm(DERAIL ~ Avg_Train + VL_TrnSpd^2 + Stn_Count + Seg_Length + offset(log(Years)) + log_TrnSpd_AvgTrain, family = binomial(link = "logit"))
logit_model64 <- glm(DERAIL ~ Avg_Train + Avg_Train^2 + VL_TrnSpd + Stn_Count + Seg_Length + offset(log(Years)) + log_TrnSpd_AvgTrain, family = binomial(link = "logit"))

#Model 32 Refined
logit_model65 <- glm(DERAIL ~ VL_Count + log_VL_TrnSpd + Stn_Count + log_Seg_Length + offset(log(Years)) + log_TrnSpd_VLCount + TrnSpd_StnCount, family = binomial(link = "logit"))
logit_model66 <- glm(DERAIL ~ VL_Count + log_VL_TrnSpd + Stn_Count + log_Seg_Length + offset(log(Years)) + TrnSpd_VLCount + TrnSpd_StnCount, family = binomial(link = "logit"))
logit_model67 <- glm(DERAIL ~ VL_Count + VL_TrnSpd + Stn_Count + Seg_Length + offset(log(Years)) + TrnSpd_VLCount, family = binomial(link = "logit"))
logit_model68 <- glm(DERAIL ~ VL_Count + VL_TrnSpd + Stn_Count + Seg_Length + offset(log(Years)) + TrnSpd_VLCount + TrnSpd_StnCount, family = binomial(link = "logit"))
logit_model69 <- glm(DERAIL ~ log_VL_Count + VL_TrnSpd + Stn_Count + Seg_Length + offset(log(Years)) + TrnSpd_VLCount, family = binomial(link = "logit"))
logit_model70 <- glm(DERAIL ~ log_VL_Count + VL_TrnSpd + Stn_Count + Seg_Length + offset(log(Years)) + log_TrnSpd_VLCount, family = binomial(link = "logit"))
logit_model71 <- glm(DERAIL ~ VL_Count + VL_TrnSpd + Stn_Count + Seg_Length + offset(log(Years)) + StnCount_SegLen + log(VLCount_SegLen), family = binomial(link = "logit"))
logit_model72 <- glm(DERAIL ~ VL_Count + VL_TrnSpd + Stn_Count + Seg_Length + offset(log(Years)) + log(VLCount_SegLen), family = binomial(link = "logit"))
```

Candidate Models for Western Canada

```
# Linear combination of variables using years as offset and Max Train Count as exposure
logit_model1 <- glm(DERAIL ~ log_VL_Count  + offset(log(Years)), family = binomial(link = "logit"))
logit_model2 <- glm(DERAIL ~ log_VL_Count  + VL_TrnSpd + offset(log(Years)), family = binomial(link = "logit"))
logit_model3 <- glm(DERAIL ~ log_VL_Count  + VL_TrnSpd + Stn_Count + offset(log(Years)), family = binomial(link = "logit"))
logit_model4 <- glm(DERAIL ~ log_VL_Count  + VL_TrnSpd + Stn_Count + Seg_Length + offset(log(Years)), family = binomial(link = "logit"))
logit_model5 <- glm(DERAIL ~ log_VL_Count  + log_VL_TrnSpd + offset(log(Years)), family = binomial(link = "logit"))
logit_model6 <- glm(DERAIL ~ log_VL_Count  + log_VL_TrnSpd + Stn_Count + offset(log(Years)), family = binomial(link = "logit"))
logit_model7 <- glm(DERAIL ~ log_VL_Count  + log_VL_TrnSpd + Stn_Count + Seg_Length + offset(log(Years)), family = binomial(link = "logit"))
logit_model8 <- glm(DERAIL ~ log_VL_Count  + log_VL_TrnSpd + Stn_Count + log_Seg_Length + offset(log(Years)), family = binomial(link = "logit"))
logit_model9 <- glm(DERAIL ~ log_VL_Count  + VL_TrnSpd + Stn_Count + log_Seg_Length + offset(log(Years)), family = binomial(link = "logit"))

# Linear combination of variables using years as offset and Average Train Count as exposure
logit_model10 <- glm(DERAIL ~ log_AvgTrain + offset(log(Years)), family = binomial(link = "logit"))
logit_model11 <- glm(DERAIL ~ log_AvgTrain + VL_TrnSpd + offset(log(Years)), family = binomial(link = "logit"))
logit_model12 <- glm(DERAIL ~ log_AvgTrain + VL_TrnSpd + Stn_Count + offset(log(Years)), family = binomial(link = "logit"))
logit_model13 <- glm(DERAIL ~ log_AvgTrain + VL_TrnSpd + Stn_Count + Seg_Length + offset(log(Years)), family = binomial(link = "logit"))
logit_model14 <- glm(DERAIL ~ log_AvgTrain + log_VL_TrnSpd + offset(log(Years)), family = binomial(link = "logit"))
logit_model15 <- glm(DERAIL ~ log_AvgTrain + log_VL_TrnSpd + Stn_Count + offset(log(Years)), family = binomial(link = "logit"))
logit_model16 <- glm(DERAIL ~ log_AvgTrain + log_VL_TrnSpd + Stn_Count + Seg_Length + offset(log(Years)), family = binomial(link = "logit"))
logit_model17 <- glm(DERAIL ~ log_AvgTrain + log_VL_TrnSpd + Stn_Count + log_Seg_Length + offset(log(Years)), family = binomial(link = "logit"))
logit_model18 <- glm(DERAIL ~ log_AvgTrain + VL_TrnSpd + Stn_Count + log_Seg_Length + offset(log(Years)), family = binomial(link = "logit"))

# Linear combination of variables using years and segment length as offsets and max Train Count
logit_model19 <- glm(DERAIL ~ log_VL_Count  + offset(log(Years)) + offset(log_Seg_Length), family = binomial(link = "logit"))
logit_model20 <- glm(DERAIL ~ log_VL_Count  + VL_TrnSpd + offset(log(Years)) + offset(log_Seg_Length), family = binomial(link = "logit"))
logit_model21 <- glm(DERAIL ~ log_VL_Count  + VL_TrnSpd + Stn_Count + offset(log(Years)) + offset(log_Seg_Length), family = binomial(link = "logit"))
logit_model22 <- glm(DERAIL ~ log_VL_Count  + log_VL_TrnSpd + offset(log(Years)) + offset(log_Seg_Length), family = binomial(link = "logit"))
logit_model23 <- glm(DERAIL ~ log_VL_Count  + log_VL_TrnSpd + Stn_Count + offset(log(Years)) + offset(log_Seg_Length), family = binomial(link = "logit"))
```

```
# Linear combination of variables using years and segment length as offsets and Average Train Count
logit_model24 <- glm(DERAIL ~ log_AvgTrain + offset(log(Years)) + offset(log_Seg_Length), family = binomial(link = "logit"))
logit_model25 <- glm(DERAIL ~ log_AvgTrain + VL_TrnSpd + offset(log(Years)) + offset(log_Seg_Length), family = binomial(link = "logit"))
logit_model26 <- glm(DERAIL ~ log_AvgTrain + VL_TrnSpd + Stn_Count + offset(log(Years)) + offset(log_Seg_Length), family = binomial(link = "logit"))
logit_model27 <- glm(DERAIL ~ log_AvgTrain + log_VL_TrnSpd + offset(log(Years)) + offset(log_Seg_Length), family = binomial(link = "logit"))
logit_model28 <- glm(DERAIL ~ log_AvgTrain + log_VL_TrnSpd + Stn_Count + offset(log(Years)) + offset(log_Seg_Length), family = binomial(link = "logit"))

# Linear combination of variables using years as offset, Max Train Count as exposure and TrnSpd_VLCount as interaction variable
logit_model29 <- glm(DERAIL ~ log_VL_Count  + VL_TrnSpd + Stn_Count + Seg_Length + offset(log(Years))+ log_TrnSpd_VLCount + log_VLCount_SegLen, family =
binomial(link = "logit"))
logit_model30 <- glm(DERAIL ~ log_VL_Count  + VL_TrnSpd + Stn_Count + Seg_Length + offset(log(Years)) + VLCount_SegLen, family = binomial(link = "logit"))
logit_model31 <- glm(DERAIL ~ log_VL_Count  + VL_TrnSpd + Stn_Count + offset(log(Years)) + log_TrnSpd_VLCount, family = binomial(link = "logit"))
logit_model32 <- glm(DERAIL ~ log_VL_Count  + VL_TrnSpd + Stn_Count + Seg_Length + offset(log(Years)) + log_TrnSpd_VLCount, family = binomial(link = "logit"))
logit_model33 <- glm(DERAIL ~ log_VL_Count  + log_VL_TrnSpd + offset(log(Years)) + log_TrnSpd_VLCount, family = binomial(link = "logit"))
logit_model34 <- glm(DERAIL ~ log_VL_Count  + log_VL_TrnSpd + Stn_Count + offset(log(Years))+ log_TrnSpd_VLCount, family = binomial(link = "logit"))
logit_model35 <- glm(DERAIL ~ log_VL_Count  + log_VL_TrnSpd + Stn_Count + Seg_Length + offset(log(Years)) + log_TrnSpd_VLCount, family = binomial(link = "logit"))
logit_model36 <- glm(DERAIL ~ log_VL_Count  + log_VL_TrnSpd + Stn_Count + log_Seg_Length + offset(log(Years)) + log_TrnSpd_VLCount, family = binomial(link = "logit"))
logit_model37 <- glm(DERAIL ~ log_VL_Count  + VL_TrnSpd + Stn_Count + log_Seg_Length + offset(log(Years)) + log_TrnSpd_VLCount, family = binomial(link = "logit"))

# Linear combination of variables using years as offset, Max Train Count as exposure and log_TrnSpd_AvgTrain as interaction variable
logit_model38 <- glm(DERAIL ~ log_VL_Count  + offset(log(Years)) + log_TrnSpd_AvgTrain, family = binomial(link = "logit"))
logit_model39 <- glm(DERAIL ~ log_VL_Count  + VL_TrnSpd + offset(log(Years)) + log_TrnSpd_AvgTrain, family = binomial(link = "logit"))
logit_model40 <- glm(DERAIL ~ log_VL_Count  + VL_TrnSpd + Stn_Count + offset(log(Years)) + log_TrnSpd_AvgTrain, family = binomial(link = "logit"))
logit_model41 <- glm(DERAIL ~ log_VL_Count  + VL_TrnSpd + Stn_Count + Seg_Length + offset(log(Years)) + log_TrnSpd_AvgTrain, family = binomial(link = "logit"))
logit_model42 <- glm(DERAIL ~ log_VL_Count  + log_VL_TrnSpd + offset(log(Years)) + log_TrnSpd_AvgTrain, family = binomial(link = "logit"))
logit_model43 <- glm(DERAIL ~ log_VL_Count  + log_VL_TrnSpd + Stn_Count + offset(log(Years)) + log_TrnSpd_AvgTrain, family = binomial(link = "logit"))
logit_model44 <- glm(DERAIL ~ log_VL_Count  + log_VL_TrnSpd + Stn_Count + Seg_Length + offset(log(Years)) + log_TrnSpd_AvgTrain, family = binomial(link = "logit"))
logit_model45 <- glm(DERAIL ~ log_VL_Count  + log_VL_TrnSpd + Stn_Count + log_Seg_Length + offset(log(Years)) + log_TrnSpd_AvgTrain, family = binomial(link = "logit"))
logit_model46 <- glm(DERAIL ~ log_VL_Count  + VL_TrnSpd + Stn_Count + log_Seg_Length + offset(log(Years)) + log_TrnSpd_AvgTrain, family = binomial(link = "logit"))
```

```
# Linear combination of variables using years and segment length as offsets, max Train Count as exposure and log_TrnSpd_VLCount as interaction variable
logit_model47 <- glm(DERAIL ~ log_VL_Count  + offset(log(Years)) + offset(log_Seg_Length) + log_TrnSpd_VLCount, family = binomial(link = "logit"))
logit_model48 <- glm(DERAIL ~ log_VL_Count  + VL_TrnSpd + offset(log(Years)) + offset(log_Seg_Length) + log_TrnSpd_VLCount, family = binomial(link = "logit"))
logit_model49 <- glm(DERAIL ~ log_VL_Count  + VL_TrnSpd + Stn_Count + offset(log(Years)) + offset(log_Seg_Length) + log_TrnSpd_VLCount, family = binomial(link = "logit"))
logit_model50 <- glm(DERAIL ~ log_VL_Count  + log_VL_TrnSpd + offset(log(Years)) + offset(log_Seg_Length) + log_TrnSpd_VLCount, family = binomial(link = "logit"))
logit_model51 <- glm(DERAIL ~ log_VL_Count  + log_VL_TrnSpd + Stn_Count + offset(log(Years)) + offset(log_Seg_Length) + log_TrnSpd_VLCount, family = binomial(link = "logit"))

 # Linear combination of variables using years and segment length as offsets, Average Train Count as exposure and log_TrnSpd_VLCount as interaction variable
logit_model52 <- glm(DERAIL ~ log_AvgTrain + offset(log(Years)) + offset(log_Seg_Length) + log_TrnSpd_VLCount, family = binomial(link = "logit"))
logit_model53 <- glm(DERAIL ~ log_AvgTrain + VL_TrnSpd + offset(log(Years)) + offset(log_Seg_Length) + log_TrnSpd_VLCount, family = binomial(link = "logit"))
logit_model54 <- glm(DERAIL ~ log_AvgTrain + VL_TrnSpd + Stn_Count + offset(log(Years)) + offset(log_Seg_Length) + log_TrnSpd_VLCount, family = binomial(link = "logit"))
logit_model55 <- glm(DERAIL ~ log_AvgTrain + log_VL_TrnSpd + offset(log(Years)) + offset(log_Seg_Length) + log_TrnSpd_VLCount, family = binomial(link = "logit"))
logit_model56 <- glm(DERAIL ~ log_AvgTrain + log_VL_TrnSpd + Stn_Count + offset(log(Years)) + offset(log_Seg_Length) + log_TrnSpd_VLCount, family = binomial(link = "logit"))

#Square of train speed or train count
logit_model57 <- glm(DERAIL ~ log_VL_Count  + VL_TrnSpd^2 + Stn_Count + Seg_Length + offset(log(Years)), family = binomial(link = "logit"))
logit_model58 <- glm(DERAIL ~ log_VL_Count  + VL_Count^2 + VL_TrnSpd + Stn_Count + Seg_Length + offset(log(Years)), family = binomial(link = "logit"))
logit_model59 <- glm(DERAIL ~ log_AvgTrain + VL_TrnSpd^2 + Stn_Count + Seg_Length + offset(log(Years)), family = binomial(link = "logit"))
logit_model60 <- glm(DERAIL ~ log_AvgTrain + Avg_Train^2 + VL_TrnSpd + Stn_Count + Seg_Length + offset(log(Years)), family = binomial(link = "logit"))

#Square of train speed or train count and log_TrnSpd_VLCount as interaction variable
logit_model57 <- glm(DERAIL ~ log_VL_Count  + VL_TrnSpd^2 + Stn_Count + Seg_Length + offset(log(Years)) + log_TrnSpd_VLCount, family = binomial(link = "logit"))
logit_model58 <- glm(DERAIL ~ log_VL_Count  + VL_Count^2 + VL_TrnSpd + Stn_Count + Seg_Length + offset(log(Years)) + log_TrnSpd_VLCount, family = binomial(link = "logit"))
logit_model59 <- glm(DERAIL ~ log_AvgTrain + VL_TrnSpd^2 + Stn_Count + Seg_Length + offset(log(Years)) + log_TrnSpd_AvgTrain, family = binomial(link = "logit"))
logit_model60 <- glm(DERAIL ~ log_AvgTrain + Avg_Train^2 + VL_TrnSpd + Stn_Count + Seg_Length + offset(log(Years)) + log_TrnSpd_AvgTrain, family = binomial(link = "logit"))
```

```
#Square of train speed or train count and log_TrnSpd_AvgTrain as interaction variable
logit_model61 <- glm(DERAIL ~ log_VL_Count  + VL_TrnSpd^2 + Stn_Count + Seg_Length + offset(log(Years)) + log_TrnSpd_AvgTrain, family = binomial(link = "logit"))
logit_model62 <- glm(DERAIL ~ log_VL_Count  + VL_Count^2 + VL_TrnSpd + Stn_Count + Seg_Length + offset(log(Years)) + log_TrnSpd_AvgTrain, family = binomial(link = "logit"))
logit_model63 <- glm(DERAIL ~ log_AvgTrain + VL_TrnSpd^2 + Stn_Count + Seg_Length + offset(log(Years)) + log_TrnSpd_AvgTrain, family = binomial(link = "logit"))
logit_model64 <- glm(DERAIL ~ log_AvgTrain + Avg_Train^2 + VL_TrnSpd + Stn_Count + Seg_Length + offset(log(Years)) + log_TrnSpd_AvgTrain, family = binomial(link = "logit"))

#Refined Models
logit_model65 <- glm(DERAIL ~ Avg_Train + VL_TrnSpd + Stn_Count + Seg_Length + offset(log(Years)), family = binomial(link = "logit"))
logit_model66 <- glm(DERAIL ~ Avg_Train + VL_TrnSpd + Stn_Count + Seg_Length + offset(log(Years)) + log_TrnSpd_AvgTrain, family = binomial(link = "logit"))
logit_model67 <- glm(DERAIL ~ log_AvgTrain + VL_TrnSpd + Stn_Count + Seg_Length + offset(log(Years)), family = binomial(link = "logit"))
logit_model68 <- glm(DERAIL ~ log_AvgTrain + VL_TrnSpd + Stn_Count + Seg_Length + offset(log(Years)) + log_TrnSpd_AvgTrain, family = binomial(link = "logit"))
logit_model69 <- glm(DERAIL ~ Avg_Train + VL_TrnSpd + Stn_Count + Seg_Length + offset(log(Years)) + AvgTrain_SegLen, family = binomial(link = "logit"))
logit_model70 <- glm(DERAIL ~ Avg_Train + VL_TrnSpd + Stn_Count + Seg_Length + offset(log(Years)) + AvgTrain_StnCount, family = binomial(link = "logit"))
logit_model71 <- glm(DERAIL ~ log_VL_Count  + VL_TrnSpd + Stn_Count + Seg_Length + offset(log(Years)) + StnCount_SegLen + log(VLCount_SegLen), family = binomial(link = "logit"))
logit_model72 <- glm(DERAIL ~ Avg_Train + VL_TrnSpd + Stn_Count + Seg_Length + offset(log(Years)) + AvgTrain_SegLen + log_TrnSpd_AvgTrain, family = binomial(link = "logit"))
```

Candidate Models for Canadian National Railway

```
# Linear combination of variables using years as offset and Max Train Count as exposure
logit_model1 <- glm(DERAIL ~ VL_Count + offset(log(Years)), family = binomial(link = "logit"))
logit_model2 <- glm(DERAIL ~ VL_Count + VL_TrnSpd + offset(log(Years)), family = binomial(link = "logit"))
logit_model3 <- glm(DERAIL ~ VL_Count + VL_TrnSpd + Stn_Count + offset(log(Years)), family = binomial(link = "logit"))
logit_model4 <- glm(DERAIL ~ VL_Count + VL_TrnSpd + Stn_Count + Seg_Length + offset(log(Years)), family = binomial(link = "logit"))
logit_model5 <- glm(DERAIL ~ VL_Count + log_VL_TrnSpd + offset(log(Years)), family = binomial(link = "logit"))
logit_model6 <- glm(DERAIL ~ VL_Count + log_VL_TrnSpd + Stn_Count + offset(log(Years)), family = binomial(link = "logit"))
logit_model7 <- glm(DERAIL ~ VL_Count + log_VL_TrnSpd + Stn_Count + Seg_Length + offset(log(Years)), family = binomial(link = "logit"))
logit_model8 <- glm(DERAIL ~ VL_Count + log_VL_TrnSpd + Stn_Count + log_Seg_Length + offset(log(Years)), family = binomial(link = "logit"))
logit_model9 <- glm(DERAIL ~ VL_Count + VL_TrnSpd + Stn_Count + log_Seg_Length + offset(log(Years)), family = binomial(link = "logit"))

# Linear combination of variables using years as offset and Average Train Count as exposure
logit_model10 <- glm(DERAIL ~ Avg_Train + offset(log(Years)), family = binomial(link = "logit"))
logit_model11 <- glm(DERAIL ~ Avg_Train + VL_TrnSpd + offset(log(Years)), family = binomial(link = "logit"))
logit_model12 <- glm(DERAIL ~ Avg_Train + VL_TrnSpd + Stn_Count + offset(log(Years)), family = binomial(link = "logit"))
logit_model13 <- glm(DERAIL ~ Avg_Train + VL_TrnSpd + Stn_Count + Seg_Length + offset(log(Years)), family = binomial(link = "logit"))
logit_model14 <- glm(DERAIL ~ Avg_Train + log_VL_TrnSpd + offset(log(Years)), family = binomial(link = "logit"))
logit_model15 <- glm(DERAIL ~ Avg_Train + log_VL_TrnSpd + Stn_Count + offset(log(Years)), family = binomial(link = "logit"))
logit_model16 <- glm(DERAIL ~ Avg_Train + log_VL_TrnSpd + Stn_Count + Seg_Length + offset(log(Years)), family = binomial(link = "logit"))
logit_model17 <- glm(DERAIL ~ Avg_Train + log_VL_TrnSpd + Stn_Count + log_Seg_Length + offset(log(Years)), family = binomial(link = "logit"))
logit_model18 <- glm(DERAIL ~ Avg_Train + VL_TrnSpd + Stn_Count + log_Seg_Length + offset(log(Years)), family = binomial(link = "logit"))

# Linear combination of variables using years and segment length as offsets and max Train Count
logit_model19 <- glm(DERAIL ~ VL_Count + offset(log(Years)) + offset(log_Seg_Length), family = binomial(link = "logit"))
logit_model20 <- glm(DERAIL ~ VL_Count + VL_TrnSpd + offset(log(Years)) + offset(log_Seg_Length), family = binomial(link = "logit"))
logit_model21 <- glm(DERAIL ~ VL_Count + VL_TrnSpd + Stn_Count + offset(log(Years)) + offset(log_Seg_Length), family = binomial(link = "logit"))
logit_model22 <- glm(DERAIL ~ VL_Count + log_VL_TrnSpd + offset(log(Years)) + offset(log_Seg_Length), family = binomial(link = "logit"))
logit_model23 <- glm(DERAIL ~ VL_Count + log_VL_TrnSpd + Stn_Count + offset(log(Years)) + offset(log_Seg_Length), family = binomial(link = "logit"))
```

```
# Linear combination of variables using years and segment length as offsets and Average Train Count
logit_model24 <- glm(DERAIL ~ Avg_Train + offset(log(Years)) + offset(log_Seg_Length), family = binomial(link = "logit"))
logit_model25 <- glm(DERAIL ~ Avg_Train + VL_TrnSpd + offset(log(Years)) + offset(log_Seg_Length), family = binomial(link = "logit"))
logit_model26 <- glm(DERAIL ~ Avg_Train + VL_TrnSpd + Stn_Count + offset(log(Years)) + offset(log_Seg_Length), family = binomial(link = "logit"))
logit_model27 <- glm(DERAIL ~ Avg_Train + log_VL_TrnSpd + offset(log(Years)) + offset(log_Seg_Length), family = binomial(link = "logit"))
logit_model28 <- glm(DERAIL ~ Avg_Train + log_VL_TrnSpd + Stn_Count + offset(log(Years)) + offset(log_Seg_Length), family = binomial(link = "logit"))

# Linear combination of variables using years as offset, Max Train Count as exposure and log_TrnSpd_VLCount as interaction variable
logit_model29 <- glm(DERAIL ~ VL_Count + offset(log(Years)) + log_TrnSpd_VLCount, family = binomial(link = "logit"))
logit_model30 <- glm(DERAIL ~ VL_Count + VL_TrnSpd + offset(log(Years)) + log_TrnSpd_VLCount, family = binomial(link = "logit"))
logit_model31 <- glm(DERAIL ~ VL_Count + VL_TrnSpd + Stn_Count + offset(log(Years)) + log_TrnSpd_VLCount, family = binomial(link = "logit"))
logit_model32 <- glm(DERAIL ~ VL_Count + VL_TrnSpd + Stn_Count + Seg_Length + offset(log(Years)) + log_TrnSpd_VLCount, family = binomial(link = "logit"))
logit_model33 <- glm(DERAIL ~ VL_Count + log_VL_TrnSpd + offset(log(Years)) + log_TrnSpd_VLCount, family = binomial(link = "logit"))
logit_model34 <- glm(DERAIL ~ VL_Count + log_VL_TrnSpd + Stn_Count + offset(log(Years))+ log_TrnSpd_VLCount, family = binomial(link = "logit"))
logit_model35 <- glm(DERAIL ~ VL_Count + log_VL_TrnSpd + Stn_Count + Seg_Length + offset(log(Years)) + log_TrnSpd_VLCount, family = binomial(link = "logit"))
logit_model36 <- glm(DERAIL ~ VL_Count + log_VL_TrnSpd + Stn_Count + log_Seg_Length + offset(log(Years)) + log_TrnSpd_VLCount, family = binomial(link = "logit"))
logit_model37 <- glm(DERAIL ~ VL_Count + VL_TrnSpd + Stn_Count + log_Seg_Length + offset(log(Years)) + log_TrnSpd_VLCount, family = binomial(link = "logit"))

# Linear combination of variables using years as offset, Max Train Count as exposure and log_TrnSpd_AvgTrain as interaction variable
logit_model38 <- glm(DERAIL ~ VL_Count + offset(log(Years)) + log_TrnSpd_AvgTrain, family = binomial(link = "logit"))
logit_model39 <- glm(DERAIL ~ VL_Count + VL_TrnSpd + offset(log(Years)) + log_TrnSpd_AvgTrain, family = binomial(link = "logit"))
logit_model40 <- glm(DERAIL ~ VL_Count + VL_TrnSpd + Stn_Count + offset(log(Years)) + log_TrnSpd_AvgTrain, family = binomial(link = "logit"))
logit_model41 <- glm(DERAIL ~ VL_Count + VL_TrnSpd + Stn_Count + Seg_Length + offset(log(Years)) + log_TrnSpd_AvgTrain, family = binomial(link = "logit"))
logit_model42 <- glm(DERAIL ~ VL_Count + log_VL_TrnSpd + offset(log(Years)) + log_TrnSpd_AvgTrain, family = binomial(link = "logit"))
logit_model43 <- glm(DERAIL ~ VL_Count + log_VL_TrnSpd + Stn_Count + offset(log(Years)) + log_TrnSpd_AvgTrain, family = binomial(link = "logit"))
logit_model44 <- glm(DERAIL ~ VL_Count + log_VL_TrnSpd + Stn_Count + Seg_Length + offset(log(Years)) + log_TrnSpd_AvgTrain, family = binomial(link = "logit"))
logit_model45 <- glm(DERAIL ~ VL_Count + log_VL_TrnSpd + Stn_Count + log_Seg_Length + offset(log(Years)) + log_TrnSpd_AvgTrain, family = binomial(link = "logit"))
logit_model46 <- glm(DERAIL ~ VL_Count + VL_TrnSpd + Stn_Count + log_Seg_Length + offset(log(Years)) + log_TrnSpd_AvgTrain, family = binomial(link = "logit"))
```

```
# Linear combination of variables using years and segment length as offsets, max Train Count as exposure and log_TrnSpd_VLCount as interaction variable
logit_model47 <- glm(DERAIL ~ VL_Count + offset(log(Years)) + offset(log_Seg_Length) + log_TrnSpd_VLCount, family = binomial(link = "logit"))
logit_model48 <- glm(DERAIL ~ VL_Count + VL_TrnSpd + offset(log(Years)) + offset(log_Seg_Length) + log_TrnSpd_VLCount, family = binomial(link = "logit"))
logit_model49 <- glm(DERAIL ~ VL_Count + VL_TrnSpd + Stn_Count + offset(log(Years)) + offset(log_Seg_Length) + log_TrnSpd_VLCount, family = binomial(link = "logit"))
logit_model50 <- glm(DERAIL ~ VL_Count + log_VL_TrnSpd + offset(log(Years)) + offset(log_Seg_Length) + log_TrnSpd_VLCount, family = binomial(link = "logit"))
logit_model51 <- glm(DERAIL ~ VL_Count + log_VL_TrnSpd + Stn_Count + offset(log(Years)) + offset(log_Seg_Length) + log_TrnSpd_VLCount, family = binomial(link = "logit"))

# Linear combination of variables using years and segment length as offsets, Average Train Count as exposure and log_TrnSpd_VLCount as interaction variable
logit_model52 <- glm(DERAIL ~ Avg_Train + offset(log(Years)) + offset(log_Seg_Length) + log_TrnSpd_VLCount, family = binomial(link = "logit"))
logit_model53 <- glm(DERAIL ~ Avg_Train + VL_TrnSpd + offset(log(Years)) + offset(log_Seg_Length) + log_TrnSpd_VLCount, family = binomial(link = "logit"))
logit_model54 <- glm(DERAIL ~ Avg_Train + VL_TrnSpd + Stn_Count + offset(log(Years)) + offset(log_Seg_Length) + log_TrnSpd_VLCount, family = binomial(link = "logit"))
logit_model55 <- glm(DERAIL ~ Avg_Train + log_VL_TrnSpd + offset(log(Years)) + offset(log_Seg_Length) + log_TrnSpd_VLCount, family = binomial(link = "logit"))
logit_model56 <- glm(DERAIL ~ Avg_Train + log_VL_TrnSpd + Stn_Count + offset(log(Years)) + offset(log_Seg_Length) + log_TrnSpd_VLCount, family = binomial(link = "logit"))

#Square of train speed or train count
logit_model57 <- glm(DERAIL ~ VL_Count + VL_TrnSpd^2 + Stn_Count + Seg_Length + offset(log(Years)), family = binomial(link = "logit"))
logit_model58 <- glm(DERAIL ~ VL_Count + VL_Count^2 + VL_TrnSpd + Stn_Count + Seg_Length + offset(log(Years)), family = binomial(link = "logit"))
logit_model59 <- glm(DERAIL ~ Avg_Train + VL_TrnSpd^2 + Stn_Count + Seg_Length + offset(log(Years)), family = binomial(link = "logit"))
logit_model60 <- glm(DERAIL ~ Avg_Train + Avg_Train^2 + VL_TrnSpd + Stn_Count + Seg_Length + offset(log(Years)), family = binomial(link = "logit"))

#Square of train speed or train count and log_TrnSpd_VLCount as interaction variable
logit_model57 <- glm(DERAIL ~ VL_Count + VL_TrnSpd^2 + Stn_Count + Seg_Length + offset(log(Years)) + log_TrnSpd_VLCount, family = binomial(link = "logit"))
logit_model58 <- glm(DERAIL ~ VL_Count + VL_Count^2 + VL_TrnSpd + Stn_Count + Seg_Length + offset(log(Years)) + log_TrnSpd_VLCount, family = binomial(link = "logit"))
logit_model59 <- glm(DERAIL ~ Avg_Train + VL_TrnSpd^2 + Stn_Count + Seg_Length + offset(log(Years)) + log_TrnSpd_VLCount, family = binomial(link = "logit"))
logit_model60 <- glm(DERAIL ~ Avg_Train + Avg_Train^2 + VL_TrnSpd + Stn_Count + Seg_Length + offset(log(Years)) + log_TrnSpd_VLCount, family = binomial(link = "logit"))
```

```
#Square of train speed or train count and log_TrnSpd_AvgTrain as interaction variable
logit_model61 <- glm(DERAIL ~ VL_Count + VL_TrnSpd^2 + Stn_Count + Seg_Length + offset(log(Years)) + log_TrnSpd_AvgTrain, family = binomial(link = "logit"))
logit_model62 <- glm(DERAIL ~ VL_Count + VL_Count^2 + VL_TrnSpd + Stn_Count + Seg_Length + offset(log(Years)) + log_TrnSpd_AvgTrain, family = binomial(link = "logit"))
logit_model63 <- glm(DERAIL ~ Avg_Train + VL_TrnSpd^2 + Stn_Count + Seg_Length + offset(log(Years)) + log_TrnSpd_AvgTrain, family = binomial(link = "logit"))
logit_model64 <- glm(DERAIL ~ Avg_Train + Avg_Train^2 + VL_TrnSpd + Stn_Count + Seg_Length + offset(log(Years)) + log_TrnSpd_AvgTrain, family = binomial(link = "logit"))

#Refined Models
logit_model65 <- glm(DERAIL ~ VL_Count + log_VL_TrnSpd + Stn_Count + log_Seg_Length + offset(log(Years)) + log_TrnSpd_VLCount + TrnSpd_StnCount, family = binomial(link = "logit"))
logit_model66 <- glm(DERAIL ~ VL_Count + log_VL_TrnSpd + Stn_Count + log_Seg_Length + offset(log(Years)) + TrnSpd_VLCount + TrnSpd_StnCount, family = binomial(link = "logit"))
logit_model67 <- glm(DERAIL ~ VL_Count + VL_TrnSpd + Stn_Count + Seg_Length + offset(log(Years)) + TrnSpd_VLCount, family = binomial(link = "logit"))
logit_model68 <- glm(DERAIL ~ VL_Count + VL_TrnSpd + Stn_Count + Seg_Length + offset(log(Years)) + TrnSpd_VLCount + TrnSpd_StnCount, family = binomial(link = "logit"))
logit_model69 <- glm(DERAIL ~ VL_Count + VL_TrnSpd + Stn_Count + Seg_Length + offset(log(Years)) + TrnSpd_VLCount, family = binomial(link = "logit"))
logit_model70 <- glm(DERAIL ~ VL_Count + VL_TrnSpd + Stn_Count + Seg_Length + offset(log(Years)) + log_TrnSpd_VLCount, family = binomial(link = "logit"))
logit_model71 <- glm(DERAIL ~ VL_Count + VL_TrnSpd + Stn_Count + Seg_Length + offset(log(Years)) + StnCount_SegLen + log(VLCount_SegLen), family = binomial(link = "logit"))
logit_model72 <- glm(DERAIL ~ VL_Count + VL_TrnSpd + Stn_Count + Seg_Length + offset(log(Years)) + log(VLCount_SegLen), family = binomial(link = "logit"))
```

Candidate Models for Canadian Pacific Railway

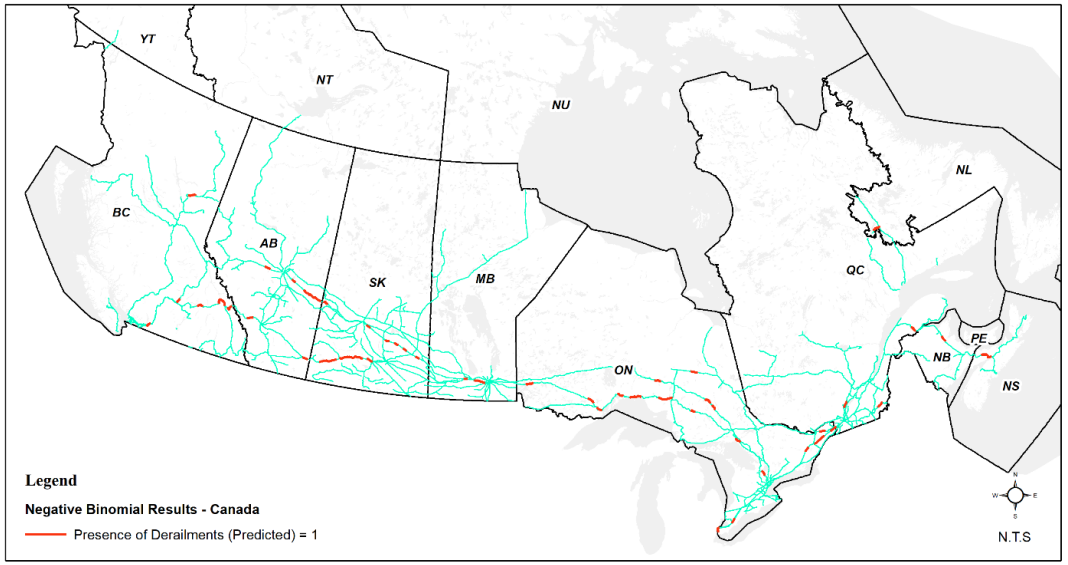
```
# Linear combination of variables using years as offset and Max Train Count as exposure
logit_model1 <- glm(DERAIL ~ VL_Count + offset(log(Years)), family = binomial(link = "logit"))
logit_model2 <- glm(DERAIL ~ VL_Count + VL_TrnSpd + offset(log(Years)), family = binomial(link = "logit"))
logit_model3 <- glm(DERAIL ~ VL_Count + VL_TrnSpd + Stn_Count + offset(log(Years)), family = binomial(link = "logit"))
logit_model4 <- glm(DERAIL ~ VL_Count + VL_TrnSpd + Stn_Count + Seg_Length + offset(log(Years)), family = binomial(link = "logit"))
logit_model5 <- glm(DERAIL ~ VL_Count + log_VL_TrnSpd + Stn_Count + offset(log(Years)), family = binomial(link = "logit"))
logit_model6 <- glm(DERAIL ~ VL_Count + log_VL_TrnSpd + Stn_Count + Seg_Length + offset(log(Years)), family = binomial(link = "logit"))
logit_model7 <- glm(DERAIL ~ VL_Count + log_VL_TrnSpd + Seg_Length + offset(log(Years)), family = binomial(link = "logit"))
logit_model8 <- glm(DERAIL ~ VL_Count + log_VL_TrnSpd + log_Seg_Length + offset(log(Years)), family = binomial(link = "logit"))
logit_model9 <- glm(DERAIL ~ VL_Count + VL_TrnSpd + log_Seg_Length + offset(log(Years)), family = binomial(link = "logit"))

# Linear combination of variables using years as offset and Average Train Count as exposure
logit_model10 <- glm(DERAIL ~ Avg_Train + offset(log(Years)), family = binomial(link = "logit"))
logit_model11 <- glm(DERAIL ~ Avg_Train + VL_TrnSpd + offset(log(Years)), family = binomial(link = "logit"))
logit_model12 <- glm(DERAIL ~ Avg_Train + VL_TrnSpd + Stn_Count + offset(log(Years)), family = binomial(link = "logit"))
logit_model13 <- glm(DERAIL ~ Avg_Train + VL_TrnSpd + Stn_Count + Seg_Length + offset(log(Years)), family = binomial(link = "logit"))
logit_model14 <- glm(DERAIL ~ Avg_Train + log_VL_TrnSpd + Stn_Count + offset(log(Years)), family = binomial(link = "logit"))
logit_model15 <- glm(DERAIL ~ Avg_Train + log_VL_TrnSpd + Stn_Count + offset(log(Years)), family = binomial(link = "logit"))
logit_model16 <- glm(DERAIL ~ Avg_Train + log_VL_TrnSpd + Seg_Length + offset(log(Years)), family = binomial(link = "logit"))
logit_model17 <- glm(DERAIL ~ Avg_Train + log_VL_TrnSpd + log_Seg_Length + offset(log(Years)), family = binomial(link = "logit"))
logit_model18 <- glm(DERAIL ~ Avg_Train + VL_TrnSpd + log_Seg_Length + offset(log(Years)), family = binomial(link = "logit"))

# Linear combination of variables using years and segment length as offsets and max Train Count
logit_model19 <- glm(DERAIL ~ VL_Count + offset(log(Years)) + offset(log_Seg_Length), family = binomial(link = "logit"))
logit_model20 <- glm(DERAIL ~ VL_Count + VL_TrnSpd + offset(log(Years)) + offset(log_Seg_Length), family = binomial(link = "logit"))
logit_model21 <- glm(DERAIL ~ VL_Count + VL_TrnSpd + Stn_Count + offset(log(Years)) + offset(log_Seg_Length), family = binomial(link = "logit"))
logit_model22 <- glm(DERAIL ~ VL_Count + log_VL_TrnSpd + Stn_Count + offset(log(Years)) + offset(log_Seg_Length), family = binomial(link = "logit"))
logit_model23 <- glm(DERAIL ~ VL_Count + log_VL_TrnSpd + Stn_Count + Seg_Length + offset(log(Years)) + offset(log_Seg_Length), family = binomial(link = "logit"))
```

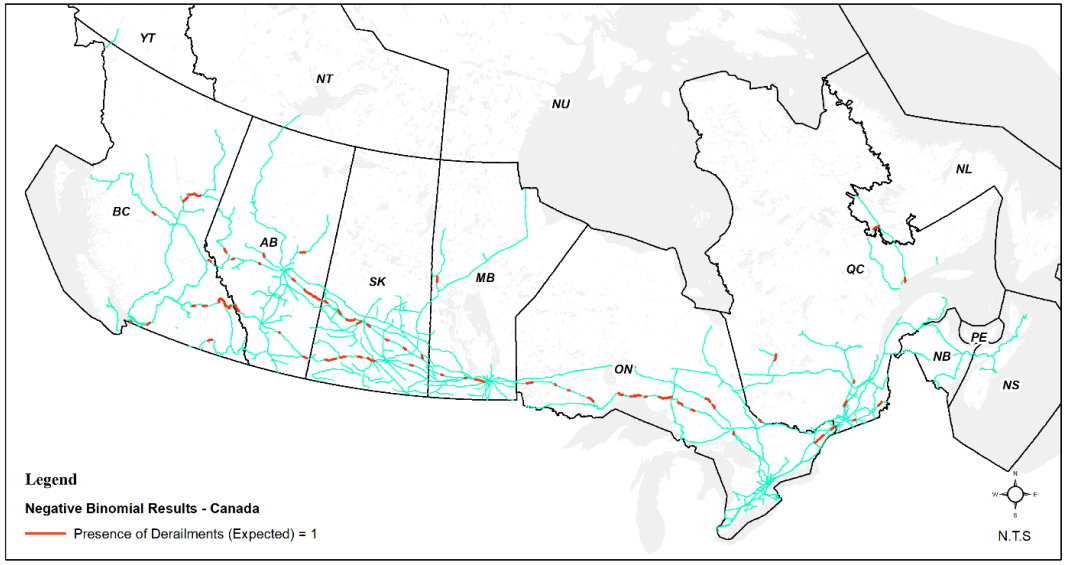
```
# Linear combination of variables using years and segment length as offsets and Average Train Count
logit_model24 <- glm(DERAIL ~ Avg_Train + offset(log(Years)) + offset(log_Seg_Length), family = binomial(link = "logit"))
logit_model25 <- glm(DERAIL ~ Avg_Train + VL_TrnSpd + offset(log(Years)) + offset(log_Seg_Length), family = binomial(link = "logit"))
logit_model26 <- glm(DERAIL ~ Avg_Train + VL_TrnSpd + Stn_Count + offset(log(Years)) + offset(log_Seg_Length), family = binomial(link = "logit"))
logit_model27 <- glm(DERAIL ~ Avg_Train + log_VL_TrnSpd + Stn_Count + offset(log(Years)) + offset(log_Seg_Length), family = binomial(link = "logit"))
logit_model28 <- glm(DERAIL ~ Avg_Train + log_VL_TrnSpd + Stn_Count + offset(log(Years)) + offset(log_Seg_Length), family = binomial(link = "logit"))

# Linear combination of variables using years as offset, Max Train Count as exposure and log_TrnSpd_VLCount as interaction variable
logit_model29 <- glm(DERAIL ~ VL_Count + offset(log(Years)) + log_TrnSpd_VLCount, family = binomial(link = "logit"))
logit_model30 <- glm(DERAIL ~ VL_Count + VL_TrnSpd + offset(log(Years)) + log_TrnSpd_VLCount, family = binomial(link = "logit"))
logit_model31 <- glm(DERAIL ~ VL_Count + VL_TrnSpd + Stn_Count + offset(log(Years)) + log_TrnSpd_VLCount, family = binomial(link = "logit"))
logit_model32 <- glm(DERAIL ~ VL_Count + VL_TrnSpd + Stn_Count + Seg_Length + offset(log(Years)) + log_TrnSpd_VLCount, family = binomial(link = "logit"))
logit_model33 <- glm(DERAIL ~ VL_Count + log_VL_TrnSpd + Stn_Count + offset(log(Years)) + log_TrnSpd_VLCount, family = binomial(link = "logit"))
logit_model34 <- glm(DERAIL ~ VL_Count + log_VL_TrnSpd + Stn_Count + Seg_Length + offset(log(Years)) + log_TrnSpd_VLCount, family = binomial(link = "logit"))
logit_model35 <- glm(DERAIL ~ VL_Count + log_VL_TrnSpd + Seg_Length + offset(log(Years)) + log_TrnSpd_VLCount, family = binomial(link = "logit"))
logit_model36 <- glm(DERAIL ~ VL_Count + log_VL_TrnSpd + log_Seg_Length + offset(log(Years)) + log_TrnSpd_VLCount, family = binomial(link = "logit"))
logit_model37 <- glm(DERAIL ~ VL_Count + VL_TrnSpd + log_Seg_Length + offset(log(Years)) + log_TrnSpd_VLCount, family = binomial(link = "logit"))

# Linear combination of variables using years as offset, Max Train Count as exposure and log_TrnSpd_AvgTrain as interaction variable
logit_model38 <- glm(DERAIL ~ VL_Count + offset(log(Years)) + log_TrnSpd_AvgTrain, family = binomial(link = "logit"))
logit_model39 <- glm(DERAIL ~ VL_Count + VL_TrnSpd + offset(log(Years)) + log_TrnSpd_AvgTrain, family = binomial(link = "logit"))
logit_model40 <- glm(DERAIL ~ VL_Count + VL_TrnSpd + Stn_Count + offset(log(Years)) + log_TrnSpd_AvgTrain, family = binomial(link = "logit"))
logit_model41 <- glm(DERAIL ~ VL_Count + VL_TrnSpd + Stn_Count + Seg_Length + offset(log(Years)) + log_TrnSpd_AvgTrain, family = binomial(link = "logit"))
logit_model42 <- glm(DERAIL ~ VL_Count + log_VL_TrnSpd + Stn_Count + Seg_Length + offset(log(Years)) + log_TrnSpd_AvgTrain, family = binomial(link = "logit"))
logit_model43 <- glm(DERAIL ~ VL_Count + log_VL_TrnSpd + Stn_Count + offset(log(Years)) + log_TrnSpd_AvgTrain, family = binomial(link = "logit"))
logit_model44 <- glm(DERAIL ~ VL_Count + log_VL_TrnSpd + Seg_Length + offset(log(Years)) + log_TrnSpd_AvgTrain, family = binomial(link = "logit"))
logit_model45 <- glm(DERAIL ~ VL_Count + log_VL_TrnSpd + log_Seg_Length + offset(log(Years)) + log_TrnSpd_AvgTrain, family = binomial(link = "logit"))
logit_model46 <- glm(DERAIL ~ VL_Count + VL_TrnSpd + log_Seg_Length + offset(log(Years)) + log_TrnSpd_AvgTrain, family = binomial(link = "logit"))
```

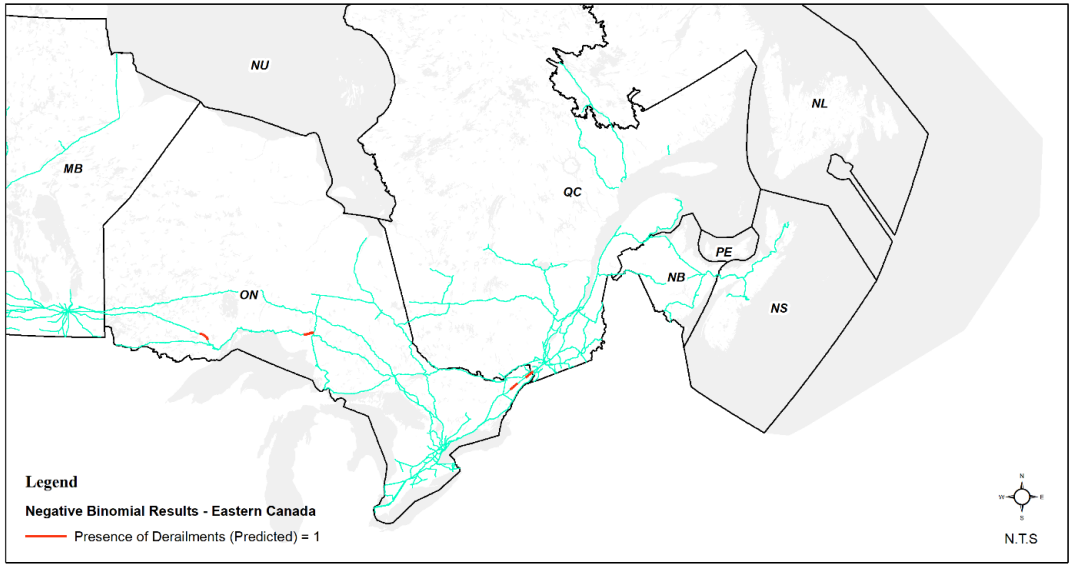
```
# Linear combination of variables using years and segment length as offsets, max Train Count as exposure and log_TrnSpd_VLCount as interaction variable
logit_model47 <- glm(DERAIL ~ VL_Count + offset(log(Years)) + offset(log_Seg_Length) + log_TrnSpd_VLCount, family = binomial(link = "logit"))
logit_model48 <- glm(DERAIL ~ VL_Count + VL_TrnSpd + offset(log(Years)) + offset(log_Seg_Length) + log_TrnSpd_VLCount, family = binomial(link = "logit"))
logit_model49 <- glm(DERAIL ~ VL_Count + VL_TrnSpd + Stn_Count + offset(log(Years)) + offset(log_Seg_Length) + log_TrnSpd_VLCount, family = binomial(link = "logit"))
logit_model50 <- glm(DERAIL ~ VL_Count + log_VL_TrnSpd + Stn_Count + offset(log(Years)) + offset(log_Seg_Length) + log_TrnSpd_VLCount, family = binomial(link = "logit"))
logit_model51 <- glm(DERAIL ~ VL_Count + log_VL_TrnSpd + Stn_Count + Seg_Length + offset(log(Years)) + offset(log_Seg_Length) + log_TrnSpd_VLCount, family = binomial(link = "logit"))

# Linear combination of variables using years and segment length as offsets, Average Train Count as exposure and log_TrnSpd_VLCount as interaction variable
logit_model52 <- glm(DERAIL ~ Avg_Train + offset(log(Years)) + offset(log_Seg_Length) + log_TrnSpd_VLCount, family = binomial(link = "logit"))
logit_model53 <- glm(DERAIL ~ Avg_Train + VL_TrnSpd + offset(log(Years)) + offset(log_Seg_Length) + log_TrnSpd_VLCount, family = binomial(link = "logit"))
logit_model54 <- glm(DERAIL ~ Avg_Train + VL_TrnSpd + Stn_Count + offset(log(Years)) + offset(log_Seg_Length) + log_TrnSpd_VLCount, family = binomial(link = "logit"))
logit_model55 <- glm(DERAIL ~ Avg_Train + log_VL_TrnSpd + Stn_Count + offset(log(Years)) + offset(log_Seg_Length) + log_TrnSpd_VLCount, family = binomial(link = "logit"))
logit_model56 <- glm(DERAIL ~ Avg_Train + log_VL_TrnSpd + Stn_Count + Seg_Length + offset(log(Years)) + offset(log_Seg_Length) + log_TrnSpd_VLCount, family = binomial(link = "logit"))

#Square of train speed or train count
logit_model57 <- glm(DERAIL ~ VL_Count + VL_TrnSpd^2 + Seg_Length + offset(log(Years)), family = binomial(link = "logit"))
logit_model58 <- glm(DERAIL ~ VL_Count + VL_Count^2 + VL_TrnSpd + Stn_Count + Seg_Length + offset(log(Years)), family = binomial(link = "logit"))
logit_model59 <- glm(DERAIL ~ Avg_Train + VL_TrnSpd^2 + Seg_Length + offset(log(Years)), family = binomial(link = "logit"))
logit_model60 <- glm(DERAIL ~ Avg_Train + Avg_Train^2 + VL_TrnSpd + Stn_Count + Seg_Length + offset(log(Years)), family = binomial(link = "logit"))

#Square of train speed or train count and log_TrnSpd_VLCount as interaction variable
logit_model57 <- glm(DERAIL ~ VL_Count + VL_TrnSpd^2 + Seg_Length + offset(log(Years)) + log_TrnSpd_VLCount, family = binomial(link = "logit"))
logit_model58 <- glm(DERAIL ~ VL_Count + VL_Count^2 + VL_TrnSpd + Stn_Count + Seg_Length + offset(log(Years)) + log_TrnSpd_VLCount, family = binomial(link = "logit"))
logit_model59 <- glm(DERAIL ~ Avg_Train + VL_TrnSpd^2 + Seg_Length + offset(log(Years)) + log_TrnSpd_VLCount, family = binomial(link = "logit"))
logit_model60 <- glm(DERAIL ~ Avg_Train + Avg_Train^2 + VL_TrnSpd + Stn_Count + Seg_Length + offset(log(Years)) + log_TrnSpd_VLCount, family = binomial(link = "logit"))
```

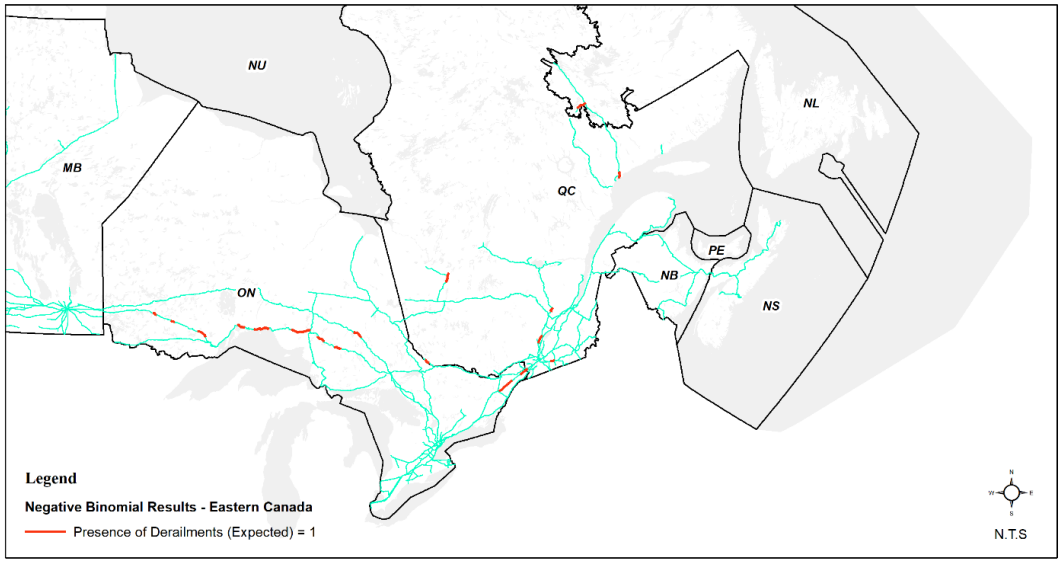
```
#Square of train speed or train count and log_TrnSpd_AvgTrain as interaction variable
logit_model61 <- glm(DERAIL ~ VL_Count + VL_TrnSpd^2 + Seg_Length + offset(log(Years)) + log_TrnSpd_AvgTrain, family = binomial(link = "logit"))
logit_model62 <- glm(DERAIL ~ VL_Count + VL_Count^2 + VL_TrnSpd + Stn_Count + Seg_Length + offset(log(Years)) + log_TrnSpd_AvgTrain, family = binomial(link = "logit"))
logit_model63 <- glm(DERAIL ~ Avg_Train + VL_TrnSpd^2 + Seg_Length + offset(log(Years)) + log_TrnSpd_AvgTrain, family = binomial(link = "logit"))
logit_model64 <- glm(DERAIL ~ VL_Count + VL_TrnSpd + log_Seg_Length + offset(log(Years)), family = binomial(link = "logit"))
#Refined Models
logit_model65 <- glm(DERAIL ~ VL_Count + VL_TrnSpd + Stn_Count + offset(log_Seg_Length) + offset(log(Years)) + VLCount_SegLen + log_TrnSpd_SegLen, family = binomial(link = "logit"))
logit_model66 <- glm(DERAIL ~ VL_Count + VL_TrnSpd + Stn_Count + offset(log_Seg_Length) + offset(log(Years)) + log_VLCount_SegLen + log_TrnSpd_SegLen, family = binomial(link = "logit"))
logit_model67 <- glm(DERAIL ~ VL_Count + VL_TrnSpd + Stn_Count + offset(log_Seg_Length) + offset(log(Years)) + log_VLCount_SegLen + TrnSpd_SegLen, family = binomial(link = "logit"))
logit_model68 <- glm(DERAIL ~ VL_Count + VL_TrnSpd + Stn_Count + offset(log_Seg_Length) + offset(log(Years)) + TrnSpd_VLCount + TrnSpd_SegLen, family = binomial(link = "logit"))
logit_model69 <- glm(DERAIL ~ VL_Count + VL_TrnSpd + Stn_Count + offset(log_Seg_Length) + offset(log(Years)) + StnCount_SegLen + TrnSpd_VLCount + TrnSpd_SegLen, family = binomial(link = "logit"))
logit_model70 <- glm(DERAIL ~ VL_Count + log_VL_TrnSpd + Stn_Count + offset(log_Seg_Length) + offset(log(Years)) + StnCount_SegLen + TrnSpd_VLCount + TrnSpd_SegLen, family = binomial(link = "logit"))
logit_model71 <- glm(DERAIL ~ log_VL_Count + log_VL_TrnSpd + Stn_Count + offset(log_Seg_Length) + offset(log(Years)) + StnCount_SegLen + TrnSpd_VLCount + TrnSpd_SegLen, family = binomial(link = "logit"))
logit_model72 <- glm(DERAIL ~ VL_Count + VL_TrnSpd + Stn_Count + offset(log_Seg_Length) + offset(log(Years)) + StnCount_SegLen + TrnSpd_VLCount + TrnSpd_SegLen, family = binomial(link = "logit"))
```

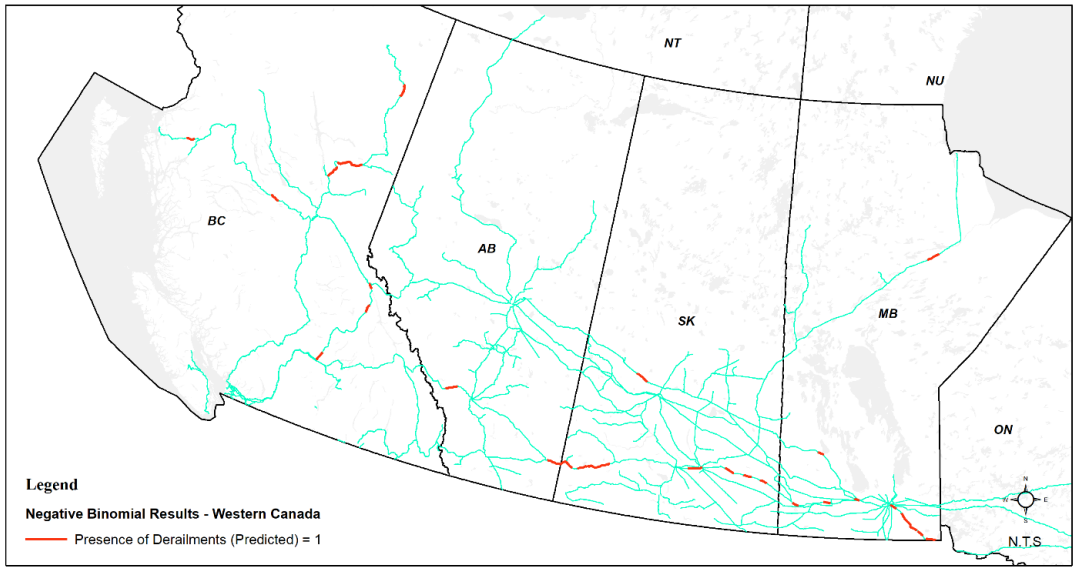
Appendix E: Converted Prediction Outcomes for Negative Binomial Models

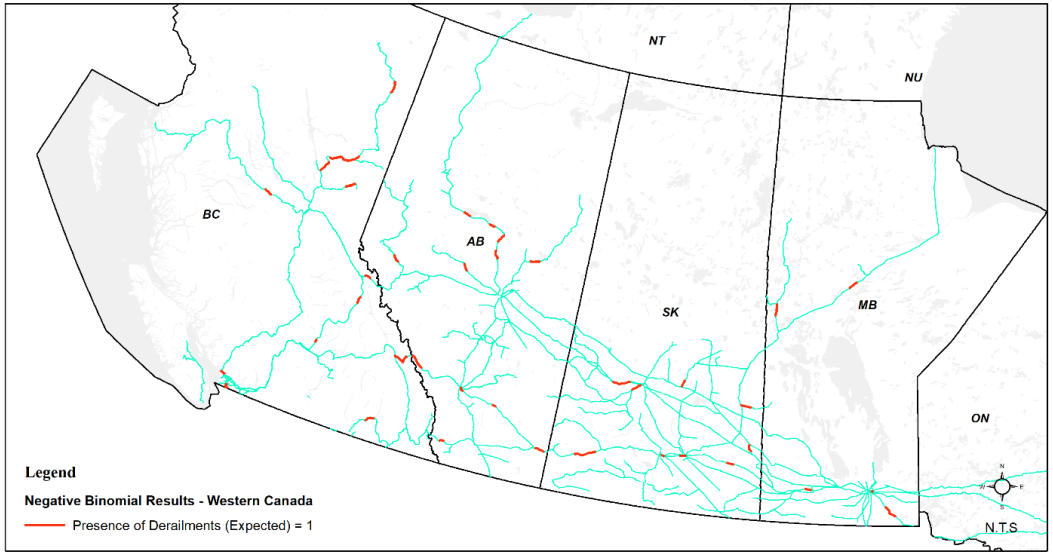


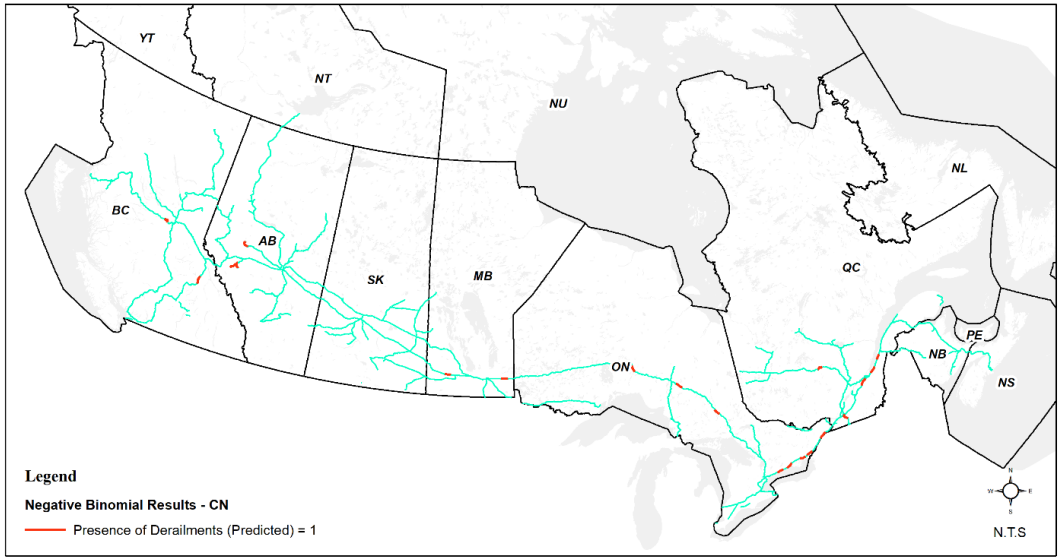


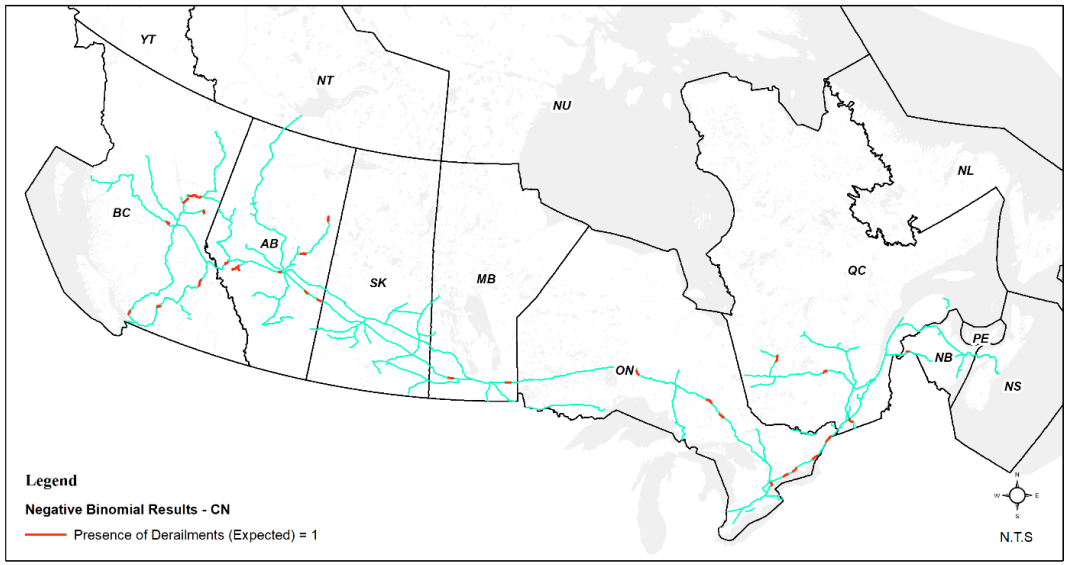

NU
MB
QC
NL
PE
NB
NS
ON
Legend
Negative Binomial Results - Eastern Canada
Presence of Derailments (Predicted) = 1
N.T.S

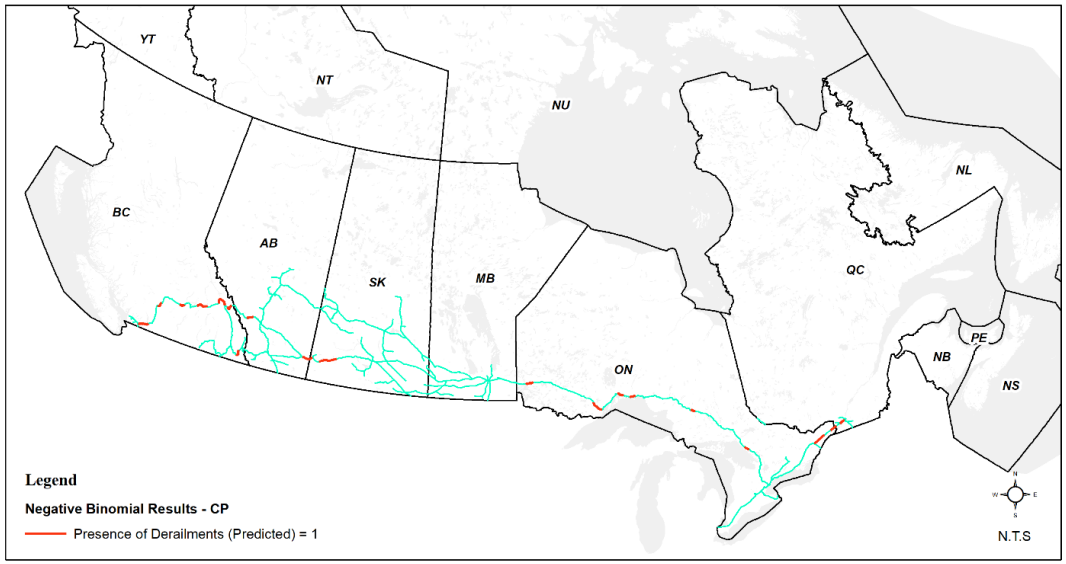

NU
MB
QC
NL
PE
NB
NS
ON
Legend
Negative Binomial Results - Eastern Canada
Presence of Derailments (Expected) = 1
N.T.S

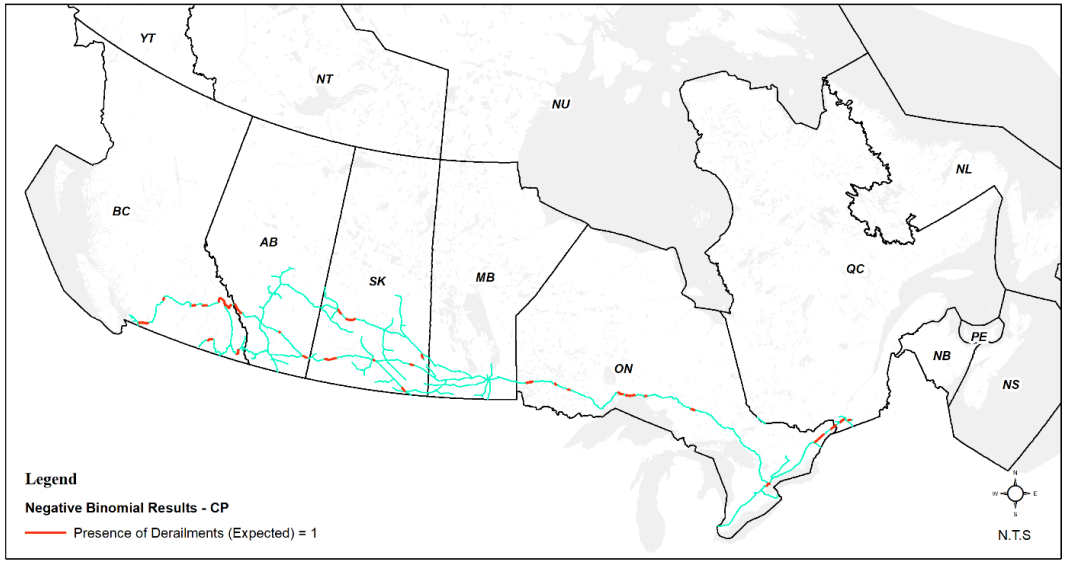

NT
NU
BC
AB
SK
MB
ON
Legend
Negative Binomial Results - Western Canada
Presence of Derailments (Predicted) = 1
N.T.S


NT
NU
BC
AB
SK
MB
ON
Legend
Negative Binomial Results - Western Canada
Presence of Derailments (Expected) = 1
N.T.S

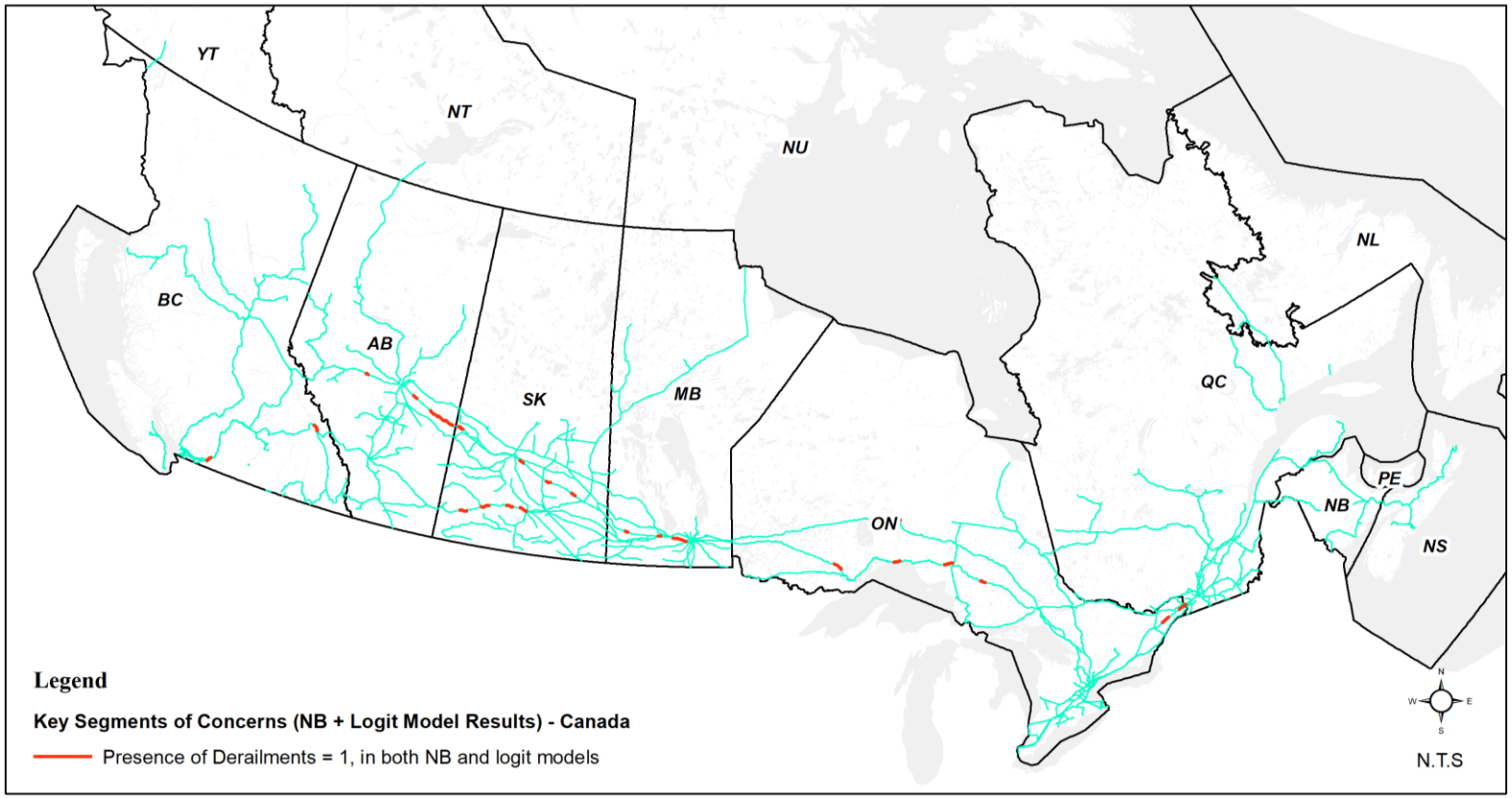

YT
NT
NU
NL
BC
AB
SK
MB
QC
PE
NB
NS
ON
Legend
Negative Binomial Results - CN
Presence of Derailments (Predicted) = 1
N.T.S

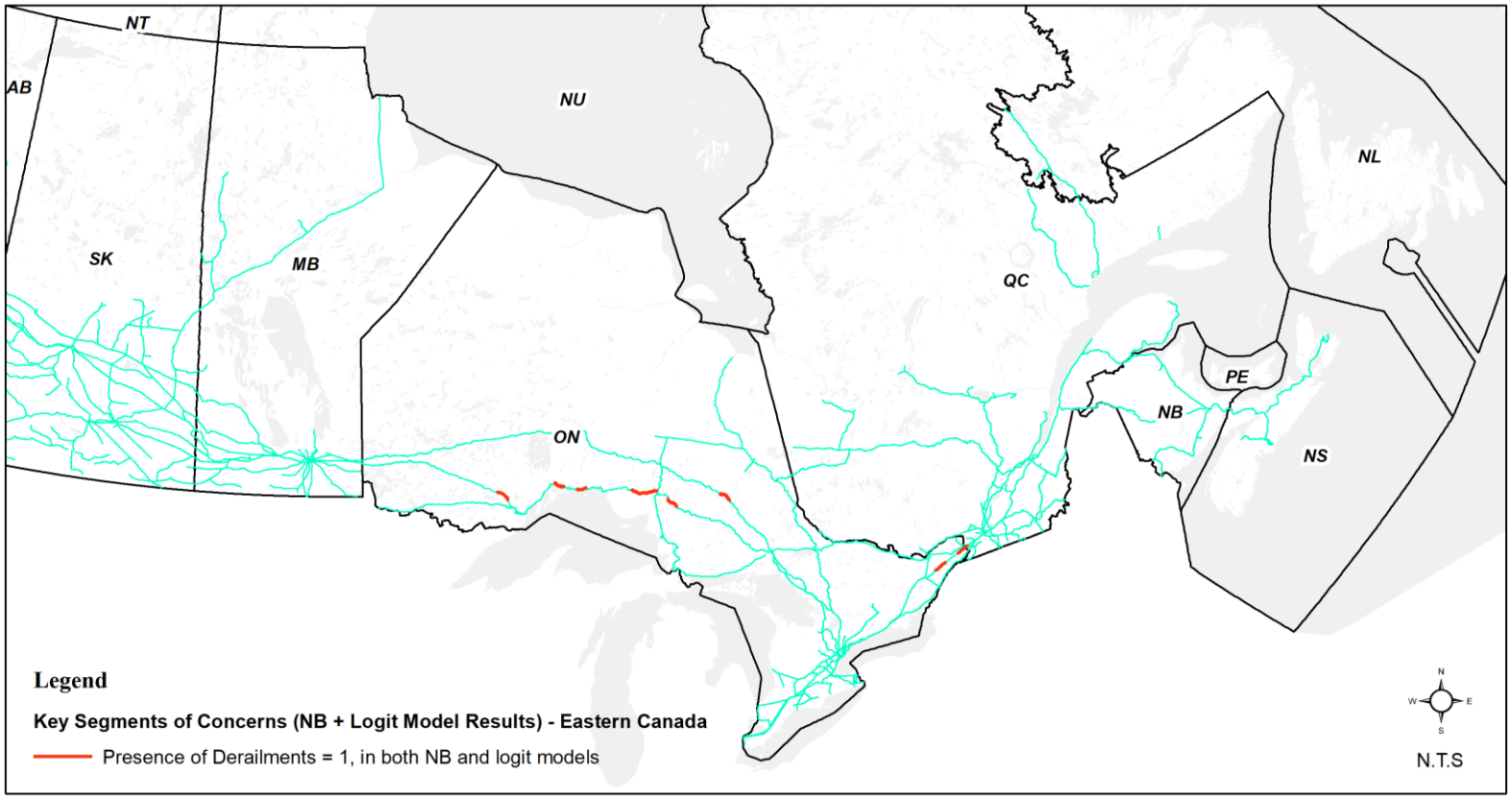

YT
NT
NU
NL
BC
AB
SK
MB
QC
PE
NB
NS
ON
Legend
Negative Binomial Results - CN
Presence of Derailments (Expected) = 1
N.T.S

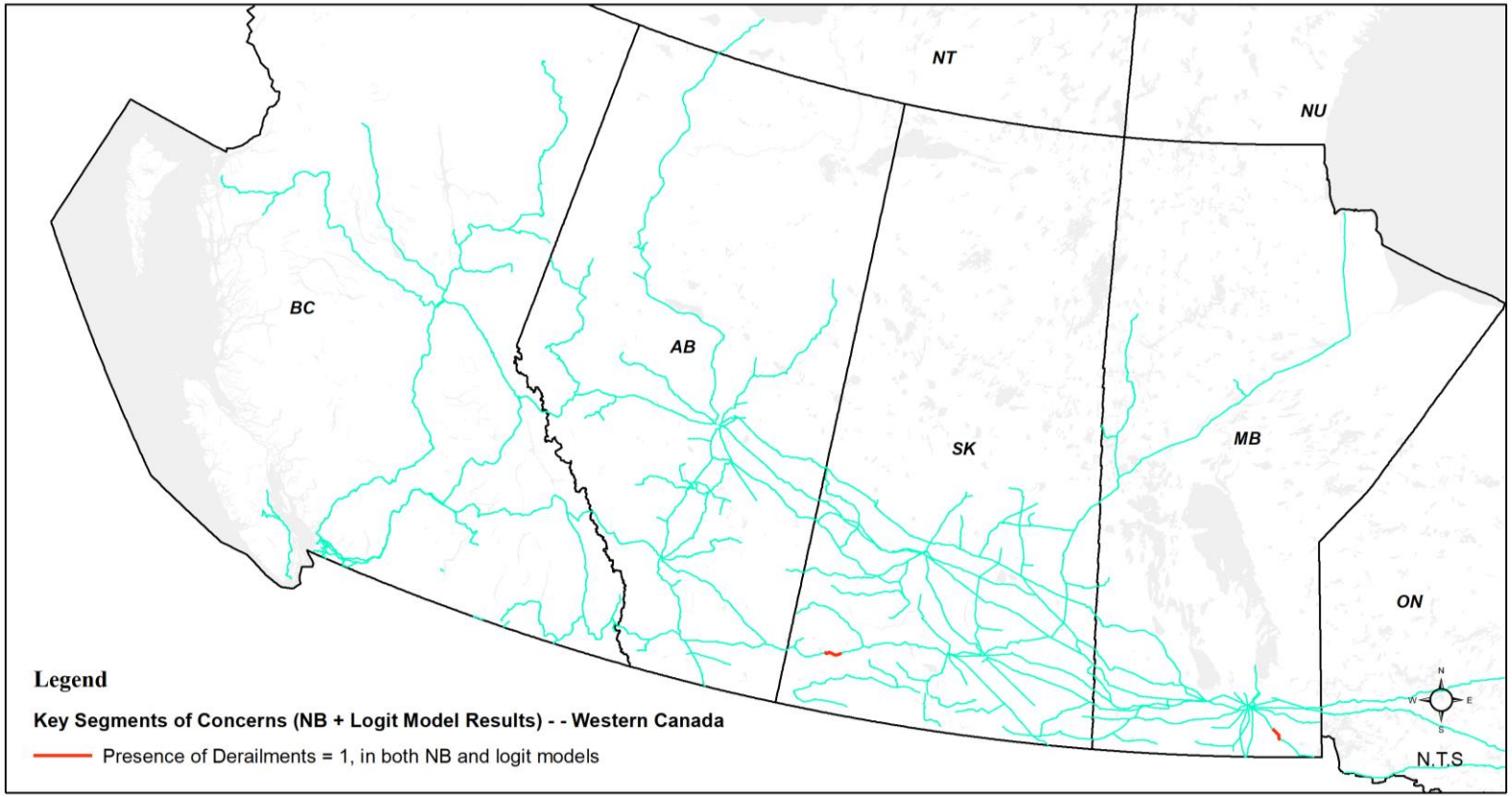

YT
NT
NU
BC
AB
SK
MB
ON
QC
NL
PE
NB
NS
Legend
Negative Binomial Results - CP
Presence of Derailments (Predicted) = 1
N.T.S

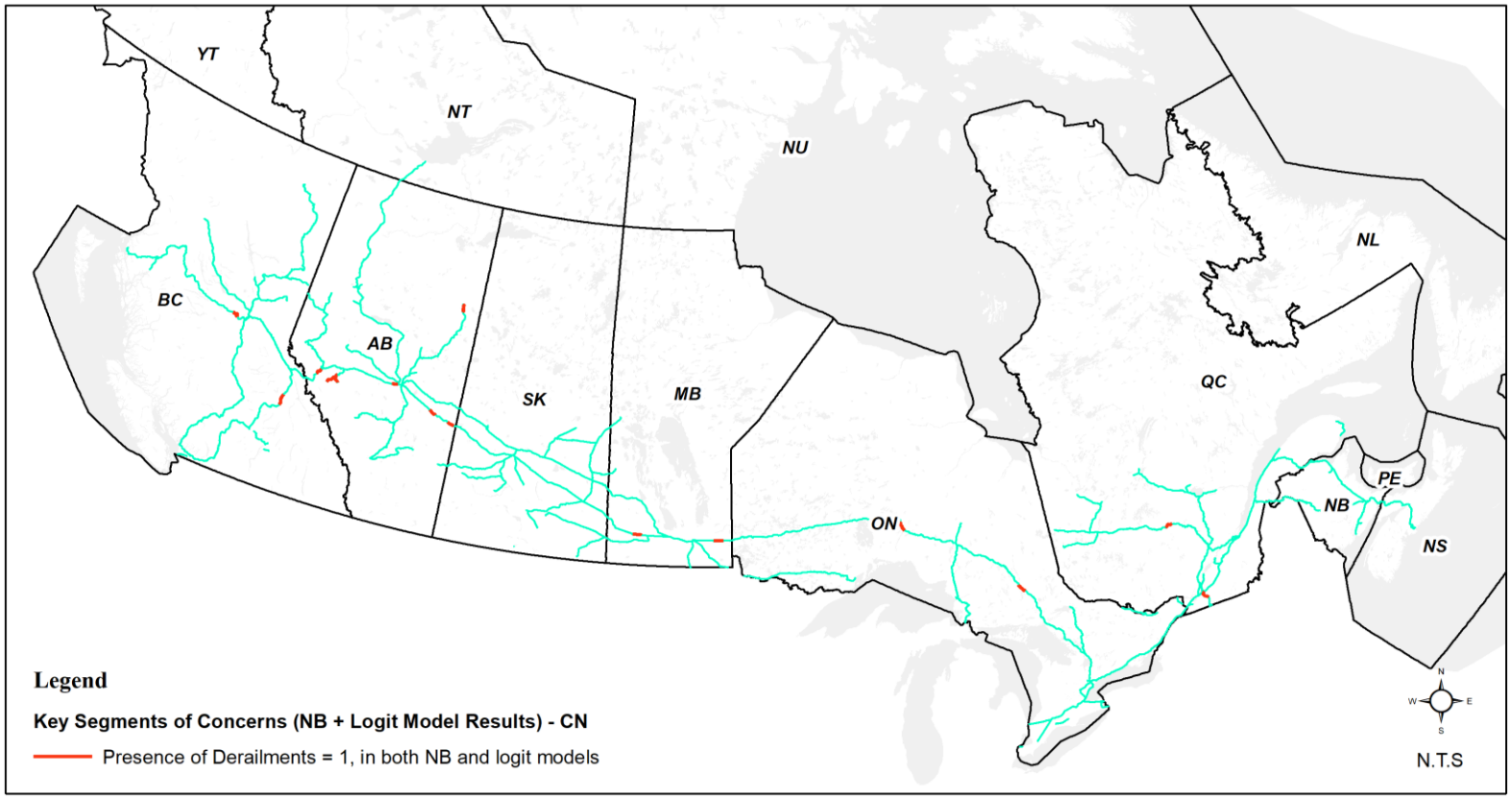

YT
NT
NU
BC
AB
SK
MB
ON
QC
NL
PE
NB
NS
Legend
Negative Binomial Results - CP
Presence of Derailments (Expected) = 1
N.T.S

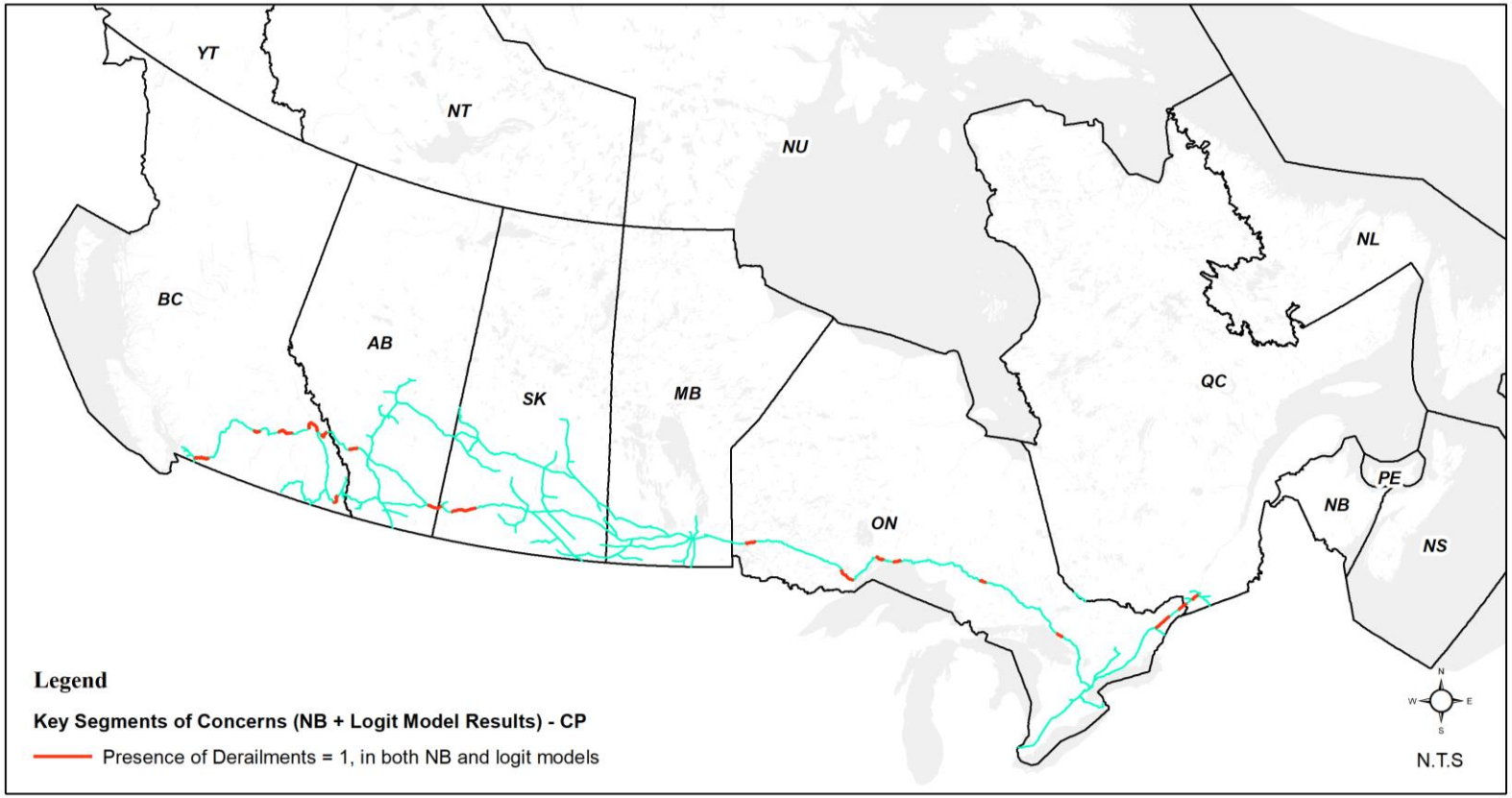
Appendix F: Key Segments of Concerns from Both Negative Binomial and Logit Models




NT
AB
NU
NL
SK
MB
QC
PE
NB
ON
NS
Legend
Key Segments of Concerns (NB + Logit Model Results) - Eastern Canada
Presence of Derailments = 1, in both NB and logit models
N
W
E
S
N.T.S


NT
NU
BC
AB
SK
MB
ON
Legend
Key Segments of Concerns (NB + Logit Model Results) - - Western Canada
Presence of Derailments = 1, in both NB and logit models
N
W
E
S
N.T.S


YT
NT
NU
BC
AB
SK
MB
ON
QC
NL
PE
NB
NS
Legend
Key Segments of Concerns (NB + Logit Model Results) - CN
Presence of Derailments = 1, in both NB and logit models
N
W
E
S
N.T.S


YT
NT
NU
BC
AB
SK
MB
ON
QC
NL
PE
NB
NS
Legend
Key Segments of Concerns (NB + Logit Model Results) - CP
Presence of Derailments = 1, in both NB and logit models
N
W
E
S
N.T.S