Development of Collision Prediction Models for Rural New Brunswick Highways

by

Jacqueline Kelly Connors

BScE, University of New Brunswick, 2018

A Thesis Submitted in Partial Fulfillment
of the Requirements for the Degree of

Master of Science in Engineering

in the Graduate Academic Unit of Civil Engineering

Supervisor: Eric D. Hildebrand, PhD, PEng, Department of Civil Engineering

Examining Board: James S. Christie, PhD, PEng, Department of Civil Engineering
Trevor Hanson, PhD, PEng, Department of Civil Engineering
Yuri Yevdokimov, PhD, Department of Economics

This thesis is accepted by the
Dean of Graduate Studies

THE UNIVERSITY OF NEW BRUNSWICK

February 2021

©Jacqueline Kelly Connors, 2021

Abstract

The Province of New Brunswick currently lacks the capability to identify road segments requiring safety remediation using state-of-the-art methodologies. This empirical research project enables collision locations to be objectively prioritized for cost-effective treatments. It is vital that only the most problematic locations be identified for improvement due to finite maintenance budgets.

Collision history data were analyzed, and roadway, roadside, and traffic data were compiled for a sample of rural locations with the goal of developing collision prediction models (CPMs). Two collision prediction models were developed for divided and undivided road segments using negative binomial regression. Variables found to have the most significant effect on collisions include AADT, segment length, grade, posted speed, lane width, plant hardiness zone, surface age, surface type, and wildlife fencing.

Predicted collision frequencies derived from the models can be compared to actual collision data to calculate potential for improvement (PFI) values to identify locations experiencing more collisions than expected. This will enable road authorities to make appropriate decisions to prioritize road locations that can benefit from remediation.

A case study was completed to illustrate how the models can be applied to road segments outside of the model development dataset. A demonstration was provided to show how road segments can be ranked using different collision performance techniques. The case study was also used to determine how the proposed models would perform in comparison with the default procedure in the Highway Safety Manual (HSM). It was determined that the newly developed CPMs resulted in smaller PFI values than the HSM SPFs.

Acknowledgements

I would like to acknowledge the following for their support and contributions to this research:

- Dr. Eric Hildebrand for his continuous encouragement and guidance throughout my studies and through the duration of this research project. Thank you for the insightful feedback and invaluable expertise that you provided to make this research possible.
- Dr. Jim Christie for his assistance with the statistical aspects of my research and generously contributing his time and extensive knowledge to help with the regression analyses.
- Diane Nash and Justine M^{ac}Leod with the NBDTI for their support and assistance locating data resources during my summer work with the Operations Branch.
- Roger Dick with the NBDTI Information Management & Technology Branch for providing me with access to the ArcGIS Data Warehouse and associated ArcGIS training resources necessary for data collection.
- New Brunswick Department of Transportation and Infrastructure for providing me with the required traffic volume, road geometry, and collision data for provincial roadways.
- Friends and family for supporting and encouraging me throughout my graduate studies.
- My husband, Andrew Connors, for his support and patience during my graduate studies. This work would not have been possible without your confidence in my abilities.
- All members of the UNB Transportation Group for their encouragement and valuable advice during this research project.

Table of Contents

ABSTRACT II
ACKNOWLEDGEMENTS III
LIST OF TABLES VI
LIST OF FIGURES VII
LIST OF ABBREVIATIONS AND ACRONYMS VIII

1. INTRODUCTION 1.1
1.1. Statement of the Motivating Problem 1.2
1.2. Research Purpose 1.3
1.3. Significance of the Study 1.3
1.4. Research Question and Hypothesis 1.5
1.5. Research Assumptions 1.6
1.6. Scope and Limitations 1.6
1.7. Expected Results 1.8

2. LITERATURE REVIEW 2.1
2.1. Collision Prediction Models 2.1
2.2. Data Collection and Segment Length 2.1
2.3. Regression Models 2.2
2.3.1. Poisson and Negative Binomial Models 2.2
2.3.2. Multivariate Models 2.4
2.3.3. Bayesian Methods 2.5
2.3.4. Hierarchical Models 2.7
2.3.5. Current Research 2.8
2.4. SAS Data Analysis 2.8
2.4.1. Correlation 2.9
2.4.2. Selection of Variables 2.9
2.4.3. SAS Output 2.10
2.4.4. Cross-validation 2.11
2.5. HSM Predictive Method 2.11
2.5.1. Negative Binomial Models 2.11
2.5.2. Crash Modifications Factors 2.12
2.5.3. Calibration 2.12
2.5.4. Jurisdiction-Specific SPFs 2.13
2.5.5. HSM2 ..... 2.16

3. METHODOLOGY 3.1
3.1. Data Collection 3.1
3.1.1. Representative Sampling 3.1
3.1.2. Data Resources 3.2
3.1.3. Potential Explanatory Variables 3.4
3.2. Collision Prediction Model Development 3.5
3.2.1. Correlation Matrix 3.6
3.2.2. Negative Binomial Regression Analysis 3.6
3.2.3. Model Validation 3.8

3.3. Case Study 3.10
3.3.1. Collision Frequency 3.11
3.3.2. Collision Rate 3.11
3.3.3. PFI Values from the Collision Prediction Models 3.11
3.3.4. PFI Values from the HSM SPFs 3.12

4. ANALYSIS AND RESULTS 4.1
4.1. Sample of Roadway Segments 4.1
4.1.1. Location of Segments 4.1
4.1.2. Representative Sampling 4.3
4.2. Collision Prediction Model Development 4.4
4.2.1. Variable Data 4.4
4.2.2. Correlation Matrix 4.4
4.2.3. Negative Binomial Regression 4.5
4.2.4. Linear Regression 4.10
4.3. Case Study 4.13
4.3.1. Collision Frequency 4.13
4.3.2. Collision Rate 4.15
4.3.3. PFI Values from the Collision Prediction Models 4.18
4.3.4. PFI Values from the HSM SPFs 4.22
4.3.5. Case Study Summary 4.26

5. CONCLUSIONS AND RECOMMENDATIONS 5.1
5.1. Conclusions 5.1
5.2. Recommendations 5.4

6. REFERENCES 6.1
A. APPENDIX A: SELECTED CONTROL SECTIONS A.1
B. APPENDIX B: VARIABLE DATA B.1
C. APPENDIX C: CORRELATIONAL STUDY C.1
D. APPENDIX D: SAS GENERALIZED LINEAR MODELS RESULTS D.1
E. APPENDIX E: SAS LINEAR REGRESSION RESULTS E.1
F. APPENDIX F: CASE STUDY RESULTS F.1
CURRICULUM VITAE

List of Tables

Table 3.1: Potential Explanatory Variables 3.5
Table 4.1: Generalized Linear Model Results for Divided Segments 4.7
Table 4.2: Generalized Linear Model Results for Undivided Segments 4.7
Table 4.3: Linear Regression Results for Divided Segments 4.11
Table 4.4: Linear Regression Results for Undivided Segments 4.11
Table 4.5: Divided Case Study Observations Ranked by Collision Frequency 4.14
Table 4.6: Undivided Case Study Observations Ranked by Collision Frequency 4.14
Table 4.7: Case Study Segments Ranked by Collision Frequency (Annual Average)...4.15
Table 4.8: Undivided Case Study Observations Ranked by Collision Rate....4.16
Table 4.9: Divided Case Study Observations Ranked by Collision Rate....4.17
Table 4.10: Case Study Segments Ranked by Collision Rate (Annual Average) 4.18
Table 4.11: PFI Values from the Divided Collision Prediction Model 4.19
Table 4.12: PFI Values from the Undivided Collision Prediction Model 4.20
Table 4.13: PFI Values from the Collision Prediction Models (Annual Average)....4.22
Table 4.14: PFI Values from the HSM rural 2-lane, 2-way undivided SPF 4.24
Table 4.15: PFI Values from the HSM rural 4-lane freeway SPF 4.25
Table 4.16: PFI Values from the HSM SPFs (Annual Average)....4.26
Table 4.17: Ranking of Undivided Case Study Segments from Different Approaches .4.27
Table A.1: List of Control Sections Selected for Model Development....A.1
Table A.2: List of Control Sections Selected for the Case Study.... A.2
Table B.1: Variable Data for Divided Segments B.1
Table B.2: Variable Data for Undivided Segments B.4
Table B.3: Variable Data for Case Study Segments....B.13
Table C.1: Correlation Matrix of Potential Explanatory Variables....C.1
Table D.1: Generalized Linear Model Results for Divided Segments D.1
Table D.2: Generalized Linear Model Results for Undivided Segments D.3
Table E.1: Linear Regression Results for Divided Segments....E.1
Table E.2: Linear Regression Results for Undivided Segments....E.2
Table F.1: Case Study Results from the Collision Prediction Models F.1
Table F.2: Statistical Significance of PFI Values from the CPMs and the HSM SPFs... F.3
Table F.3: Example of t-Test for PFI from the Collision Prediction Models F.4
Table F.4: Case Study Data for the HSM SPFs for 2-Lane Highways.... F.5
Table F.5: Case Study Data for HSM SPFs for Freeways (PDO Multiple Vehicle) F.6
Table F.6: Case Study Data for HSM SPFs for 4-Lane Freeways (FI Multiple Vehicle)F.6
Table F.7: Case Study Data for HSM SPFs for 4-Lane Freeways (FI Single Vehicle)... F.7
Table F.8: Case Study Data for HSM SPFs for Freeways (PDO Single Vehicle)..... F.7
Table F.9: Case Study Results for HSM SPFs for 4-Lane Freeways by Severity F.8

List of Figures

Figure 4.1: Location of selected segments on the GeoNB Map 4.2
Figure 4.2: Goodness-of-fit Plot for Divided Segments 4.12
Figure 4.3: Goodness-of-fit Plot for Undivided Segments 4.12
Figure A.1: County Map A.3
Figure A.2: District Map A.4
Figure E.1: Fit Plot for Actual vs Predicted Collisions (Divided Segments) E.1
Figure E.2: Fit Plot for Actual vs Predicted Collisions (Undivided Segments) E.2

List of Abbreviations and Acronyms

AADT	Annual Average Daily Traffic
AIC	Akaike information criterion
BIC	Bayesian information criterion
CMF	Crash Modification Factor
CPM	Collision Prediction Model
GLM	Generalized linear models
HSM	Highway Safety Manual
NB	Negative binomial
NBDTI	New Brunswick Department of Transportation and Infrastructure
PDO	Property Damage Only
PFI	Potential for Improvement
SPF	Safety Performance Function

1. Introduction

Road traffic safety practices are essential for the province of New Brunswick to reduce the number of motor vehicle collisions occurring within the road network. Given finite maintenance budgets, the government can only invest in treatments targeted to improve safety for a limited number of road locations. It is therefore imperative that only the most problematic locations be identified for improvement.

Determining where and why motor vehicle collisions occur within the road network is an important element to road safety engineering. The ability to estimate collision frequencies along highways allows for effective road safety improvements. The reduction of collisions within a highway segment requires focus on risk factors with explanatory value for determining collision frequencies.

The process of network screening is required to review locations within the road network to determine whether countermeasures have a high potential to reduce collision frequency in those locations. Two techniques within network screening which involve the use of actual historical data are the measurement of collision frequencies and the measurement of collision rates; however, the use of observed data alone creates biases.

The identification of locations with high collision frequencies based on observed collisions is biased toward high volume locations, i.e. locations with higher volumes will naturally have a higher collision frequency than locations with lower volumes. To account for the effect of volume on the results, collision rates can be developed. A common collision rate is calculated as frequency divided by volume and segment length. This technique is biased

toward low volume locations. It is based on the assumption that frequency and volume have a linear relationship; however, lower volume locations tend to have higher collision rates.

Due to the flaws of these two techniques, an expected number of collisions can be calculated for comparison purposes to determine a potential for improvement (PFI). This is currently the most widely accepted approach and requires the development of collision prediction models (CPM) to compare the results with the actual data (AASHTO, 2010). The PFI values are calculated as the difference between expected (modelled) number of collisions and actual number of collisions. Through the use of collision prediction modelling, locations experiencing more collisions than expected can be identified. The determination of such locations allows the focus to be properly placed on locations with safety issues that require remediation.

Various safety improvement projects can be initiated based upon the results of the models and the identification of target locations will allow for better allocation of funding. To estimate the benefits of the models, the percentage reduction in collisions can be calculated upon implementation of countermeasures. This will justify remediation of locations that are prioritized by the models.

1.1. Statement of the Motivating Problem

The problem addressed by this study is the premise that the current methods used for countermeasure implementation along New Brunswick's road network lead to inefficient allocation of resources. There is currently no objective approach employed to identify and

address locations experiencing high collision frequencies. The current approach results in the inefficient use of limited funds available for safety projects.

An extension of this problem is to create locally calibrated models to identify locations experiencing more collisions than expected to ensure the proper use of treatments at locations that are underperforming in terms of safety. This focus will aid in the allocation of investments rather than reactively implementing countermeasures based on public or political pressures.

1.2. Research Purpose

The goal of the research project is to develop empirically based models that objectively identify road network locations with unusually poor collision performance. The appropriate prediction of collisions and comparison with actual collisions will aid in identifying locations that are amenable to cost-effective countermeasures.

1.3. Significance of the Study

This research provides an objective approach to identify locations in New Brunswick's road network that are under-performing in terms of safety. Currently, the government of New Brunswick has an unstructured approach to address problem areas in the road network which leads to reactionary measures. This research project enables collision locations to be identified and properly prioritized for cost-effective treatments. This area of research is important for the Traffic Engineering and Operations Branch in the New Brunswick Department of Transportation and Infrastructure (NBDTI).

The public will also benefit from this research due to the primary goal of improving public safety. This research has the potential to reduce vehicle collisions in New Brunswick, and therefore decrease the casualties resulting from collisions. It was estimated by Statistics Canada that in New Brunswick in 2016, there were 5.8 fatalities per billion vehicle-kilometres and 314 injuries per billion vehicle-kilometres (Transport Canada, 2018).

The benefits and contributions resulting from this research can be measured by reduced cost through better allocation of resources by implementing a higher quality system to address specific road locations. Cost savings can be measured through collision costs as well as provincial budget allocation.

Due to the high costs involved with serious collisions, the cost savings from preventing and mitigating collisions can be very impactful. The cost of collisions has been estimated by several researchers; however, a relevant Canadian estimate was included in the economic analysis using data from New Brunswick in the dissertation titled “Road Safety Audits: Quantifying and Comparing the Benefits and Costs for Freeway Projects” by Peter Lougheed (2016). A collision resulting in a fatality was valued at $5,500,000, a collision resulting in an injury was valued at $150,000, and a collision resulting in property damage only was valued at $15,000.

This research will also reduce costs by allowing for the proper allocation of capital investment. This is due to the prioritization of locations experiencing more collisions than expected that will benefit the most from an investment in a targeted countermeasure. The countermeasures will be justified for each case by the objective approach. Based on a fixed

budget, this system will allow for more effective use of spending on treatments that provide positive results.

According to the Ministerial Statement from May 2019, the ordinary budget for 2019-2020 for the Department of Transportation and Infrastructure was $312 million; however, only a portion of that budget can be allocated to road safety projects (GNB, 2020). For example, in 2018-2019, $122 million was budgeted towards road rehabilitation of asphalt and chip seal surfaces, which was less than the annual amount recommended by the Asset Management model to bring the roads to good condition in the next 10 years (GNB, 2019). The current methods for road rehabilitation are aiming to address comfort and structural deficiencies, but they are not necessarily targeting road safety issues.

The procedure for determining which areas will benefit the most from countermeasures provides a higher quality screening system. This research is worth pursuing because a more proactive approach allows for improved safety performance and proper investment.

1.4. Research Question and Hypothesis

The research question addressed by this study was to determine how the proposed models would perform in comparison with the procedure in the Highway Safety Manual (HSM) in terms of properly identifying locations with high potential for improvement (PFI) values. It was hypothesized that the collision predictions from the proposed models would better represent local conditions due to the jurisdiction-specific approach used during model development as well as transferability issues with the HSM procedure. The models were developed from a sample of New Brunswick road segments. It was expected that the new models would use descriptive variables to properly identify locations with more collisions

than expected. Regression analyses were completed to determine the goodness-of-fit for each model.

1.5. Research Assumptions

A few assumptions were required to be made for the research to be completed. It was assumed that road and traffic characteristics were consistent along each road segment. The control sections were assumed to be homogeneous roadway segments with similar volumes, geometric design, and traffic control features. Another assumption was that no significant changes in road geometry or road condition had occurred throughout the duration of the study sufficient to affect collision outcome.

1.6. Scope and Limitations

Time and budget constraints limited the depth of this research study. The data for inclusion in the models were restricted to NBDTI's database for traffic volumes, collision reports, and road life study diagrams. Further data collection in the field was not practical for the study time period. The physical boundaries of the project were restricted to rural areas outside of urban growth boundaries. Any roads classified as local (route numbers greater than 199) were excluded from the project. Roads classified as collector (route numbers 100-199) or arterial (route numbers 1-99) were included. The scope was also restricted to road segments only, excluding at-grade intersections and interchange systems.

The data analysis included five years of collision data. The collision database was also restrictive in that only collisions reported by the police are included. The reporting

threshold includes any collisions resulting in injury, fatality or property damage exceeding $1,000.

Specific variables were considered for inclusion in the models. Data availability from the government of New Brunswick constrained the choice of variables. The research was also confined to provincially owned roads.

The models were developed from a sample of the road network. The evaluation of the models was completed through a case study on a separate sample of highway segments to provide an example of application and justification of the recommended method. The selected sections for the case study were not included in the data used for model development.

For comparison purposes, the case study involved the contrasting of segments identified for remediation using the following techniques:

- collision frequencies from actual collision data,
- collision rates from actual collision data,
- potential for improvement (PFI) values from the newly developed predictive models, and
- PFI values from the Highway Safety Manual (HSM) procedure.

Recommendations are presented for implementation of the models to other locations, but the current study does not include an analysis of all roads in New Brunswick.

1.7. Expected Results

The models consist of safety performance functions for different road types to be used by the Operations Branch at NBDTI. It is expected that the models will allow for a more efficient way of utilizing funds by identifying locations that are treatable with a countermeasure. The appropriate prediction of collisions and comparison with actual data will aid in identifying locations that are more likely to benefit from investments in countermeasures.

The models and associated recommendations for implementation will allow for a better method of identifying road safety concerns that cause collisions along New Brunswick's road network. This research provides an objective approach to identify and address locations experiencing higher collision frequencies than expected and it will aid in the efficient allocation of resources according to the annual budget.

2. Literature Review

In order to identify locations in a road network that are under-performing in terms of safety, there is a wide range and diversity of techniques applied to detect locations experiencing more collisions. A literature review was conducted to evaluate relevant research to find examples of experimental methodologies and models in similar studies. Current methodologies used to model collisions were reviewed and existing resources that are available were referenced. The literature review was conducted to describe other researchers' perspectives and previous findings related to this research.

2.1. Collision Prediction Models

Collision prediction models (CPMs) refer to mathematical equations that relate the number of collisions to specific roadway, roadside, and traffic characteristics (Srinivasan and Bauer, 2013). The number of collisions is the dependent variable in the equation. The safety performance functions (SPFs) in the Highway Safety Manual (HSM) are CPMs that can be used to determine the expected average crash frequency for existing conditions or the safety impacts of changes in design from roadway projects (Srinivasan and Bauer, 2013). The expected average crash frequency is based on the predicted average crash frequency from a predictive model and the observed annual crash frequency.

2.2. Data Collection and Segment Length

In the application of the HSM SPFs and in the development of jurisdiction-specific SPFs, it is recommended that roadway segments be divided into homogeneous sections with the value for each variable remaining unchanged throughout the length of the segment

(AASHTO, 2010). With variables that tend to change frequently along the roadway, however, this results in very short segments and issues with accounting for collisions in the proper segment (Srinivasan and Bauer, 2013). A study by Bonneson *et al.* (2012) used databases that contained non-homogeneous segments in terms of the variables being considered for inclusion in their models for freeways and interchanges. They found that many of the geometric road characteristics would change throughout the length of a segment. Instead of separating the segments into very short segments, they decided to include the proportion of segment lengths associated with the different roadway characteristics (such as horizontal curve presence) in their data collection (Bonneson *et al*., 2012). By using longer non-homogeneous segments, it is more likely that collisions can be located in the appropriate segment rather than span the boundary between two shorter segments.

2.3. Regression Models

The dependent variable in collision prediction models is the number of collisions. Collisions are classified as count data and can therefore be modeled using count data models (Srinivasan and Bauer, 2013).

2.3.1. Poisson and Negative Binomial Models

Regression models that are most commonly used for collision count data are Poisson and negative binomial models, which are generalized linear models (GLM) with log-linear model relationships between the explanatory variables and the expected number of collisions. The log-linear modelling form results in a linear combination of the independent

variables with their associated coefficients estimated by the maximum likelihood method (Srinivasan and Bauer, 2013).

Poisson regression is the foundation of many methodologies in this field of research due to the count nature of collision data (the dependent variable) (Lord and Mannering, 2010). A number of evaluations were performed by Lord and Mannering (2010) to compare different methodologies for analysing crash frequency data and they concluded that Poisson regression is the most basic model, but it cannot handle over- and under-dispersion. The Poisson distribution restricts the variance to be equal to the mean; however, it is common for collision data to be over-dispersed, meaning the variance of the data exceeds the mean.

Negative binomial (Poisson-gamma) regression is a form of the Poisson model and is more appropriate than Poisson regression if over-dispersion is present due to its extra parameter that adjusts the variance separately from the mean. Over-dispersion occurs when there is greater variability in the data than expected. The error term in the negative binomial model is gamma-distributed and allows the variance to differ from the mean. The negative binomial model is considered the current standard method for the development of safety performance functions (SPFs).

Young and Park (2013) used a Poisson-gamma distribution to develop SPFs for Regina, Saskatchewan. They included variables for exposure (traffic volume) and road segment length in their SPFs (Young and Park, 2013). Kwon *et al*. (2013) also used a negative binomial regression model for parameter estimation for the safety performance functions in their study on segmenting methods. The new SPFs developed by Brimley *et al.* (2012) in Utah also used negative binomial regression.

Another alternative regression model is the Poisson-lognormal model, which differs from the negative binomial model by using a lognormal distributed Poisson parameter rather than gamma-distribution. This results in a more flexible model to handle over-dispersion (Lord and Mannering, 2010). Sacchi *et al*. (2015) applied Poisson–lognormal regression models to their traffic and collision data by means of the full Bayesian method. Multivariate and univariate approaches were both used for comparison purposes and two Poisson–lognormal models were developed with time-varying coefficients. The Poisson-lognormal distribution does not, however, have a closed form, resulting in a more complex model estimation than the simple Poisson or negative binomial distributions (Lord and Mannering, 2010).

Dong *et al*. (2016) argue that models that account for space-time effects outperform the Poisson-lognormal model in terms of safety ranking. The researchers applied the Poisson-lognormal model in Florida to fit the crash rate for comparison purposes and the research concluded that the spatio-temporal interaction effect had better goodness-of-fit than the Poisson-lognormal model. A couple of cons to these types of models include that they may not be easy to transfer to other datasets and that the results of the correlation can be difficult to interpret (Lord and Mannering, 2010).

2.3.2. Multivariate Models

Multivariate models are statistical procedures that are not limited to just number of crashes. Several variables that affect crash frequency are accounted for "based on the premise that differences in the characteristics of features result in different crash outcomes" (AASHTO, 2010). These models can simultaneously model different types of crashes; however, they

involve a complex estimation process that requires correlation matrices (Lord and Mannering, 2010).

Multivariate models were addressed by Sacchi *et al*. (2015) as a means of modelling multiple response variables jointly to prioritize sites for safety improvements. The results from the study justify the use of the multivariate structure, supported by a consistency test. It was concluded that the multivariate method can extend the univariate concept of the potential for safety improvement; however, the researchers acknowledge that their results are based on a single dataset and further research would be required to confirm their conclusions. It is also addressed that it is difficult to determine the accuracy of the identified collision hot spots due to the fact that in analyzing the real data, it is not known whether the sites are actually hazardous (Sacchi *et al*., 2015).

Although they are complex to estimate, Dong *et al*. (2016) recommended employing multivariate models in future research in order to confirm their conclusions. They evaluated and compared three methods for hotspot identification using crash data from Florida. In their research, the results were only based on crash frequencies and could benefit from including severity levels.

2.3.3. Bayesian Methods

Bayesian methods are used to fit probability models to datasets to determine the probability distribution of the parameters included in the models. The probability distribution can be used for the prediction of unobserved quantities (Gelman *et al*., 1995). The steps involved in Bayesian data analysis are to set up the probability model, calculate and interpret the posterior distribution (i.e. the conditional probability distribution of the unobserved

quantities based on the observed data), and evaluate the fit of the model. Iterations can then be performed if necessary (Gelman *et al*., 1995). The model comparison paper previously discussed by Lord and Mannering (2010) concluded that although Bayesian methods involve a complex estimation process, they usually have a better fit than traditional parametric models.

The Empirical Bayes (EB) technique allows for a model based on historical data to predict the expected number of collisions. The EB method is addressed in Appendix A of the HSM as a method of combining the estimate calculated using the HSM SPFs and observed crash frequencies to "obtain a more reliable estimate of expected average crash frequency" (AASHTO, 2010). This technique can be used to compensate for any regression-to-the mean bias.

The method proposed by Sacchi *et al.* (2015) was developed in a full Bayesian context and when evaluated against the more common procedures, it provided more realistic results. Sacchi *et al.* used safety performance functions from the same types of road facilities to aid in the estimation of expected collision frequencies. The macroscopic hotspots identification research completed by Dong *et al*. (2016) also used full Bayesian methods.

The new cluster-based algorithm proposed by Bi *et al.* (2014) was compared with the EB algorithm. The new method considered more than just collision counts in order to quantify roadway safety performance. Examples of the safety-related factors were fatality, injuries, and collision duration. It was determined that for a small sample of segments, the EB algorithm is better and for large numbers of samples, the new algorithm is justified.

2.3.4. Hierarchical Models

Hierarchical models include many parameters that are dependent on each other, resulting in joint probability models that reflect the correlations among the parameters (Lord and Mannering, 2010; Gelman *et al*., 1995). When the data are related by hierarchical relationships, the use of non-hierarchical models is insufficient due to their inability to accurately fit the data when few parameters are included and the risk of over-fitting the data when many parameters are included. Over-fitting the data refers to the case when the models fit the actual data well but cannot accurately predict new data. The use of hierarchical models allows for the inclusion of sufficient parameters for a proper fit to the data and accounts for dependence between parameters through the use of population distribution (Gelman *et al*., 1995).

The full Bayesian approach explored by Sacchi *et al*. (2015) highlighted its capability of multi-level interpretation for hierarchical models. Hierarchical modelling was also emphasized by Bi *et al*. (2014), who concluded that it outperforms Empirical Bayes for identification of locations with high collision frequencies. Hierarchical models are listed as one of the commonly used clustering methods for different severity level representation.

Dong *et al*. (2016) also referenced various studies that have identified hierarchical modeling techniques as better fits for crash rates. One downfall stated by Dong *et al*. regarding hierarchical modeling approaches is that they treat spatial and temporal effects separately, which prevents the acknowledgement of the interaction of space and time to collision occurrence. Lord and Mannering (2010) also stated that hierarchical models do

not always provide better results than traditional techniques and that it may be difficult to transfer to other datasets and to interpret correlation results.

2.3.5. Current Research

In relation to the current study, the above-noted model comparisons aid in identifying key factors when choosing appropriate models for prioritizing road segment locations that warrant remediations. This study is concerned with the development of collision prediction models for rural highways in New Brunswick and the comparison of the results with the HSM SPF results.

Due to the nature of the available data in New Brunswick and the available resources for statistical analysis, Poisson and negative binomial regression were considered for the model development procedure. Findings from these previous studies suggested that the complexity of multivariate, Bayesian, and hierarchical models would be too resource intensive for the current study.

2.4. SAS Data Analysis

SAS software can be used to carry out the statistical analysis required for model development. Specifically, it is useful in determining whether certain independent variables are good candidates for predicting the number of collisions. Another use of the software is to determine if over-dispersion is present, indicating whether a negative binomial model would be appropriate (UCLA: Statistical Consulting Group). SAS software is capable of fitting generalized linear models using the GENMOD procedure and the option to select the distribution allows users to specify a negative binomial distribution.

2.4.1. Correlation

One concern during model development is the potential for correlation between variables (Srinivasan and Bauer, 2013). For example, a road's speed limit is likely to be strongly correlated with some geometric features such as shoulder width or presence of a directional divider. The correlation between pairs of variables can be assessed using SAS to develop a correlation matrix. Only one of two highly correlated variables should be included in the model.

2.4.2. Selection of Variables

The selection of independent variables is the main component of developing CPMs. To identify the significant variables, a stepwise regression approach can be used, using a forward selection procedure or a backward elimination procedure (Srinivasan and Bauer, 2013).

The forward selection procedure begins with no variables in the model and each new variable is added individually to determine the effects of that variable. For each step, the variable that improves the model the most is selected to be added until the outstanding variables no longer improve the model with statistical significance. The backward elimination procedure begins with a model that includes all potentially explanatory variables and each variable is removed individually to determine the effects of deleting that variable. For each step, the variable that is determined to be the least significant to the model is selected to be eliminated until the remaining variables are all statistically significant (Srinivasan and Bauer, 2013).

The addition/elimination of each variable is tested using model comparison criteria such as t-statistic, chi-square statistic, Akaike's information criterion (AIC), and Bayesian Information Criterion (BIC) (Srinivasan and Bauer, 2013).

2.4.3. SAS Output

The output results from the generalized linear models task contain the Criteria for Assessing Goodness of Fit table. Pearson chi-square is included in this table and is used to test the model form to determine whether the data are over-dispersed when modeled with a negative binomial distribution. The negative binomial model is confirmed to be a good fit for the data if the p-value is not significant (UCLA: Statistical Consulting Group).

In order to assess the relative quality of successive models when employing a stepwise regression approach, the AIC and the BIC can be referenced. AIC and BIC take into account the number of variables included and provide a method of selecting the best model by dealing with the compromise between the model's goodness-of-fit and its complexity (Srinivasan and Bauer, 2013).

Another resulting output table is the Analysis of Maximum Likelihood Parameter Estimates table, which presents the coefficient estimates along with confidence intervals, chi-square tests and p-values for each of the independent variables. The dispersion coefficient estimate is also presented and if the confidence interval for the dispersion coefficient does not include zero, then the negative binomial model form is determined to be more appropriate than the Poisson model form. The LR Statistics for Type 3 Analysis table aids in determining if categorical variables are statistically significant.

2.4.4. Cross-validation

The data can be split randomly into two datasets as a method of ensuring that overfitting is not an issue and to demonstrate the use of the model. The first set of data is used for the model development procedure and the second set is used for model validation (Srinivasan and Bauer, 2013).

2.5. HSM Predictive Method

It is noted that one common practice is to follow the calibration procedure in the Highway Safety Manual (HSM) (AASHTO, 2010). The predictive models in the HSM include safety performance functions (SPFs) that estimate the average collision frequency of a network, facility, or individual site. The models were developed based on data from a few jurisdictions and the authors of the HSM acknowledge the need for local calibration of the safety performance functions to be applicable to different jurisdictions. The SPFs provided in the HSM are accompanied by crash modification factors (CMF) which are multiplied by the original functions to account for site-specific conditions. It is further recommended that jurisdiction-specific SPFs be developed if possible (AASHTO, 2010).

2.5.1. Negative Binomial Models

The SPFs in the HSM were developed using negative binomial regression with a log-linear relationship between collision frequency and road site characteristics (AASHTO, 2010). The facility types for which SPFs are provided in the HSM that are relevant to the present study are rural two-lane, two-way roads (undivided) and rural four-lane freeways. For roadway segments, the estimates are measured in number of crashes per mile per year.

A SPF Decision Guide was developed by Srinivasan *et al.* (2013) at the University of North Carolina Highway Safety Research Center to aid agencies with the decision of whether to calibrate the existing SPFs from the HSM or develop jurisdiction-specific SPFs. The HSM SPFs are negative binomial (NB) regression models that were estimated using a generalized linear models (GLM) procedure (Srinivasan *et al.*, 2013).

2.5.2. Crash Modifications Factors

The HSM prediction methods involve the use of SPFs to predict the crash frequency for a "base" condition, with AADT as the only independent variable. Then, crash modification factors (CMFs) can be applied to adjust the predictions for any conditions that differ from the base condition (AASHTO, 2010).

SPF results are multiplied by all of the applicable CMFs to adjust the predicted collision frequencies to account for site-specific conditions (AASHTO, 2010). Sites are defined as project locations, such as intersections or roadway segments. Each CMF represents the effects of any features differing from the base condition and all of these effects are assumed to be independent; however, it is noted that little research has been conducted about the independence of these effects (AASHTO, 2010).

2.5.3. Calibration

In the HSM, there is also a calibration process to account for local geographic conditions due to differences in "climate, animal population, driver populations, crash reporting theshold, and crash reporting practices" (AASHTO, 2010). Calibration factors can be used to adjust the base models to a jurisdiction outside of the jurisdictions used to develop the

models. The calibration factor for roadway segments (C_r) is based on the ratio of the observed collision frequencies for a chosen set of sections to the expected average collision frequency estimated for the same sections in the same time frame.

The National Cooperative Highway Research Program (NCHRP) prepared a guide in 2014 to aid HSM users with the process of developing calibration factors, titled "User's Guide to Develop Highway Safety Manual Safety Performance Function Calibration Factors". Calibration factor development integrates existing data on collisions, traffic volumes, and roadway elements in order to provide better collision estimates for jurisdiction-specific conditions (Bahar, 2014).

2.5.4. Jurisdiction-Specific SPFs

It is recommended by AASHTO to develop new SPFs unique to each jurisdiction to more accurately predict crashes if time permits and sufficient data are available. A number of studies have been completed in various jurisdictions to compare the HSM's procedure with newly developed SPFs. Two examples of such studies were performed in Tennessee and South Dakota, which both concluded that jurisdiction-specific models outperformed the calibrated functions from the HSM in terms of accuracy (Qin *et al*., 2018; Wali *et al*., 2018). Using data that reflect specific characteristics from a particular region allows for models that more accurately reflect the road and traffic conditions of that region.

Even with the calibration procedure, often the HSM SPFs lack accuracy and in some cases, it is determined that the calibrated models do not appropriately predict collision counts. The HSM authors acknowledge that SPFs developed directly with data from a specific jurisdiction may provide more reliable and accurate estimates (AASHTO, 2010).

It is recommended to use negative binomial regression models using generalized linear modeling techniques to develop jurisdiction-specific SPFs. To develop the models, a sample of 100-200 miles is needed with at least 300 crashes per year and at least 3 years of data (Srinivasan *et al*., 2013).

Different studies from states (and internationally) have compared the procedure in the HSM to models they have developed to test the reliability of the calibration factors. In many studies that were reviewed, the accuracy of models from the HSM in comparison with jurisdiction-specific SPFs often resulted in criticisms of the HSM procedure. Developing jurisdiction-specific SPFs allows for additional variables to be tested for explanatory value.

A number of studies have discussed the advantages of jurisdiction-specific SPFs over the models from the HSM. A study conducted by Lubliner (2011) tested the effectiveness of the HSM predictive models for highways in Kansas. Alternative calibration procedures were derived by the researcher and their outcomes were validated in comparison with the HSM calibration procedure. Lubliner's prediction models using local data and site-specific characteristics proved to have better accuracy. In applying the HSM and incorporating specific qualities of the Kansas highway system, Lubliner was able to conclude that the newly developed calibration procedures outperformed the HSM calibration procedure. The county-specific calibration developed by Lubliner resulted in improved model accuracy in comparison with the single statewide calibration procedure developed based on the HSM methodology. The results were measured in terms of percentage improvement of the differences between the predicted and actual crashes in comparison to the statewide calibration results. The county-specific calibration resulted in an average percent-

difference that was 5.7 percent more accurate than the statewide calibration (Lubliner, 2011).

The HSM calibration procedure is used based on the underlying assumption that the CMFs in Part C of the HSM are universal, which has been challenged by researchers in Ontario and Italy (Sacchi *et al.*, 2012). This assumption implies that the CMFs for adjusting the base condition models are applicable to all states and are not a function of site characteristics (Srinivasan *et al.*, 2013). A paper prepared by Sacchi *et al.* focused on a methodology to assess the international transferability of SPFs and CMFs from the HSM. The study presented tools that could be used by other countries to determine whether the HSM models are compatible with their locations (Sacchi *et al.*, 2012). In their study, the calibration factors estimated from the HSM procedure resulted in a wide range of values even within jurisdictions. The study concluded that for segments with few crashes, the HSM models tended to overestimate crash frequency and for segments with higher number of crashes, the models underestimated crash frequency (Sacchi *et al.*, 2012). The authors recommended evaluating the validity of the SPFs when the HSM models are attempted to be applied in another jurisdiction, especially in other countries.

A study by Brimley *et al.* (2012) performed in Utah also followed the calibration procedure in the HSM and compared the results with new SPFs that were developed using negative binomial regression. The resulting calibration factor from applying the HSM calibration procedure was determined to be 1.16, which indicated that the HSM SPF underpredicted the number of collisions in Utah. The newly developed models included some additional variables that are not used explicitly in the HSM. The jurisdiction-specific SPFs developed by Brimley *et al.* resulted in more accurate crash predictions. It is also to be noted that the

recommended model for implementation in Utah that was developed by the researchers included the traffic flow variable as the natural log of AADT, as opposed to the linear relationship of AADT to number of collisions in the HSM. This preferred model was chosen through the analysis of minimizing the BIC of each of the proposed models. (Brimley *et al*., 2012).

Similar results were found in the study by Young and Park (2013), which compared the performance of uncalibrated HSM SPFs, calibrated HSM SPFs, and jurisdiction-specific SPFs in Regina, Saskatchewan using goodness-of-fit tests. It was found that the HSM typically under-predicted collisions in the city of Regina and it was concluded that the newly developed models provided the best fit to their data. Overall, the new models developed by many researchers showed that the indicative roadway characteristics for predicting collisions in their respective locations were different from the characteristics in the HSM.

2.5.5. HSM2

The state of the research has advanced in the years since the HSM was published in 2010. The second edition of the HSM is expected to be published in 2021 (AASHTO, 2019). New research will be incorporated into the HSM2 and the aim is to improve its usability, accessibility and technical content. The involvement of many transportation professionals is essential to the development of the new manual and the drafts are currently being reviewed by over 140 volunteers (Tillander, 2018).

The crash modification factors (CMFs) in the HSM are required to be updated for the second edition. A project is currently underway by the University of North Carolina, led

by Raghavan Srinivasan, to revise the criteria for which CMFs to include in the second edition of the HSM (Transportation Research Board, 2018).

The goal of the study is to identify and assess all potential CMFs and then to assemble a list of the applicable CMFs for inclusion in HSM2. The research deliverables will include newly developed guidelines for how to incorporate the updated CMFs in HSM2. The research was originally set to be completed by the end of 2020, but a continuation request has been approved and the researchers have stated that phase 2 is underway (Transportation Research Board, 2018).

3. Methodology

The overall approach to the research was to develop, refine, and validate collision prediction models (CPM) to calculate expected numbers of collisions for rural New Brunswick highway segments. The expected collisions were compared with actual collision data to identify locations experiencing more collisions than expected. The available data for traffic volumes, road characteristics, and vehicle collisions on sample sections of rural arterial and collector roads were provided by the New Brunswick Department of Transportation and Infrastructure (NBDTI).

3.1. Data Collection

The data collection largely involved the identification and collection of independent variables to be used as predictors for expected collision frequencies. The New Brunswick road network was subdivided by control sections in order to collect data for each roadway segment separately. The data were collected for control sections that were determined to be relatively homogeneous in character and had not been part of any major upgrade projects within the timeframe of the study. The data that were gathered for each control section included roadway and roadside characteristics, geographic locations, traffic volumes, and collisions to serve as inputs for the development of the collision prediction models.

3.1.1. Representative Sampling

The roadway segments that were included in the model development were chosen in order to provide representation of the full road network of rural New Brunswick arterial and collector roadways. The objective was to provide a broad representation of geographic

location, speed distribution, functional class, and volume distribution. Any segments with intersections were excluded from the sample because they are traditionally modelled as a separate element. Segments of highways that had been reconfigured or rerouted within the timeframe of 2011 to 2016 were also excluded due to lack of consistency in the traffic and road geometry data. Consecutive segments were avoided to allow for a good spread of locations and to avoid any segments with characteristics that were too similar. Due to the restrictions of the collision statistics database, the control sections could not be divided and therefore the data collection was limited to segments that run the entire length of a control section. The segments were chosen to reflect a standard sample of rural highways in New Brunswick.

3.1.2. Data Resources

Resources from NBDTI were used to determine which data were available and the variables were operationally defined for measurement methods. NBDTI was contacted to request traffic volume data, collision history data, road life study diagrams, and ArcGIS Maps data.

The data that were considered in the study design included road geometry characteristics obtained from road life study diagrams and traffic volumes obtained from NBDTI's 2016 Traffic Map for annual average daily traffic (AADT). Alongside the road life study diagrams, NBDTI's GIS Data Warehouse was used to identify segments that fit the criteria for the study. Due to the fact that the most up-to-date road life study diagrams were from the year 2009, variables such as segment length, surface type, and surface age were also collected from the GIS Data Warehouse. Any road segments that had been resurfaced since 2009 required data on surface age and surface type from the GIS Data Warehouse. Any

control sections that had been renamed, reconstructed, or changed lengths in the Control Section Manual between 2009 and 2016 were not included in the selected segments for data collection due to a lack of updated data from the road life study diagrams. The changes to certain control sections created consistency issues between the road life study diagrams and the updated ArcGIS Map data.

Another dataset that was required during model development was the collision statistics dataset obtained through police reports of collisions that occurred on provincial roads during the years 2011 to 2016, excluding 2013 due to lack of data. The collision data are organised by county, district, route number, and control section (as delineated by the Control Section Manual). The collision database includes all reported collisions which represent collisions resulting in a fatality, injury or in property damages greater than $1,000 in total.

Some variables required data that were not readily available from NBDTI and therefore online resources were used for data collection purposes. Google Maps was used to collect data on driveway density, wildlife fencing, and other data requirements for the model development process. A map of Plant Hardiness Zones was also used as a general proxy for the different climate zones in New Brunswick. The zones divide the province based on the geographic regions with certain ranges of average annual minimum winter temperatures. The Plant Hardiness Zones are based on the USDA Hardiness Zone Map Designations and the zones applicable to the highway segments involved in this study are the following (NB Plant Hardiness Zone Map, 2020):

- Zone 3b: -35°F to -30°F (-37°C to -34°C)

- Zone 4a: -30°F to -25°F (-34°C to -32°C)
- Zone 4b: -25°F to -20°F (-32°C to -29°C)
- Zone 5a: -20°F to -15°F (-29°C to -26°C)
- Zone 5b: -15°F to -10°F (-26°C to -23°C)

3.1.3. Potential Explanatory Variables

From the datasets noted above, the variables to be considered for inclusion in the collision prediction models were determined. The list of potential explanatory variables provided in Table 3.1 includes the data resources associated with each variable.

Certain road characteristics were inconsistent throughout the length of some segments. For simplicity and to remain consistent with the control section labelling, surface age, shoulder width, surface and lane width, horizontal curvature, and grade were all calculated as weighted averages over the entire segment. The earliest study year for data collection was 2011; therefore, the surface age was first determined from the GIS Data Warehouse if the year resurfaced was on or before the year 2011. The road life study diagram had to be referenced if the road had been resurfaced more recently than the study year to determine the year of work before 2011 to calculate the surface age. Shoulder width, surface and lane width, horizontal curvature, and grade were all calculated based on the road life study diagrams.

Table 3.1: Potential Explanatory Variables

Variable Type	Variable	Units/Classification	Data Resource(s)
Independent Variables	Road classification	• Arterial • Collector	Control Section Manual
	Traffic volume (AADT)	Vehicles per day	2011-2016 AADT Maps
	Segment length	Kilometres	Road Life Study Diagrams and ArcGIS Data Warehouse
	Surface type	• Asphalt • Chip seal	
	Surface age	Years	
	Directional divider	• Divided • Undivided	Road Life Study Diagrams
	Shoulder width	Metres	
	Lane width	Metres	
	Horizontal curvature	Degrees	
	Grade	Percent	
	Posted speed	Kilometres per hour	
	Plant Hardiness Zone	• Zone 3b • Zone 4a • Zone 4b • Zone 5a • Zone 5b	New Brunswick Plant Hardiness Zone Map
	Driveway density	Driveways per kilometre	Google Maps
	Wildlife fencing	Percentage of segment fenced	
Dependent Variable	Number of expected annual collisions	Discrete number	Determined from the models and compared to actual collisions from NBDTI Collision Statistics spreadsheets

3.2. Collision Prediction Model Development

The data were analyzed through the model development procedure. The models were developed by removing threats of multicollinearity, selecting the appropriate combination of explanatory variables, and assessing the goodness-of-fit of the derived models. The data

collected for a sample of highway locations were used to perform the regression analysis in order to calculate expected collision frequencies in the segments and a separate sample of highway locations was used to complete a case study to demonstrate the application of the models.

3.2.1. Correlation Matrix

The first step in narrowing down the potential explanatory variables for inclusion in the models was a correlational study to reduce the likelihood of multicollinearity. Pearson correlation was used to evaluate the interrelationship of the independent variables.

A correlation matrix was prepared using SAS software to analyze the complete list of independent variables for all of the control sections included in the model development. A strong correlation between variables was considered to exist if the correlation coefficient was 0.8 or greater (Devore, 2011). The statistical significance of any potential explanatory variables found to be strongly correlated was analyzed to determine which should be excluded from the models.

3.2.2. Negative Binomial Regression Analysis

Regression analysis was used to model the effect of the independent variables on collision frequency. The regression procedure statistically determined which variables were strong indicators for the prediction of the number of collisions. The relevant road and traffic characteristics were selected based upon how much explanatory value they had when tested for their relation to number of collisions in combination with the other variables. The models were optimized to solely include indicative variables.

There is a general consensus among researchers for the most appropriate approach for modelling collision data. The state-of-the-art and most generally accepted modelling format is negative binomial regression (Kweon and Lim 2014, Young and Park 2013, Tegge *et al.* 2010, Lord *et al.* 2016, and Saito *et al.* 2011). Negative binomial regression was used due to the non-negative, random integer nature of collision count data and the over-dispersion of the data. Thorough descriptions of the theory behind negative binomial regression are provided by Srinivasan and Bauer (2013) and Saito *et al.* (2011). The regression analysis was completed using SAS software to configure and fit the negative binomial regression models on the training datasets. Two different combinations of models were considered. It was determined that there would either be separate models for 2-lane undivided segments and 4-lane divided segments or just one model with the presence of a directional divider as a variable.

In order to select the explanatory variables to be included in the final models, a sequence of models was fitted, beginning with a simple model with only an intercept term. One additional variable was then included with each successive model (SAS Institute Inc., 2009). The effect of adding in each variable was measured by observing the difference in p-values for the combination of variables. The variables were analyzed to determine whether they had explanatory value by the statistical significance of their associated coefficient p-values. Due to the natural variation in collision data and the objective of accounting for all significant variables, a 90% confidence interval was used. Any variables that produced p-values greater than 0.1 were determined not to have explanatory value and were therefore rejected. Forward selection and backward elimination techniques were both used to determine the optimal combination of explanatory variables. Many iterations were

completed to determine which variables to include. The optimal combination of variables was determined by observing the resulting p-values of each variable to ensure that they were all less than 0.1 to be within the 90% confidence limits. The final models were selected to be the models that maximized the number of variables included while meeting the p-value threshold and producing a good linear regression output of observed versus predicted collisions.

3.2.3. Model Validation

Model validation involved a statistical analysis for goodness-of-fit to determine how well the models had trained on the training datasets. Values of deviance, degrees of freedom, Akaike information criterion (AIC), Bayesian information criterion (BIC), and Pearson Chi-Square were analyzed to compare each successive model. The deviance is defined as "twice the difference between the maximum attainable log likelihood and the log likelihood of the model under consideration" (SAS Institute Inc., 2019). The ratio of deviance to the degrees of freedom was compared with each new model. AIC is the measurement of likelihood for the model fitting the overall data and was used to measure the relative goodness-of-fit between different models. BIC accounts for the number of variables included in the data. The Pearson Chi-Square is a test of the model form that determines if the data are over-dispersed when modeled with a negative binomial distribution (UCLA: Statistical Consulting Group).

Each SAS output included estimate values for the coefficients for each variable and their statistical significance. There was also an estimate of the dispersion parameter which indicated whether the dependent variable was over-dispersed. A Type 3 Analysis was selected in order to analyze the 2-degree-of-freedom Chi-square test and associated p-

values. This was used to determine if each independent variable was a statistically significant predictor of number of collisions.

Once the combination of independent variables was optimized, a goodness-of-fit test was performed to compare observed and predicted collisions using a linear plot. Observed collisions were plotted on the y-axis and predicted collisions were plotted on the x-axis to assess the fit of the model iterations to the actual data. The statistical difference between the slope of the plot and the expected value of 1.0 was calculated at a 90% confidence level. The statistical difference between the intercept and the expected value of zero was also calculated at the same confidence level. The objective was to obtain models that have no statistical difference (the null hypothesis is true), indicating a realistic prediction of collision frequencies. The coefficient of determination (R^2) was also used to analyze the goodness-of-fit. An R^2 value close to one would indicate little variation between observed and predicted collisions.

The model equations could then be written using the coefficient estimates and the negative binomial model form provided in Equations 3.1 and 3.2. The log of the dependent variable is predicted with a linear combination of the independent variables (UCLA: Statistical Consulting Group).

$$\log(N) = Intercept + a_1(V_1) + \cdots + a_n(V_n) \quad [3.1]$$

$$N = e^{Intercept + a_1(V_1) + \cdots + a_n(V_n)} \quad [3.2]$$

Where:

- N = predicted number of collisions
- a = coefficient value
- V = independent variable measurement

3.3. Case Study

Once the models were developed and validated, a case study was completed to test the performance of the models and to demonstrate their use for identifying candidate locations for remediation based on poor safety performance. Statistical analyses were used to interpret the segments experiencing more or less collisions than the models predicted.

A small sample of segments was removed from the dataset prior to model development to be used as a test case. The test case sample was chosen in order to have a range of volumes (low, medium, and high), speeds (80 km/h, 90 km/h, 100 km/h, and 110 km/h), and geographic distribution (North, West, East, and South). The models were run on the test dataset to generate predicted collision counts and compare them with the actual collision counts.

For comparison purposes, the case study involved the contrasting of segments identified using four different techniques. The segments were prioritized by each technique according to how they were performing during the study period in order of worst to best. The collision performance was measured for each of the segments and then they were arranged to determine which segments were performing worse using the following techniques:

1) Collision frequencies from actual collision data

2) Collision rates from actual collision data

3) Potential for Improvement (PFI) values from the newly developed models

4) Potential for Improvement (PFI) values from the Highway Safety Manual (HSM) safety performance functions (SPFs) (AASHTO, 2010 and AASHTO, 2014)

3.3.1. Collision Frequency

The case study segments were first ranked from highest to lowest actual collision frequency. The selection of locations with high collision frequencies for remediation based on observed collisions has been proven to have a bias toward high volume locations.

3.3.2. Collision Rate

To account for the effect of volume on the frequency results, the collision rates were also measured, calculated as frequency divided by volume and segment length. The case study segments were then ranked from highest to lowest based on their collision rates. The use of this technique to identify candidate locations for remediation tends to be biased toward low volume locations. It is based on the assumption that frequency and volume have a linear relationship; however, lower volume locations tend to have higher collision rates.

3.3.3. PFI Values from the Collision Prediction Models

PFI values were calculated as the difference between the expected number of collisions and the actual number of observed collisions. Statistical methods were used to analyze PFI values of the road segments. T-tests were used with a null hypothesis that PFI equals zero at a 95% confidence interval. Results of the statistical significance of each PFI value were

used to identify segments that were performing statistically different than expected. This aided in identifying the locations that have statistically more actual collisions than predicted by the model and were therefore considered to be performing worse than expected.

3.3.4. PFI Values from the HSM SPFs

The SPFs provided in the HSM for 2-lane highway segments (AASHTO, 2010) were applied to the case study data for undivided segments and the SPFs for 4-lane freeway segments (AASHTO, 2014) were applied to the case study data for divided segments. The SPFs are accompanied by crash modification factors (CMFs) to account for jurisdiction-specific conditions.

The HSM procedure is very data intensive and many of the data requirements were not readily available from the NBDTI resources. Additional data collection was required as the variables included in the HSM SPFs and CMFs differed from the previously collected variable data. Google Maps was used to collect the road geometry and roadside data that could be measured or observed. The collected data were used to calculate the relevant CMFs, which were then multiplied by the original SPFs. PFI values were subsequently calculated to compare the SPF results to the actual number of collisions. The calibration procedure was not applied due to the lack of availability of appropriate local data. The minimum sample size in the HSM for the calibration dataset is 30 to 50 segments, which is greater than the number of segments in the case study.

4. Analysis and Results

The implementation of the collision prediction model development process resulted in two separate models for divided and undivided segments. The explanatory variables were selected for each model using SAS software and the available data from NBDTI. The model validation process involved a case study to apply the models to different segments that were not included in the model development dataset.

4.1. Sample of Roadway Segments

The segments for model development and for the case study were chosen to be a representative sample of rural highway control sections in New Brunswick. The characteristics that were considered to ensure proper representation were geographic location, speed limit, road classification, and traffic volume along the segments.

4.1.1. Location of Segments

Data were collected and analyzed for 55 roadway segments in total. A total of 50 segments were used in the model development procedure (11 divided segments and 39 undivided segments). The remaining 5 segments were excluded from the model development dataset to be used in the case study (1 divided segment and 4 undivided segments).

A map of the locations of the included segments is provided in Figure 4.1. A full list of all the control sections can be found in Table A.1 and Table A.2 in Appendix A.

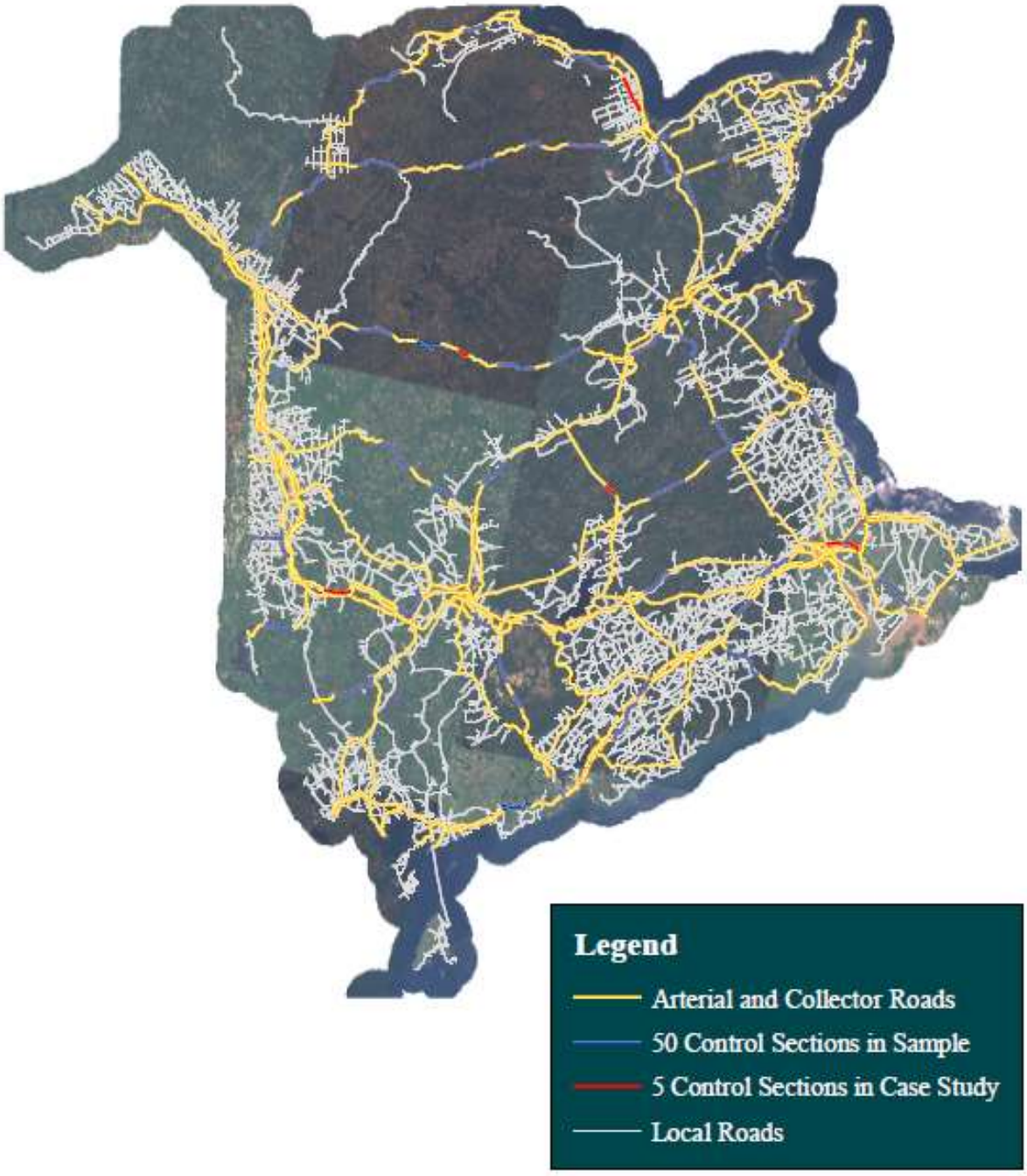

Figure 4.1: Location of selected segments on the GeoNB Map

4.1.2. Representative Sampling

The geographic spread of all 55 segments was examined to ensure that there was at least one segment in each county and a somewhat even distribution between the NBDTI districts. The final sample contained between one and seven segments in each county. A map divided into counties is provided in Figure A.1 in Appendix A and a map divided into districts in provided in Figure A.2. The representation of functional class resulted in a total of 26 arterials and 29 collectors in the final sample of road segments.

The speed distribution was analyzed to determine how many segments were represented at each posted speed limit. It was determined that two segments were in a 70 km/h zone, 20 segments were in a 80 km/h zone, 12 segments were in a 90 km/h zone, 13 segments were in a 100 km/h zone, and eight segments were in a 110 km/h zone. This was assumed to be a good representation of the distribution of posted speed limits on highways throughout the province.

The volume distribution was examined in bins to compare the number of segments with low traffic volume (<8,000 vehicles per day), medium traffic volume (8,000 to 15,000 vehicles per day), and high traffic volume (>15,000 vehicles per day). The distribution was skewed with the largest proportion of segments in the low-volume category. This was consistent with the overall distribution of volumes for all of the volume data available for provincial roads in the study years 2011-2016 (excluding 2013). The traffic volume values for each segment are presented in Tables B.1, B.2, and B.3 in Appendix B for each study year.

The segments that were removed from the full dataset to be used in the case study were chosen somewhat randomly but with the goal of representing typical segments within the sample. The case study comprised of two segments on arterial roads and three segments on collector roads. The locations of the case study segments are presented in red in Figure 4.1.

4.2. Collision Prediction Model Development

The collision prediction models were developed in SAS for undivided and divided segments separately. Tables of variables were required to be formatted such that they could be uploaded into SAS to be analyzed. A correlation matrix was developed to analyze the relationship between the independent variables before using generalized linear modeling to refine the models to only include explanatory variables. The models were finalized by running linear regression to test how well the models were predicting collisions.

4.2.1. Variable Data

All of the available variable data from the road life study diagrams, the AADT maps, the ArcGIS Data Warehouse, the New Brunswick plant hardiness zone map, Google Maps, and the collision statistics spreadsheets were compiled into tables of variables for each model and for the case study. The tables of variables are provided in Tables B.1 to B.3 in Appendix B.

4.2.2. Correlation Matrix

A correlational study was completed to determine the relationship between the independent variables. In the case that two variables were strongly correlated, one of the variables was excluded from the models to reduce the likelihood of multicollinearity.

Using the CORR procedure in SAS University Edition, 14 variables were input to produce a correlation matrix. The variable values for all 50 segments being used for model development were included in the procedure. A strong correlation was identified if the correlation coefficient was 0.8 or greater between two variables. The Pearson correlation coefficients were output in a table and the resulting correlation matrix is presented in Table C.1 in Appendix C.

The road classification variable was determined to have a high correlation with both shoulder width and posted speed. This is logical as arterial roads have wider shoulders and higher posted speeds than collectors so those variables would have a similar effect on collision frequency to road classification. Road classification was consequently removed as an independent variable in the following steps of the model development procedure.

4.2.3. Negative Binomial Regression

Regression analysis was used to model the effect of the independent variables on collision frequency. Negative binomial regression was chosen to determine which variables had explanatory value to model the expected collisions. The optimal combinations of variables were selected by determining the models that solely included indicative variables.

For the purposes of the statistical analysis, the dataset was required to contain non-negative integer count values. The dependent variable is the count of collisions occurring on New Brunswick provincial road segments. Instead of modeling the average collisions per year using average volume and collision data, the data for each year were separated into individual observations. The training dataset consisted of 50 roadway segments with five years of collision data and traffic data for each segment, thus resulting in a total of 250

observations for model development. The five case study segments were also separated into yearly observations, thus resulting in a total of 25 observations for the case study.

The generalized linear model task (GENMOD procedure) in SAS University Edition was used to perform the negative binomial modelling analysis. The response variable was selected to be the actual collisions and was regressed against different combinations of the independent variables. Plant hardiness zone and surface type were selected as classification variables and the remaining variables were selected as continuous variables. Along with the intercept term, each of the explanatory variables were added as model effects.

Both forward selection and backward elimination techniques were used to refine the models and select the best combinations of variables. Forward selection began with a model with only an intercept term and the variables were added one at a time. Backward elimination began with all of the variables included and the variables were removed one at a time according to their statistical significance. The combinations of variables were modified after each successive model run to only include variables that were statistically significant according to the coefficient p-values with 90% confidence limits.

It was attempted to develop just one model with all 250 observations included with the presence of a directional divider as a categorical variable. Separate models were also developed for divided segments (55 observations) and undivided segments (195 observations). The best results were achieved when divided and undivided segments were modeled separately. Once each model analysis was run in SAS, the results tables were displayed, and the results of the final models are provided in Tables 4.1 and 4.2 with detailed output tables in Appendix D. The final models maximized the number of variables

included while ensuring that each variable was statistically significant. The goodness-of-fit between successive models was measured by comparing deviance, Akaike information criterion (AIC), Bayesian information criterion (BIC), and Pearson Chi-Square. Each of these values was attempted to be minimized while also meeting the individual variable requirements.

The final model for divided segments included seven independent variables: plant hardiness zone, AADT, segment length, grade, posted speed, lane width, and surface age. The final model for undivided segments also included seven independent variables: plant hardiness zone, surface type, grade, surface age, AADT, segment length, and wildlife fencing.

Table 4.1: Generalized Linear Model Results for Divided Segments

Analysis of Maximum Likelihood Parameter Estimates and LR Statistics for Type 3 Analysis		
Parameter	**Estimate**	**Pr > ChiSq**
Intercept	-12.7743	0.056
Plant Hardiness Zone 4b	1.998	<0.0001
Plant Hardiness Zone 5a	0.73	
Plant Hardiness Zone 5b	0	
AADT/1000	0.0516	0.0098
Length	0.1234	<0.0001
Grade	1.0131	0.0017
Posted speed	0.1342	0.0092
Lane width	-1.2016	0.0026
Surface age	0.0498	0.043
Dispersion	0.0391	

The variables that were determined to be statistically significant are listed within the parameter column of Tables 4.1 and 4.2. The estimate column provides the values of the coefficients for each variable. For example, since the estimate value is 0.1234 for length in

the divided segments model, for each one-kilometre increase in length, the expected log count of the number of collisions increases by 0.1234.

Table 4.2: Generalized Linear Model Results for Undivided Segments

Analysis of Maximum Likelihood Parameter Estimates and LR Statistics for Type 3 Analysis		
Parameter	**Estimate**	**Pr > ChiSq**
Intercept	-0.7669	0.0232
Plant Hardiness Zone 3b	-0.0057	0.0068
Plant Hardiness Zone 4a	-0.0522	
Plant Hardiness Zone 4b	0.3401	
Plant Hardiness Zone 5a	-0.1797	
Plant Hardiness Zone 5b	0	
Surface Type Chip Seal	-0.8741	<0.0001
Surface Type Asphalt	0	
Grade	0.1377	0.0142
Surface age	-0.0226	0.0395
AADT/1000	0.2435	<0.0001
Length	0.116	<0.0001
Wildlife fencing	-0.0086	0.0031
Dispersion	0.0858	

It is to be noted that the coefficient estimate for surface age was determined to be positive in the divided model and negative in the undivided model. For undivided segments, this suggests that the older the road surface, the better the collision performance. The negative coefficient is consistent with previous results obtained on the relationship between pavement condition and road safety on undivided roads in New Brunswick (Ghanbari, 2017). A negative correlation between pavement condition and collisions was determined, indicating that as pavement deteriorates, all other significant variables remaining constant, collisions are expected to be reduced (Ghanbari, 2017). The positive coefficient estimate for surface age for divided segments could be attributed to more stringent maintenance

requirements for divided highways. Surface condition data were not available for the study period and therefore surface age was used as a proxy.

The dispersion coefficients (α) were analyzed for both models to determine if negative binomial was the appropriate model form. Both of the dispersion coefficients were determined to be statistically different from zero as proven by the Wald 90% Confidence Limits in Tables D.1 and D.2 in Appendix D, therefore it was confirmed that there was over-dispersion in the data and the negative binomial model was more appropriate than the Poisson model (UCLA: Statistical Consulting Group). Over-dispersion is present when the variance is greater than the mean and the Poisson distribution assumes that the variance of the data is equal to the mean (Date, 2019 and NCSS Statistical Software). Over-dispersion was also confirmed by analysing the ratio of deviance to degrees of freedom. Since the ratios were greater than 1 in both of the model results, it was confirmed the data were over-dispersed.

The resulting model equations from the modeling outputs are presented in Equations 4.1 and 4.2.

Model Equation for Divided Segments:

$$N_D = e^{-12.7743+1.998Z4B+0.73Z5A+0.0516AADT+0.1234L+1.0131G+0.1342PS-1.2016LW+0.0498SA} \quad [4.3]$$

Model Equation for Undivided Segments:

$$N_U = e^{-0.7669-0.0057Z3B-0.0522Z4A+0.3401Z4B-0.1797Z5A-0.8741ST+0.1377G-0.0226SA+0.2435AADT+0.116L-0.0086WF} \quad [4.4]$$

Where:

- N = predicted number of collisions for roadway segment per year
- Z3B = Plant hardiness zone 3b (1 if segment is in zone 3b, 0 otherwise)

- Z4A = Plant hardiness zone 4a (1 if segment is in zone 4a, 0 otherwise)
- Z4B = Plant hardiness zone 4b (1 if segment is in zone 4b, 0 otherwise)
- Z5A = Plant hardiness zone 5a (1 if segment is in zone 5a, 0 otherwise)
- AADT = Annual average daily traffic (thousands of vehicles per day)
- L = Length (km)
- G = Grade (weighted average in %)
- PS = Posted speed (km/h)
- LW = Lane width (metres)
- SA = Surface age (years)
- ST = Surface type (Chip seal = 1, Asphalt = 0)
- WF = Wildlife fencing (percentage of segment fenced)

An output dataset was created from each final model in order to be used in the linear regression analysis to measure the goodness-of-fit of the models.

4.2.4. Linear Regression

To measure how well the models were predicting collisions, linear regression analysis was used to determine the fit of the model iterations to actual data. The goodness-of-fit between observed and predicted collisions was evaluated using linear plots. The linear regression task (REG procedure) in SAS University Edition was used to specify the dataset which had been previously created from the output of the generalized linear model task. The dependent variable was defined as the actual collisions collected from the NBDTI collision statistics spreadsheets. The explanatory variable was defined as the predicted collisions

from the model output. Along with the intercept term, the explanatory variable was added as a model effect. A confidence level of 90% was selected for the parameter estimates.

The model for divided segments had a linear regression R-squared of 0.679, adjusted to 0.673 to account for any variance explained simply due to chance. The model for undivided segments had a linear regression R-squared of 0.543, adjusted to 0.541. The intercept was determined for both models as well as the slope which is represented by the parameter estimate of the predicted value in Tables 4.3 and 4.4. Detailed output tables and plots are provided in Appendix E.

Table 4.3: Linear Regression Results for Divided Segments

Variable	Parameter Estimate	90% Confidence Limits	
Intercept	0.060	-0.975	1.096
Predicted Value	0.988	0.832	1.145

Table 4.4: Linear Regression Results for Undivided Segments

Variable	Parameter Estimate	90% Confidence Limits	
Intercept	0.175	-0.218	0.568
Predicted Value	0.938	0.836	1.041

For both models, the intercept was determined to be not statistically different from zero and the slope was determined to be not statistically different from one at a 90% confidence level. This indicates that the models provided a realistic prediction of collision frequencies.

The actual and predicted collisions were plotted in Figures 4.2 and 4.3 using Microsoft Excel and a line of best fit was applied to measure the goodness-of-fit. The equation on the chart in Figure 4.2 indicates a slope of 0.9878 and an intercept of 0.0602, which are very similar to the results from SAS; any difference is assumed to be from rounding error. The equation on the chart in Figure 4.2 indicates a slope of 0.9383 and an intercept of 0.175.

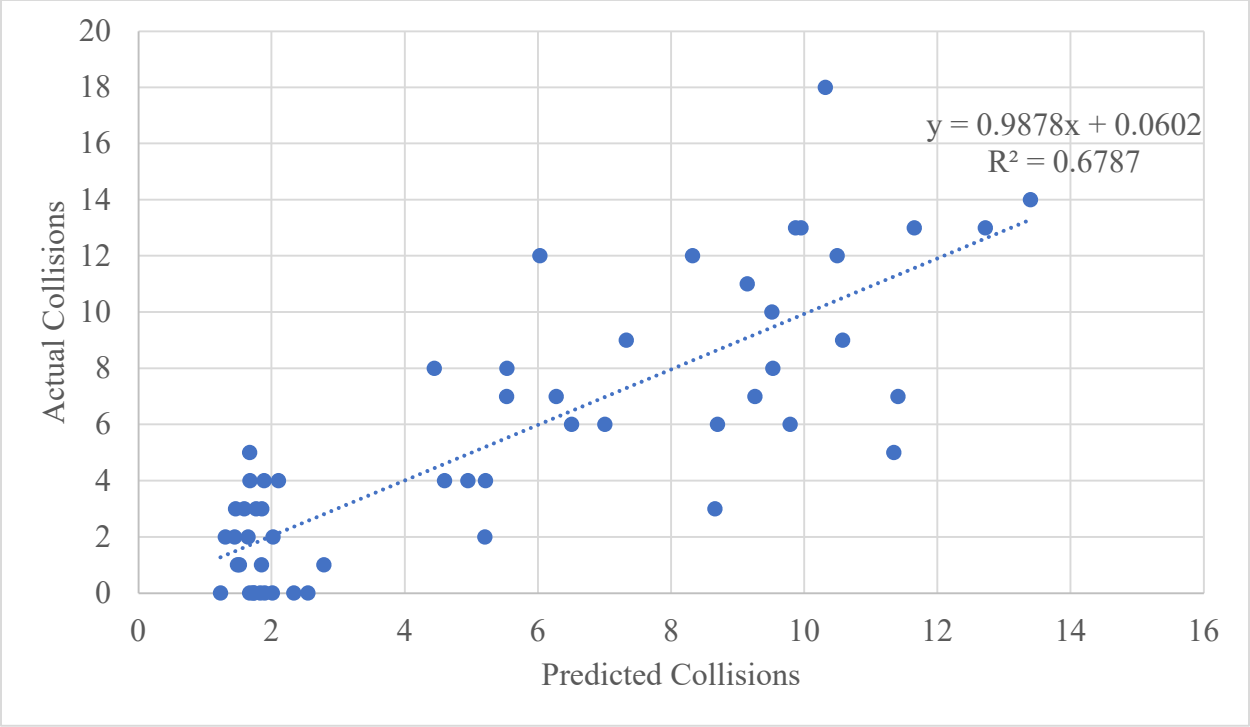

Figure 4.2: Goodness-of-fit Plot for Divided Segments

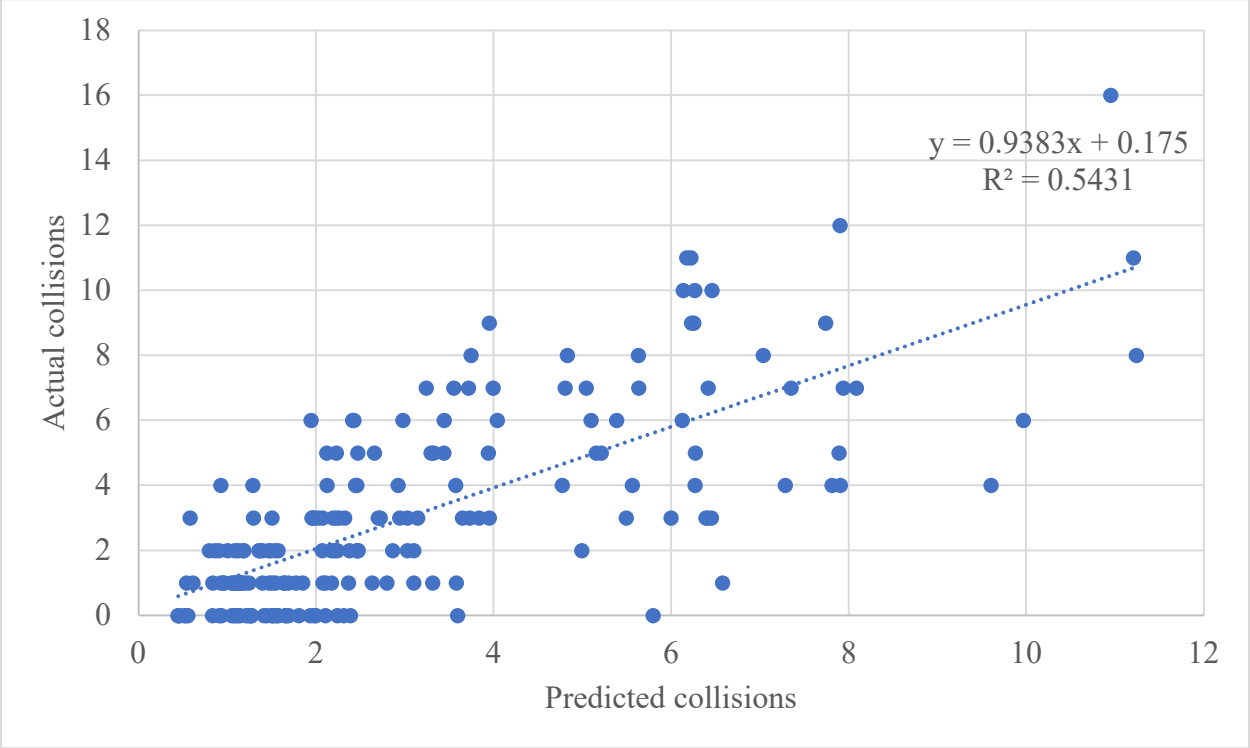

Figure 4.3: Goodness-of-fit Plot for Undivided Segments

4.3. Case Study

The case study was completed to demonstrate the application of the models to roadway segments outside of the sample dataset. Predicted collision counts were generated using the model equations in order to be compared with the actual collision counts. The segments in the case study were ranked to determine which were performing worse than the models predicted. The intent of the case study is to illustrate the application of four different techniques. The techniques that were used to rank the segments are as follows:

1) Collision frequencies from actual collision data
2) Collision rates from actual collision data
3) Potential for Improvement (PFI) values from the newly developed models
4) PFI values from the Highway Safety Manual (HSM) safety performance functions (SPFs) (AASHTO, 2010 and AASHTO, 2014)

4.3.1. Collision Frequency

The actual collision frequencies of the case study segments were collected from the collision data provided by the New Brunswick Department of Transportation and Infrastructure. The five divided case study observations were ranked from highest to lowest collision frequency in Table 4.5 and the 20 undivided case study observations were ranked from highest to lowest collision frequency in Table 4.6. The average annual collision frequencies of all five case study segments were also calculated to provide a ranking of the segments based on five years of AADT and collision data from 2011-2012 and 2014-2016 in Table 4.7. The AADT columns in Tables 4.5 to 4.7 indicate that the prioritization of

road locations with high collision frequencies for remediation is biased toward high volume locations.

Table 4.5: Divided Case Study Observations Ranked by Collision Frequency

Year	Route	Control Section	AADT	Actual Annual Collisions
2016	2	60	23200	14
2011	2	60	21300	13
2014	2	60	21400	13
2015	2	60	22500	12
2012	2	60	21500	9

Table 4.6: Undivided Case Study Observations Ranked by Collision Frequency

Year	Route	Control Section	AADT	Actual Annual Collisions
2011	11	27	3860	8
2014	11	27	3760	7
2016	11	27	4150	7
2015	11	27	4010	4
2012	11	27	3840	3
2012	105	14	2290	3
2011	105	14	2310	2
2016	105	14	2380	2
2012	123	2	350	2
2015	105	14	2380	1
2011	108	10	490	1
2015	123	2	510	1
2016	123	2	500	1
2014	105	14	2380	0
2012	108	10	430	0
2014	108	10	430	0
2015	108	10	680	0
2016	108	10	680	0
2011	123	2	350	0
2014	123	2	510	0

Table 4.7: Case Study Segments Ranked by Collision Frequency (Annual Average)

Directional Divider	Route	Control Section	AADT	Average Actual Annual Collisions
Divided	2	60	21980	12.2
Undivided	11	27	3924	5.8
	105	14	2348	1.6
	123	2	444	0.8
	108	10	542	0.2

Analyzing the average annual data is more common practice than analyzing each year of collision data separately to eliminate any biases from anomalies. Using more than just one year of collision data allows for a more accurate representation of the actual collision performance along the road segments. For example, the data for Route 105 Control Section 14 in Table 4.6 indicate that zero collisions occurred in 2014 but the average annual collisions over five years was determined to be 1.6 in Table 4.7; therefore, 2014 was not a good representation of the collision experience along that road segment.

4.3.2. Collision Rate

Due to the effect of AADT on the collision frequency results, the collision rates were also measured by dividing frequency by volume and segment length. Equation 4.3 was used to calculate collision rates for each highway segment observation in the case study, representing the average number of collisions per million vehicle-kilometers.

$$\boldsymbol{R = \frac{C(1,000,000)}{365(AADT)(L)}} \quad [4.3]$$

Where:

- R = collision rate (collisions/million veh-km)

- C = number of actual collisions
- AADT = Annual Average Daily Traffic (vehicles per day)
- L = Segment length (km)

The 20 undivided case study observations were ranked from highest to lowest collision rate in Table 4.8 and the five divided case study observations were ranked from highest to lowest collision rate in Table 4.9.

Table 4.8: Undivided Case Study Observations Ranked by Collision Rate

Year	Route	Control Section	Length	AADT	Actual collisions	Collision rate (collisions/million vehicle-kilometers)
2012	123	2	4.889	350	2	3.202
2011	108	10	4.352	490	1	1.285
2016	123	2	4.889	500	1	1.121
2015	123	2	4.889	510	1	1.099
2011	11	27	14.452	3860	8	0.393
2014	11	27	14.452	3760	7	0.353
2012	105	14	10.514	2290	3	0.341
2016	11	27	14.452	4150	7	0.320
2011	105	14	10.514	2310	2	0.226
2016	105	14	10.514	2380	2	0.219
2015	11	27	14.452	4010	4	0.189
2012	11	27	14.452	3840	3	0.148
2015	105	14	10.514	2380	1	0.109
2014	105	14	10.514	2380	0	0.000
2012	108	10	4.352	430	0	0.000
2014	108	10	4.352	430	0	0.000
2015	108	10	4.352	680	0	0.000
2016	108	10	4.352	680	0	0.000
2011	123	2	4.889	350	0	0.000
2014	123	2	4.889	510	0	0.000

Table 4.9: Divided Case Study Observations Ranked by Collision Rate

Year	Route	Control Section	Length	AADT	Actual collisions	Collision rate (collisions/million vehicle-kilometers)
2011	2	60	14.522	21300	13	0.115
2014	2	60	14.522	21400	13	0.115
2016	2	60	14.522	23200	14	0.114
2015	2	60	14.522	22500	12	0.101
2012	2	60	14.522	21500	9	0.079

The AADT columns in Tables 4.8 and 4.9 indicate that the prioritization of road locations with high collision rates for remediation is biased toward low volume locations, with the exception of the segments with zero actual collisions in a given study year in Table 4.8. This technique assumes that collision frequency and traffic volume have a linear relationship but many of the lower volume locations have higher collision rates.

The average annual collision rates of all five case study segments were also calculated to provide a ranking of the segments based on five years of AADT and collision data in Table 4.10. Using several years of collision data to calculate the collision rate provides more reliable results. For example, the data for Route 108 Control Section 10 in Table 4.8 indicate that the collision rate was 1.285 collisions per million vehicle-kilometers in 2011, which ranked it as second priority in the sample of 20 observations. This would indicate that the segment is performing worse than Route 11 Control Section 27. However, the average collision rate over five years was calculated to be 0.232 collisions per million vehicle-kilometers in Table 4.10, which placed it as a lower priority than Route 11 Control Section 27.

Table 4.10: Case Study Segments Ranked by Collision Rate (Annual Average)

Directional Divider	Route	Control Section	Length	AADT	Average actual annual collisions	Collision rate (collisions/million vehicle-kilometers)
Undivided	123	2	4.889	444	0.8	1.010
	11	27	14.4519	3924	5.8	0.280
	108	10	4.352	542	0.2	0.232
	105	14	10.514	2348	1.6	0.178
Divided	2	60	14.522	21980	12.2	0.105

4.3.3. PFI Values from the Collision Prediction Models

The intent of applying the newly developed models to the case study segments was to illustrate how the models can aid the user to identify the most problematic road locations. Identifying the most problematic control sections allows for a justification of where to allocate investments in improvement projects. The models can be applied to a small or large sample of segments in order to prioritize the road sections that are performing worse than expected based on their Potential for Improvement (PFI) values. It is also possible to apply one of the CPMs (divided or undivided) to a singular segment to determine an expected number of collisions and compare the result with the actual collision history data for that segment.

PFI values were calculated based on the results of applying the models to the case study road segments. PFI values are defined as the difference between the expected (or modelled) number of collisions and the actual number of collisions. The necessary variable data for applying the models, the resulting expected collisions, and the PFI results are presented in Table F.1 in Appendix F. A positive PFI indicates that the segment is performing worse

than expected and a negative PFI indicates that the segment is performing better than expected according to the newly developed CPMs.

The five divided case study observations were ranked from highest to lowest PFI in Table 4.11 and the 20 undivided case study observations were ranked from highest to lowest PFI in Table 4.12.

The data values in Table 4.11 show that the collisions were greatly underestimated by the divided model (around 5 to 7 collisions). This resulted in high PFI values, indicating that this segment (Route 2 Control Section 60) is performing much worse than expected. Upon further investigation of the data used to develop the divided model in comparison with the data used for the divided segment in the case study, a problem was identified with the grade values. The weighted averages of the grades in the original dataset ranged from 1.02% to 3.29% and the weighted average of the grade of the divided segment in the case study (Route 2 Control Section 60) was 0.91%. This grade value is not within the range of the grades used in the model development and therefore may be the cause of the underestimation of the annual collisions for that segment and the associated high PFI values.

Table 4.11: PFI Values from the Divided Collision Prediction Model

Year	Route	Control Section	Actual collisions	Expected collisions	PFI
2011	2	60	13	4.61	8.39
2014	2	60	13	5.38	7.62
2016	2	60	14	6.52	7.48
2015	2	60	12	5.99	6.01
2012	2	60	9	4.90	4.10

Table 4.12: PFI Values from the Undivided Collision Prediction Model

Year	Route	Control Section	Actual collisions	Expected collisions	PFI
2011	11	27	8	5.79	2.21
2012	105	14	3	1.13	1.87
2014	11	27	7	5.33	1.67
2016	11	27	7	5.60	1.40
2011	105	14	2	1.16	0.84
2016	105	14	2	1.17	0.83
2012	123	2	2	1.33	0.67
2011	108	10	1	0.75	0.25
2015	105	14	1	1.20	-0.20
2016	123	2	1	1.26	-0.26
2015	123	2	1	1.30	-0.30
2014	108	10	0	0.69	-0.69
2016	108	10	0	0.70	-0.70
2015	108	10	0	0.72	-0.72
2012	108	10	0	0.72	-0.72
2014	105	14	0	1.23	-1.23
2014	123	2	0	1.33	-1.33
2011	123	2	0	1.36	-1.36
2015	11	27	4	5.54	-1.54
2012	11	27	3	5.69	-2.69

The average annual expected collisions of all five case study segments were also calculated to provide a ranking of the segments based on five years of AADT and collision data in Table 4.13. A comparison of Tables 4.12 and 4.13 shows that in one given year (2011), Route 11 Control Section 27 would be selected as the worst-performing undivided segment, but if the five-year average is considered, then Route 105 Control Section 14 is identified as the highest priority for the undivided segments. This is despite the higher number of actual annual collisions that have occurred on Route 11 Control Section 27. As indicated by the expected annual collisions for Route 11 Control Section 27 in Table 4.13,

it was expected that this segment would experience 5.57 collisions per year based on the road geometric characteristics, traffic volumes, and roadside conditions that were included in the variables in the undivided CPM. This resulted in a PFI value that was not as high as the PFI value for Route 105 Control Section 14.

The PFI values in Table 4.13 were analyzed using t-tests to determine their statistically significant difference from the expected value of zero. The yearly observations of the PFI values from Tables 4.11 and 4.12 were analyzed to interpret whether the average PFI for each segment in Table 4.13 was statistically greater or less than zero. This aided in identifying the priority locations that have more actual collisions than predicted by the models. Positive PFI values that are statistically significant indicate locations that are performing worse than expected and would therefore benefit from investment in a treatment. The null hypothesis for each t-test was that the annual average PFI for that segment was equal to zero at a 95% confidence level. The t-stat value was calculated and compared to the t-critical to determine whether to reject the null hypothesis. This analysis allowed the formation of conclusions on the performance of the case study segments. More details on the statistical test results are provided in Tables F.2 and F.3 in Appendix F.

The divided segment (Route 2 Control Section 60) was determined to have a PFI value that was statistically greater than zero, indicating that the segment was performing worse than expected. Three of the undivided segments were determined to have PFI values that were not significantly different from zero. These segments would therefore not be identified as priorities for remediation because the annual collision experience was as expected. Route 108 Control Section 10 was determined to have a PFI value that was statistically less than zero, indicating that the segment was performing better than expected.

Table 4.13: PFI Values from the Collision Prediction Models (Annual Average)

Directional divider	Route	Control Section	Average actual annual collisions	Expected annual collisions	PFI	t-stat
Divided	2	60	12.2	5.48	6.72	8.84
Undivided	105	14	1.6	1.18	0.42	0.80
	11	27	5.8	5.59	0.21	0.22
	123	2	0.8	1.32	-0.52	-1.36
	108	10	0.2	0.72	-0.52	-2.69

4.3.4. PFI Values from the HSM SPFs

The SPFs in the Highway Safety Manual (HSM) were applied to the case study segments to compare with the results of the newly developed collision prediction models. The predictive model for rural two-lane, two-way undivided roadway segments from Chapter 10 (AASHTO, 2010) was applied to the undivided segments and the predictive model for rural four-lane freeway segments from Chapter 18 (AASHTO, 2014) was applied to the divided segments. Crash modification factors (CMFs) from the HSM were multiplied by the SPF results to account for site-specific conditions. The data requirements for applying the HSM procedure differed from the dataset that was already compiled to apply the newly developed models due to the different variables included in the HSM CMFs to predict collisions. Additional data collection was required, and Google Maps was used to collect all road geometry and roadside data that could be measured or observed. A full list of the case study data that were collected for the HSM SPFs can be found in Tables F.4 to F.8 in Appendix F.

The CMFs that were included for 2-lane highways were lane width (CMF_{1r}), shoulder width (CMF_{2r}), horizontal curves (CMF_{3r}), grade (CMF_{5r}), driveway density (CMF_{6r}), and

roadside hazard rating (CMF_{10r}). The CMF for superelevation (CMF_{4r}) was not included due to lack of data and the remaining CMFs were not applicable to the segments in the case study.

The CMFs that were included for 4-lane freeways were horizontal curve (CMF_1), lane width (CMF_2), inside shoulder width (CMF_3), median width (CMF_4), outside shoulder width (CMF_8), shoulder rumble strips (CMF_9), outside clearance (CMF_{10}), and outside barrier (CMF_{11}). The CMF for median barrier (CMF_5) was not applicable to the segments in the case study. The CMF for high volume (CMF_6) was not included due to lack of volume concentration data and the CMF for lane change (CMF_7) was not included due to lack of volume data for lane change activity on entrance and exit ramps.

Calibration of the models was not practical for this small case study as the data requirements to develop calibration factors were not met. The calibration factor for roadway segments (C_r) is based on the ratio of observed collision frequencies to expected average crash frequency. The number of observations in the case study was insufficient to apply a calibration factor. A larger sample of segments would be required to apply the calibration procedure. As mentioned, the process of applying the SPFs and associated CMFs was very data intensive and time consuming and data accessibility was also limited which restricted the size of the complete dataset. The goal was to have the maximum number of segments included in the dataset to develop the models, within a reasonable data collection timeframe. The original dataset of 55 segments (275 yearly observations) was collected and 5 of the segments (25 observations) were removed to be used in the case study. The remaining 50 segments were split into a dataset for divided segments and a dataset for undivided segments to develop the CPMs. A larger case study sample would

have reduced the quantity of segments available for inclusion in the model development procedure.

Potential for Improvement (PFI) values were calculated based on the results of applying the HSM SPFs and CMFs to the case study observations. The results of applying the predictive model for rural two-lane, two-way undivided roadway segments to the 20 undivided segments are provided in Table 4.14. The data presented in Table 4.14 indicate that five of the undivided segments had a positive PFI and the values were skewed with many negative PFI values.

Table 4.14: PFI Values from the HSM rural 2-lane, 2-way undivided SPF

Year	Route	Control Section	Actual collisions	Predicted Crash Frequency	PFI
2012	123	2	2	0.28	1.72
2011	11	27	8	7.34	0.66
2011	108	10	1	0.38	0.62
2016	123	2	1	0.40	0.60
2015	123	2	1	0.41	0.59
2014	11	27	7	7.14	-0.14
2011	123	2	0	0.28	-0.28
2012	108	10	0	0.33	-0.33
2014	108	10	0	0.33	-0.33
2014	123	2	0	0.41	-0.41
2015	108	10	0	0.53	-0.53
2016	108	10	0	0.53	-0.53
2016	11	27	7	7.92	-0.92
2015	11	27	4	7.64	-3.64
2012	11	27	3	7.30	-4.30
2012	105	14	3	7.60	-4.60
2011	105	14	2	7.66	-5.66
2016	105	14	2	7.87	-5.87
2015	105	14	1	7.87	-6.87
2014	105	14	0	7.87	-7.87

The results of applying the predictive model for rural four-lane freeway segments to the divided segments are provided in Table 4.15. As shown by the data values in the table, the predicted crash frequencies for divided segments were overestimated by the SPF (around 22 to 24 collisions). This is the opposite result from the newly developed collision prediction models. It is interesting to note that of the 22-24 predicted annual collisions, around 11-12 annual collisions were predicted to be single-vehicle property damage only (PDO) collisions by the HSM SPFs. The results for the HSM SPFs for 4-lane freeways by severity are included in Table F.9 in Appendix F. These predicted PDO collisions may account for some of the overestimation due to differences in collision reporting thresholds in New Brunswick and the hypothesized number of unreported single-vehicle collisions that do not result in a casualty.

Table 4.15: PFI Values from the HSM rural 4-lane freeway SPF

Year	Route	Control Section	Actual collisions	Predicted Crash Frequency	PFI
2011	2	60	13	22.28	-9.28
2014	2	60	13	22.38	-9.38
2016	2	60	14	24.24	-10.24
2015	2	60	12	23.51	-11.51
2012	2	60	9	22.48	-13.48

The average annual predicted collisions of all five case study segments were also calculated to provide a ranking of the segments based on five years of AADT and collision data in Table 4.16. In Table 4.14, Route 11 Control Section 27 and Route 108 Control Section 10 both had one observation (in 2011) that indicated that the segments were performing worse than expected, but in Table 4.16, the average PFI values are both negative. Route 105

Control Section 14 had consistent collision predictions that significantly exceeded the actual collisions.

The PFI values were analyzed using t-tests to determine their statistically significant difference from the expected value of zero. The data presented in Table 4.16 indicate that none of the PFI values were determined to be statistically greater than zero. The positive PFI value for Route 123 Control Section 2 would identify the segment to be performing worse than expected; however, it was not statistically significantly different from zero. Route 105 Control Section 14 and Route 2 Control Section 60 both had PFI values that were statistically less than zero, indicating that the segments were performing better than expected according to the HSM SPFs. More details on the statistical test results are provided in Table F.2 in Appendix F.

Table 4.16: PFI Values from the HSM SPFs (Annual Average)

Directional divider	Route	Control Section	Actual collisions	Predicted Crash Frequency	PFI	t-stat
Undivided	123	2	0.8	0.36	0.44	1.16
	108	10	0.2	0.42	-0.22	-1.02
	11	27	5.8	7.47	-1.67	-1.71
	105	14	1.6	7.77	-6.17	-11.09
Divided	2	60	12.2	22.98	-10.78	-13.73

4.3.5. Case Study Summary

The case study was used as an illustration of how the new models can be applied as well as how other common techniques can be applied to evaluate the collision performance of road segments. The four different techniques that were used to rank the segment observations in the case study were as follows:

1) Collision frequencies from actual collision data

2) Collision rates from actual collision data

3) Potential for Improvement (PFI) values from the newly developed models

4) PFI values from the Highway Safety Manual (HSM) safety performance functions (SPFs)

In terms of safety performance, the results from each of the approaches identify different road segments to be prioritized. The newly developed models and the HSM models include different variables and relationships between coefficients and thus result in differing ranks of the PFI values. The ranking determined by each of the four approaches are compared in Table 4.17 for undivided segments.

Table 4.17: Ranking of Undivided Case Study Segments from Different Approaches

Route	Control Section	Ranking by New Model	Ranking by HSM Model	Ranking by Collision Frequency	Ranking by Collision Rate
105	14	1	4	2	4
11	27	2	3	1	2
123	2	3	1	3	1
108	10	4	2	4	3

The segments are ranked from 1 to 4, with 1 being the worst. The “worst” segments are defined as the segments that are the best candidates for remediation according to each respective approach. For the modelling approaches, the worst segments have more actual collisions than predicted (i.e. positive potential for improvement (PFI) values). The results presented in Tables 4.12 and 4.14 indicate that the newly developed CPM for undivided

segments resulted in lower PFI values than the HSM SPF, indicating that the expected collision values are closer to the actual collision values.

The worst-performing segment in comparison to the expected collisions according to the new model was Route 105 Control Section 14. The worst-performing segment in comparison to the expected collisions according to the HSM models coincides with the segment with the highest collision rate. This may indicate that the HSM models account for issues with identifying locations using actual collision frequencies such as the bias toward high volume locations. However, the HSM models may over-accommodate for those issues and identify segments too similarly to the collision rate approach which is biased toward low volume locations due to the assumed linear relationship between frequency and volume.

Since there was only one divided segment included in the case study, it was not ranked. The new divided CPM identified Route 2 Control Section 60 as performing worse than expected and the HSM SPF identified the divided segment as performing better than expected. Upon further investigation into the location of Route 2 Control Section 60, it is hypothesized that there may be potential issues with the roadway segment. The segment is located just outside of an urban environment which may render the rural models inapplicable to this segment. The segment also contains a railway crossing which is not included as one of the independent variables in either model. A larger sample of divided segments in the case study would better demonstrate the application of the models and provide more accurate results to test the prediction capabilities of the model.

5. Conclusions and Recommendations

The main objective of this research was to develop collision prediction models for rural highways in New Brunswick to aid with the identification and prioritisation of road segments with unusually poor collision performance. The empirically based models were produced to provide an objective approach to assess the safety of locations within the road network. Research included data collection for potential variables to be included in the models, correlation analysis, regression analysis, model validation, and a case study to illustrate the application of the models.

5.1. Conclusions

A sample of 50 rural highway segments in New Brunswick (250 yearly observations with 5 years of collision data for each segment) was selected to develop the collision prediction models based on local conditions. The sample was selected at random to provide a geographical representation of all rural highway segments in New Brunswick. A total of 14 independent variables were collected for consideration in the predictive models. Pearson correlation was used to evaluate the interrelationship of the independent variables and to reduce the likelihood of multicollinearity.

Regression analysis was then used to model the effect of the independent variables on collision frequency and select the appropriate combination of variables. Negative binomial regression was chosen to determine which variables had explanatory value to model the expected collisions because of its ability to appropriately model over-dispersed count data.

Collision count data for divided and undivided roads were modeled separately to determine which roadway, roadside, and traffic characteristics had the best predictive ability to estimate the number of collisions occurring in the road segments. Results indicate that plant hardiness zone, AADT, segment length, grade, posted speed, lane width, and surface age all had a significant effect on collision frequency for divided segments. Plant hardiness zone, surface type, grade, surface age, AADT, segment length, and wildlife fencing all had a significant effect on collision frequency for undivided segments. The resulting model equations can be used by road authorities to calculate the potential for improvement (PFI) values for road segments to determine their safety performance. Road segments that are most likely to benefit from safety countermeasures correspond with the highest PFI values.

An output dataset was created from each final model in order to be used in the linear regression analysis to measure the goodness-of-fit between observed and predicted collisions. The model for divided segments had a linear regression adjusted R-squared of 0.673 and the model for undivided segments had a linear regression adjusted R-squared of 0.541. For both models, the intercept was determined to be not statistically different from zero and the slope was determined to be not statistically different from one at a 90% confidence level.

A number of studies were reviewed from various jurisdictions to compare the Highway Safety Manual's (HSM) procedure with newly developed SPFs. The studies concluded that jurisdiction-specific models outperformed the HSM models in terms of accuracy. The case study from the current research project allowed for a comparison of the newly developed models to the HSM SPFs. A separate sample of 5 highway segments (25 yearly observations with 5 years of collision data for each segment) was used to complete the case

study to demonstrate the application of the models. Applying both the new models and the HSM SPFs to the case study segments resulted in smaller PFI values from the new models.

The case study involved the ranking of road segments according to different techniques to illustrate how road locations can be prioritized based on their collision performance. The segments were ranked according to collision frequency, collision rate, the results of the new models, and the results of the HSM SPFs. The biases introduced by ranking according to frequency and rate were confirmed. The results from each of the approaches identify different road segments to be prioritized. The newly developed models and the HSM models include different variables and relationships between coefficients and thus result in differing ranks of the PFI values. Most of the segments were determined to be performing better than expected according to the HSM procedure, which could be due to transferability issues.

The new divided CPM identified the divided segment in the case study to be performing worse than expected and the HSM SPF for rural four-lane freeway segments identified the divided segment to be performing better than expected. It appears that the divided segment in the case study was a poor representation of divided roads and cannot be used alone to compare the two models. This specific segment has some unique conditions including its proximity to an urban environment and the presence of a railway crossing. These factors were not incorporated in the variable data to predict the collisions. This segment also has a very flat gradient, which is not typical according to the data for the divided segments used in the model development procedure. Due to the fact that the case study only included data for one divided segment, it could not be determined how well the newly developed model for divided segments performed in terms of accuracy in comparison to the HSM

predictive model for rural four-lane freeway segments. A larger sample of divided highway sections was needed in the case study.

5.2. Recommendations

Recommendations for using the models developed through this research, as well as for future research, are listed as follows:

- The collision prediction models (CPMs) developed in this study are applicable only for rural arterial and collector roads in the province of New Brunswick. It is recommended that any other jurisdiction calibrate the models to reflect local conditions prior to application.

- Priority rankings for the safety performance of road segments determined by the PFI values calculated from applying the CPMs should be interpreted using engineering judgement in order to maximize their effectiveness.

- The CPM for divided highways was developed from segments in plant hardiness zones 4b, 5a, and 5b. It is recommended that new models be developed or that the sample of segments be expanded to redevelop the existing divided CPM for application to divided highway segments in any other zones.

- The case study included a limited number of divided segment observations. A larger sample of divided segments in the case study would better demonstrate the application of the models and provide more accurate results to test the prediction capabilities of the model.

- The HSM models were only applied to the case study segments. Application of the HSM models to the entire dataset used for model development would allow for the calibration procedure to be applied and a more in-depth comparison between the HSM models and the newly developed models. In order for the HSM SPFs to be applicable to future projects and/or decisions, a sample of at least 30 to 50 segments would need to be used to calibrate the models.

    - The severe overestimation of the freeway (divided segments) collisions could be corrected by applying the HSM predictive method to a larger sample of segments along with applying the calibration procedure which accounts for the ratio of observed crashes to predictive crashes of the dataset. This may result in a more appropriate SPF for the local conditions to be applied to other roadways during the study years.

- Future research could involve more models for roads that were outside of the scope of this study. It is recommended that separate models be developed for local rural roads.

- This research was limited to roadway segments; however, future research could apply similar methodologies to develop models for intersections and interchanges.

- Any decisions for countermeasure implementation based on the results of applying the CPMs should be followed up with before-and-after studies of collision data to determine the effectiveness of the new treatments on road safety.

6. References

AASHTO. (2010). *Highway Safety Manual Volume 1 and 2, 1st edition.* Washington, D.C.: The American Association of State Highway and Transportation Officials.

AASHTO. (2014). *Highway Safety Manual Supplement, 1st edition.* Washington, DC: The American Association of State Highway and Transportation Officials.

AASHTO. (2019, November 18). *Highway Safety Manual 2nd Edition Update.* Retrieved from Highway Safety Manual : http://www.highwaysafetymanual.org/Documents/HSM2_GeneralUpdate_20191113.pdf

Bahar, G. (2014). *User's Guide to Develop Highway Safety Manual Safety Performance Function Calibration Factors.* National Cooperative Highway Research Program (NCHRP).

Bi, C., Ma, X., Zhang, Y., Wang, Y. (2014). Developing A Cluster-based Algorithm for Collision Hotspot Identification. *COTA International Conference of Transportation Professionals*, (pp. 2381 - 2395). Changsha. doi:https://doi.org/10.1061/9780784413623.228

Bonneson, J.A., Geedipally, S., Pratt, M.P., and Lord, D. (2012). *Safety Prediction Methodology and Analysis Tool for Freeways and Interchanges.* Washington, D.C.: Final Report for NCHRP Project 17-45.

Brimley, B. K., M. Saito, & G. G. Schultz. (2012). *Calibration of Highway Safety Manual Safety Performance Function.* Transportation Research Record 2279: Journal of the Transportation Research Board.

Date, S. (2019, October 6). *Negative Binomial Regression: A Step by Step Guide.* Retrieved from Towards Data Science: https://towardsdatascience.com/negative-binomial-regression-f99031bb25b4

Devore, J. L. (2011). *Probability and Statistics for Engineering and the Sciences (Eigth Edition).* Cengage Learning.

Dong, N., Huang, H., Lee, J., Gao, M., Abdel-Aty, M. (2016). Macroscopic hotspots identification: A Bayesian spatio-temporal interaction approach. *Accident Analysis & Prevention, Volume 92*, 256-264. doi:https://doi.org/10.1016/j.aap.2016.04.001

Gelman, A., Carlin, J. B., & Stern, H. S. (1995). *Bayesian Data Analysis, Second Edition.* Chapman & Hall/CRC Press. ProQuest Ebook Central, http://ebookcentral.proquest.com/lib/unb/detail.action?docID=199128.

Ghanbari, A. B. (2017). *Impact of Pavement Condition Indicators on Road Safety in New Brunswick.* The University of New Brunswick.

Government of New Brunswick (GNB). (2019). *Transportation and Infrastructure Annual Report 2018-2019.*

Government of New Brunswick (GNB). (2020). *2019-2020 Capital Budget Estimates.* https://www2.gnb.ca/content/gnb/en/departments/dti/publications/content/2019-2020/capital_budget_estimates_2019-20.html.

Kweon, Y.-J. and Lim, I.-K. (2014). *Development of Safety Performance Functions for Multilane Highway and Freeway Segments Maintained by the Virginia Department of Transportation.* Charlottesville: Virginia Center for Transportation Innovation & Research.

Kwon, O.H., Park, M.J., Yeo, H., Chung, K. (2013). Evaluating the performance of network screening methods for detecting high collision concentration locations on highways. *Accident Analysis & Prevention, Volume 51*, 141-149. doi:https://doi.org/10.1016/j.aap.2012.10.019

Lord, D., & Mannering, F. (2010). *The statistical analysis of crash-frequency data: A review and assessment of methodological alternatives.* Transportation Research Part A: Policy and Practice, 44(5), 291–305. https://doi.org/10.1016/j.tra.2010.02.001.

Lord, D., Geedipally, S. and Shirazi, M. (2016). *Improved Guidelines for Estimating the Highway Safety Manual Calibration Factors.* College Station, TX: Texas A&M Transportation Institute.

Lougheed, Peter James. (2016). *Road Safety Audits: Quantifying and Comparing the Benefits and Costs for Freeway Projects* . PhD Thesis, Department of Civil Engineering, University of New Brunswick, Fredericton, 99-101.

Lubliner, H. (2011). *Evaluation of the Highway Safety Manual Crash Prediction Model for Rural Two-lane Highway Segments in Kansas.* The University of Kansas.

NCSS Statistical Software. (n.d.). *Negative Binomial Regression, Chapter 326*. Retrieved August 19, 2020, from https://ncss-wpengine.netdna-ssl.com/wp-content/themes/ncss/pdf/Procedures/NCSS/Negative_Binomial_Regression.pdf

New Brunswick Plant Hardiness Zone Map. (2020). Retrieved from PlantMaps: https://www.plantmaps.com/interactive-new-brunswick-plant-zone-hardiness-map.php

Qin, X., Chen, Z., & Shaon, R. R. (2018). *Developing jurisdiction-specific SPFs and crash severity portion functions for rural two-lane, two-way intersections.* Journal of Transportation Safety & Security, 1–13. https://doi.org/10.1080/19439962.2018.1458052.

Sacchi, E., Persaud, B., & Bassani, M. (2012). *Assessing International Transferability of Highway Safety Manual Crash Prediction Algorithm and Its Components.* Transportation Research Record, 2279(1), 90–98.

Sacchi, E., Sayed, T., El-Basyouny, K. (2015). Multivariate Full Bayesian Hot Spot Identification and Ranking: New Technique. *Transportation Research Record*, 1-9. doi:http://journals.sagepub.com/doi/full/10.3141/2515-01

Saito, M., Brimley, B. K., and Schultz, G.G. (2011). *Transportation Safety Data and Analysis Volume 2: Calibration of the Highway Safety Manual and Development of New Safety Performance Functions.* Utah Department of Transportation.

SAS Institute Inc. (2009). *SAS/STAT ® 9.2 User's Guide, Second Edition.* Cary, NC: SAS Institute Inc.

SAS Institute Inc. (2019). *SAS/STAT 15.1 User's Guide.* Cary, NC: SAS Institute Inc. .

Srinivasan, R. and Bauer, K. (2013). *Safety Performance Function Development Guide: Developing Jurisdiction-Specific SPFs.* Washington: Federal Highway Administration Office of Safety.

Srinivasan, R., Carter, D., and Bauer, K. (2013). *Safety Performance Function Decision Guide: SPF Calibration vs SPF Development.* The University of North Carolina Highway Safety Research Center. Washington, DC: Federal Highway Administration Office of Safety.

Tegge, R. A., Jo, J.-H., Ouyang, Y. (2010). *Development and Application of Safety Performance Functions for Illinois.* Urbana, IL: Illinois Center for Transportation.

Tillander, T. (2018). *HSM2 Development Update, AASHTO Committee on Traffic Engineering.* Annual AASHTO Committee on Traffic Engineering Meeting in Denver, Colorado.

Transport Canada, Government of Canada. (2018). *Canadian Motor Vehicle Traffic Collision Statistics: 2016.* https://www.tc.gc.ca/eng/motorvehiclesafety/canadian-motor-vehicle-traffic-collision-statistics-2016.html.

Transportation Research Board. (2018). *The National Academies of Sciences, Engineering, and Medicine.* Retrieved from NCHRP 17-72 [Active] Update of Crash Modification Factors for the Highway Safety Manual: http://apps.trb.org/cmsfeed/TRBNetProjectDisplay.asp?ProjectID=3875

UCLA: Statistical Consulting Group. (n.d.). *Negative Binomial Regression | SAS Annotated Output*. Retrieved June 24, 2020, from https://stats.idre.ucla.edu/sas/output/negative-binomial-regression/

UCLA: Statistical Consulting Group. (n.d.). *Negative Binomial Regression | SAS Data Analysis Examples*. Retrieved August 19, 2020, from https://stats.idre.ucla.edu/sas/dae/negative-binomial-regression/

UCLA: Statistical Consulting Group. (n.d.). *Negative Binomial Regression | Stata Data Analysis Examples.* Retrieved June 24, 2020, from https://stats.idre.ucla.edu/stata/dae/negative-binomial-regression/

Wali, B., Khattak, A. J., Waters, J., Chimba, D., & Li, X. (2018). *Development of Safety Performance Functions: Incorporating Unobserved Heterogeneity and Functional Form Analysis.* Transportation Research Record: Journal of the Transportation Research Board. https://doi.org/10.1177/0361198118767409.

Washington, S., Karlaftis, M., and Mannering, F. (2011). *Statistical and Econometric Methods for Transportation Data Analysis Second Edition.* Boca Raton, FL: CRC Press Group.

Young, J. and Park, P. Y. (2013). *Benefits of small municipalities using jurisdiction-specific safety performance functions rather than the Highway Safety Manual's calibrated or uncalibrated safety performance functions.* Canadian Journal of Civil Engineering. doi:https://doi.org/10.1139/cjce-2012-0501

A. Appendix A: Selected Control Sections

Table A.1: List of Control Sections Selected for Model Development

Route	Control section	Road classification	Directional divider	County
1	9	Arterial	Divided	Saint John
1	14	Arterial	Divided	Kings
1	81	Arterial	Divided	Westmorland
2	16	Arterial	Divided	Carleton
2	42	Arterial	Divided	Westmorland
2	62	Arterial	Divided	Westmorland
4	1	Arterial	Undivided	York
7	2	Arterial	Divided	Sunbury
7	4	Arterial	Undivided	Sunbury
7	9	Arterial	Divided	Saint John
7	95	Arterial	Divided	Queens
8	6	Arterial	Undivided	York
8	20	Arterial	Undivided	Gloucester
11	3	Arterial	Undivided	Kent
11	5	Arterial	Undivided	Kent
11	25	Arterial	Undivided	Gloucester
11	29	Arterial	Undivided	Restigouche
11	31	Arterial	Undivided	Restigouche
11	34	Arterial	Undivided	Restigouche
15	97	Arterial	Divided	Westmorland
17	4	Arterial	Undivided	Restigouche
17	8	Arterial	Undivided	Victoria
17	11	Arterial	Undivided	Madawaska
95	1	Arterial	Divided	Carleton
101	7	Collector	Undivided	Charlotte
102	11	Collector	Undivided	Sunbury
105	11	Collector	Undivided	York
107	2	Collector	Undivided	York
107	4	Collector	Undivided	York
108	6	Collector	Undivided	Victoria
108	8	Collector	Undivided	Victoria
108	12	Collector	Undivided	Northumberland
108	14	Collector	Undivided	Northumberland
109	2	Collector	Undivided	Victoria
112	9	Collector	Undivided	Queens
114	2	Collector	Undivided	Kings

116	3	Collector	Undivided	Kent
116	5	Collector	Undivided	Kent
117	2	Collector	Undivided	Kent
118	2	Collector	Undivided	Northumberland
121	3	Collector	Undivided	Kings
122	3	Collector	Undivided	York
126	2	Collector	Undivided	Northumberland
126	5	Collector	Undivided	Kent
134	22	Collector	Undivided	Gloucester
160	3	Collector	Undivided	Gloucester
180	2	Collector	Undivided	Gloucester
180	4	Collector	Undivided	Restigouche
180	6	Collector	Undivided	Restigouche
180	9	Collector	Undivided	Restigouche

Table A.2: List of Control Sections Selected for the Case Study

Route	Control section	Road classification	Directional divider	County
2	60	Arterial	Divided	Westmorland
11	27	Arterial	Undivided	Gloucester
105	14	Collector	Undivided	York
108	10	Collector	Undivided	Northumberland
123	2	Collector	Undivided	Sunbury

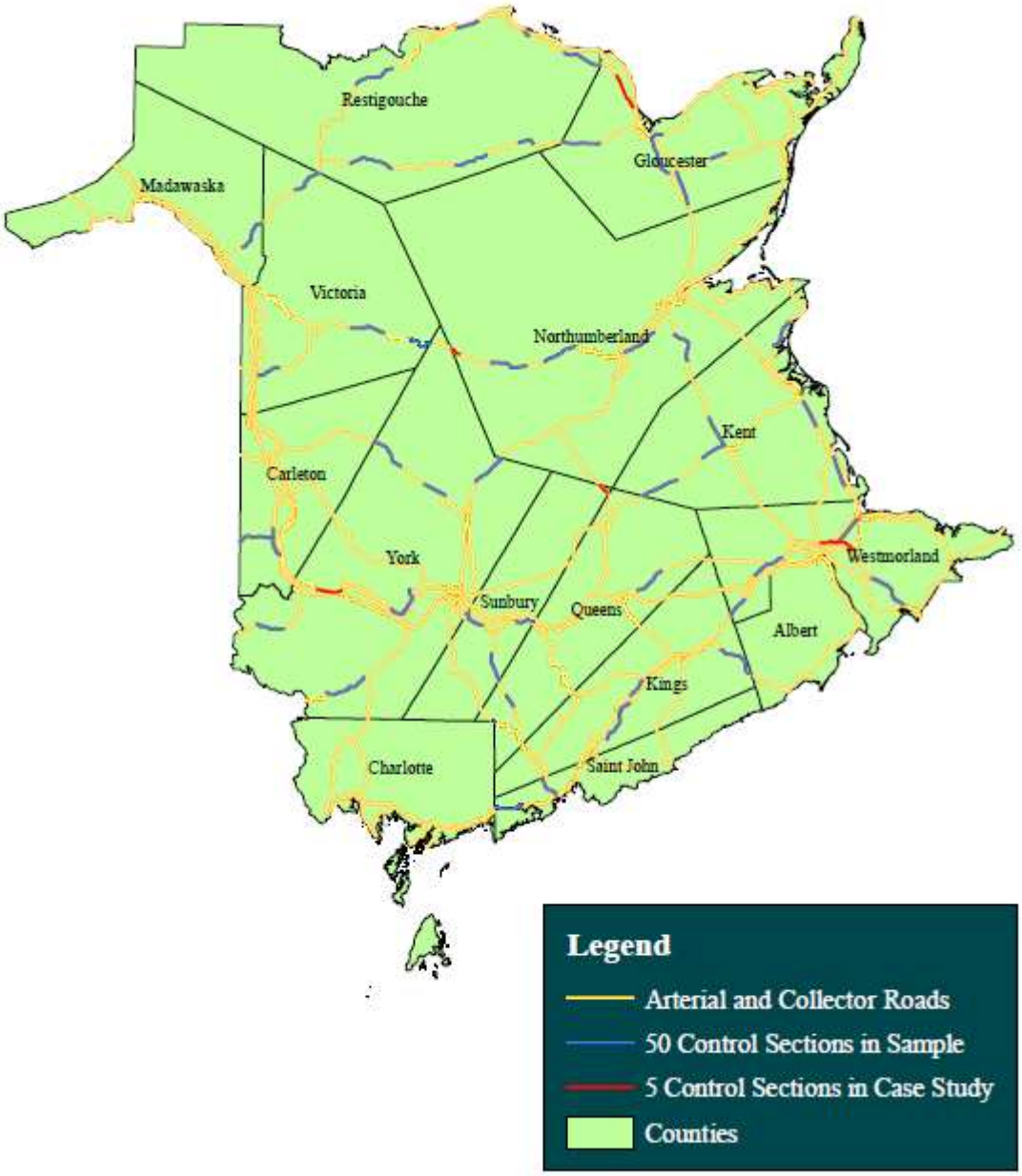

Figure A.1: County Map

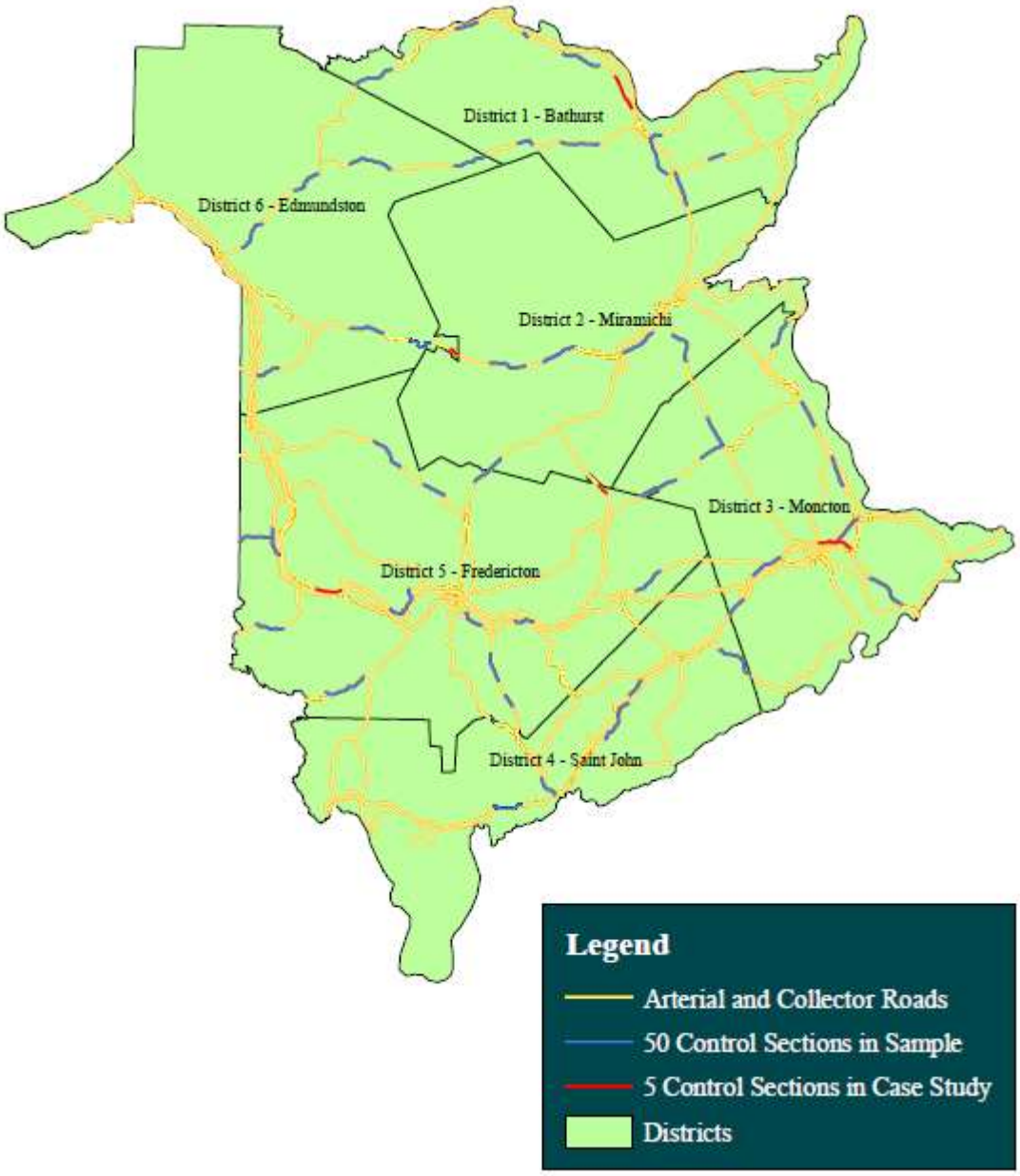

Figure A.2: District Map

B. Appendix B: Variable Data

Table B.1: Variable Data for Divided Segments

Year	Route	Control Section	Segment Length (km)	Traffic volume (vehicles per day)	Surface type	Surface age (years)	Shoulder width (m)	Lane width (m)	Horizontal Curvature	Grade	Posted speed (km/h)	Plant Hardiness Zone	Driveway density (per km)	Wildlife fencing (percentage fenced)	Actual collisions
2011	1	9	11.901	7130	Asphalt	11	2.0	3.75	0.57	1.60	110	5b	0	35.71	8
2012	1	9	11.901	6820	Asphalt	12	2.0	3.75	0.57	1.60	110	5b	0	35.71	4
2014	1	9	11.901	6310	Asphalt	14	2.0	3.75	0.57	1.60	110	5b	0	35.71	4
2015	1	9	11.901	6330	Asphalt	15	2.0	3.75	0.57	1.60	110	5b	0	35.71	2
2016	1	9	11.901	6840	Asphalt	16	2.0	3.75	0.57	1.60	110	5b	0	35.71	7
2011	1	14	18.904	9200	Asphalt	1	2.5	3.75	0.64	1.59	110	5b	0	0	6
2012	1	14	18.904	9110	Asphalt	2	2.5	3.75	0.64	1.59	110	5b	0	0	9
2014	1	14	18.904	10500	Asphalt	4	2.5	3.75	0.64	1.59	110	5b	0	0	6
2015	1	14	18.904	11300	Asphalt	5	2.5	3.75	0.64	1.59	110	5b	0	0	8
2016	1	14	18.904	11400	Asphalt	6	2.5	3.75	0.64	1.59	110	5b	0	0	13
2011	1	81	8.669	7500	Asphalt	7	2.4	3.75	0.13	1.07	110	5b	0	0	2
2012	1	81	8.669	7500	Asphalt	8	2.4	3.75	0.13	1.07	110	5b	0	0	1
2014	1	81	8.669	7500	Asphalt	10	2.4	3.75	0.13	1.07	110	5b	0	0	4
2015	1	81	8.669	7520	Asphalt	11	2.4	3.75	0.13	1.07	110	5b	0	0	3
2016	1	81	8.669	7800	Asphalt	12	2.4	3.75	0.13	1.07	110	5b	0	0	1
2011	2	16	11.7919	10495	Asphalt	2	2.4	3.75	0.20	1.94	100	4b	0	0	11
2012	2	16	11.7919	10495	Asphalt	2	2.4	3.75	0.20	1.94	100	4b	0	0	7
2014	2	16	11.7919	10275	Asphalt	4	2.4	3.75	0.20	1.94	100	4b	0	0	13

Year	Route	Control Section	Segment Length (km)	Traffic volume (vehicles per day)	Surface type	Surface age (years)	Shoulder width (m)	Lane width (m)	Horizontal Curvature	Grade	Posted speed (km/h)	Plant Hardiness Zone	Driveway density (per km)	Wildlife fencing (percentage fenced)	Actual collisions
2015	2	16	11.7919	10500	Asphalt	5	2.4	3.75	0.20	1.94	100	4b	0	0	12
2016	2	16	11.7919	10725	Asphalt	6	2.4	3.75	0.20	1.94	100	4b	0	0	7
2011	2	42	17.300	12700	Asphalt	7	2.4	3.75	0.26	1.92	110	5b	0	0	13
2012	2	42	17.300	12800	Asphalt	8	2.4	3.75	0.26	1.92	110	5b	0	0	14
2014	2	42	17.300	12900	Asphalt	5	2.4	3.75	0.26	1.92	110	5b	0	0	13
2015	2	42	17.300	13000	Asphalt	1	2.4	3.75	0.26	1.92	110	5b	0	0	6
2016	2	42	17.300	13500	Asphalt	2	2.4	3.75	0.26	1.92	110	5b	0	0	9
2011	2	62	13.343	15800	Asphalt	5	2.5	3.75	0.32	1.18	110	5a	0	0	12
2012	2	62	13.343	15600	Asphalt	6	2.5	3.75	0.32	1.18	110	5a	0	0	3
2014	2	62	13.343	15500	Asphalt	8	2.5	3.75	0.32	1.18	110	5a	0	0	10
2015	2	62	13.343	16100	Asphalt	9	2.5	3.75	0.32	1.18	110	5a	0	0	18
2016	2	62	13.343	16900	Asphalt	10	2.5	3.75	0.32	1.18	110	5a	0	0	5
2011	7	2	8.4421	20600	Asphalt	1	2.7	3.75	0.00	1.02	110	5b	0	0	0
2012	7	2	8.4421	20500	Asphalt	2	2.7	3.75	0.00	1.02	110	5b	0	0	4
2014	7	2	8.4421	20600	Asphalt	4	2.7	3.75	0.00	1.02	110	5b	0	0	0
2015	7	2	8.4421	21300	Asphalt	5	2.7	3.75	0.00	1.02	110	5b	0	0	0
2016	7	2	8.4421	22100	Asphalt	6	2.7	3.75	0.00	1.02	110	5b	0	0	1
2011	7	9	9.629	6120	Asphalt	14	2.3	3.75	0.55	2.14	100	5b	0	0	5
2012	7	9	9.629	6120	Asphalt	15	2.3	3.75	0.55	2.14	100	5b	0	0	0
2014	7	9	9.629	6050	Asphalt	17	2.3	3.75	0.55	2.14	100	5b	0	0	0
2015	7	9	9.629	6600	Asphalt	7	2.3	3.75	0.55	2.14	100	5b	0	0	0
2016	7	9	9.629	6800	Asphalt	8	2.3	3.75	0.55	2.14	100	5b	0	0	2

Year	Route	Control Section	Segment Length (km)	Traffic volume (vehicles per day)	Surface type	Surface age (years)	Shoulder width (m)	Lane width (m)	Horizontal Curvature	Grade	Posted speed (km/h)	Plant Hardiness Zone	Driveway density (per km)	Wildlife fencing (percentage fenced)	Actual collisions
2011	7	95	5.06	4420	Asphalt	1	2.5	3.75	0.14	3.29	100	5b	0	100	3
2012	7	95	5.06	3840	Asphalt	2	2.5	3.75	0.14	3.29	100	5b	0	100	1
2014	7	95	5.06	3880	Asphalt	4	2.5	3.75	0.14	3.29	100	5b	0	100	2
2015	7	95	5.06	3950	Asphalt	5	2.5	3.75	0.14	3.29	100	5b	0	100	0
2016	7	95	5.06	4160	Asphalt	6	2.5	3.75	0.14	3.29	100	5b	0	100	3
2011	15	97	12.900	22600	Asphalt	4	2.5	3.75	0.08	1.22	110	5b	0.08	0	4
2012	15	97	12.900	22800	Asphalt	5	2.5	3.75	0.08	1.22	110	5b	0.08	0	8
2014	15	97	12.900	23300	Asphalt	7	2.5	3.75	0.08	1.22	110	5b	0.08	0	7
2015	15	97	12.900	23800	Asphalt	7	2.5	3.75	0.08	1.22	110	5b	0.08	0	6
2016	15	97	12.900	24800	Asphalt	4	2.5	3.75	0.08	1.22	110	5b	0.08	0	12
2011	95	1	14.441	2180	Asphalt	3	3.0	4.86	0.52	1.59	100	4b	0.28	0	3
2012	95	1	14.441	2210	Asphalt	4	3.0	4.86	0.52	1.59	100	4b	0.28	0	0
2014	95	1	14.441	2030	Asphalt	6	3.0	4.86	0.52	1.59	100	4b	0.28	0	0
2015	95	1	14.441	1680	Asphalt	7	3.0	4.86	0.52	1.59	100	4b	0.28	0	4
2016	95	1	14.441	1600	Asphalt	8	3.0	4.86	0.52	1.59	100	4b	0.28	0	2

Table B.2: Variable Data for Undivided Segments

Year	Route	Control Section	Segment Length (km)	Traffic volume (vehicles per day)	Surface type	Surface age (years)	Shoulder width (m)	Lane width (m)	Horizontal Curvature	Grade	Posted speed (km/h)	Plant Hardiness Zone	Driveway density (per km)	Wildlife fencing (percentage fenced)	Actual collisions
2011	4	1	18.177	910	Asphalt	16	2.5	3.60	0.87	1.76	90	4a	5.58	0	6
2012	4	1	18.177	910	Asphalt	17	2.5	3.60	0.87	1.76	90	4a	5.58	0	3
2014	4	1	18.177	860	Asphalt	19	2.5	3.60	0.87	1.76	90	4a	5.58	0	3
2015	4	1	18.177	860	Asphalt	20	2.5	3.60	0.87	1.76	90	4a	5.58	0	3
2016	4	1	18.177	860	Asphalt	18	2.5	3.60	0.87	1.76	90	4a	5.58	0	3
2011	7	4	11.849	4050	Asphalt	10	3.0	3.75	0.38	2.85	100	5b	0.25	100	4
2012	7	4	11.849	4090	Asphalt	11	3.0	3.75	0.38	2.85	100	5b	0.25	100	6
2014	7	4	11.849	4190	Asphalt	13	3.0	3.75	0.38	2.85	100	5b	0.25	100	2
2015	7	4	11.849	4300	Asphalt	14	3.0	3.75	0.38	2.85	100	5b	0.25	100	0
2016	7	4	11.849	4540	Asphalt	15	3.0	3.75	0.38	2.85	100	5b	0.25	100	5
2011	8	6	14.216	2250	Asphalt	7	2.5	3.75	0.29	3.22	90	4b	8.23	0	5
2012	8	6	14.216	2300	Asphalt	8	2.5	3.75	0.29	3.22	90	4b	8.23	0	4
2014	8	6	14.216	2200	Asphalt	10	2.5	3.75	0.29	3.22	90	4b	8.23	0	4
2015	8	6	14.216	2150	Asphalt	11	2.5	3.75	0.29	3.22	90	4b	8.23	0	8
2016	8	6	14.216	2390	Asphalt	8	2.5	3.75	0.29	3.22	90	4b	8.23	0	12
2011	8	20	16.334	4010	Asphalt	1	3	3.75	0.21	1.36	100	4b	0	84.23	1
2012	8	20	16.334	3930	Asphalt	2	3	3.75	0.21	1.36	100	4b	0	84.23	9
2014	8	20	16.334	4030	Asphalt	4	3	3.75	0.21	1.36	100	4b	0	84.23	6
2015	8	20	16.334	4190	Asphalt	5	3	3.75	0.21	1.36	100	4b	0	84.23	5
2016	8	20	16.334	4280	Asphalt	6	3	3.75	0.21	1.36	100	4b	0	84.23	4

Year	Route	Control Section	Segment Length (km)	Traffic volume (vehicles per day)	Surface type	Surface age (years)	Shoulder width (m)	Lane width (m)	Horizontal Curvature	Grade	Posted speed (km/h)	Plant Hardiness Zone	Driveway density (per km)	Wildlife fencing (percentage fenced)	Actual collisions
2011	11	3	12.481	7030	Asphalt	4	3	3.75	0.08	1.31	100	5a	0.80	0	6
2012	11	3	12.481	6970	Asphalt	5	3	3.75	0.08	1.31	100	5a	0.80	0	4
2014	11	3	12.481	7320	Asphalt	3	3	3.75	0.08	1.31	100	5a	0.80	0	16
2015	11	3	12.481	7520	Asphalt	4	3	3.75	0.08	1.31	100	5a	0.80	0	8
2016	11	3	12.481	7600	Asphalt	5	3	3.75	0.08	1.31	100	5a	0.80	0	11
2011	11	5	11.013	5990	Asphalt	3	3	3.75	0.07	0.97	100	5a	0.45	0	3
2012	11	5	11.013	5940	Asphalt	4	3	3.75	0.07	0.97	100	5a	0.45	0	11
2014	11	5	11.013	6100	Asphalt	6	3	3.75	0.07	0.97	100	5a	0.45	0	10
2015	11	5	11.013	6250	Asphalt	7	3	3.75	0.07	0.97	100	5a	0.45	0	11
2016	11	5	11.013	6350	Asphalt	8	3	3.75	0.07	0.97	100	5a	0.45	0	9
2011	11	25	4.726	1940	Asphalt	6	3	3.75	0.26	2.53	100	4b	0	0	0
2012	11	25	4.726	1950	Asphalt	7	3	3.75	0.26	2.53	100	4b	0	0	3
2014	11	25	4.726	1910	Asphalt	9	3	3.75	0.26	2.53	100	4b	0	0	1
2015	11	25	4.726	2060	Asphalt	10	3	3.75	0.26	2.53	100	4b	0	0	1
2016	11	25	4.726	2120	Asphalt	11	3	3.75	0.26	2.53	100	4b	0	0	2
2011	11	29	15.799	3150	Asphalt	7	3.0	4.92	0.18	1.53	100	4b	0.13	18.04	7
2012	11	29	15.799	3320	Asphalt	8	3.0	4.92	0.18	1.53	100	4b	0.13	18.04	7
2014	11	29	15.799	3320	Asphalt	10	3.0	4.92	0.18	1.53	100	4b	0.13	18.04	9
2015	11	29	15.799	3200	Asphalt	11	3.0	4.92	0.18	1.53	100	4b	0.13	18.04	7
2016	11	29	15.799	3260	Asphalt	8	3.0	4.92	0.18	1.53	100	4b	0.13	18.04	4
2011	11	31	2.518	3420	Asphalt	2	3	3.75	0.13	1.00	100	4b	0.79	0	2
2012	11	31	2.518	3450	Asphalt	3	3	3.75	0.13	1.00	100	4b	0.79	0	2

Year	Route	Control Section	Segment Length (km)	Traffic volume (vehicles per day)	Surface type	Surface age (years)	Shoulder width (m)	Lane width (m)	Horizontal Curvature	Grade	Posted speed (km/h)	Plant Hardiness Zone	Driveway density (per km)	Wildlife fencing (percentage fenced)	Actual collisions
2014	11	31	2.518	3450	Asphalt	5	3	3.75	0.13	1.00	100	4b	0.79	0	3
2015	11	31	2.518	3640	Asphalt	6	3	3.75	0.13	1.00	100	4b	0.79	0	4
2016	11	31	2.518	3710	Asphalt	7	3	3.75	0.13	1.00	100	4b	0.79	0	0
2011	11	34	9.009	5700	Asphalt	7	3	4.37	0.11	2.20	100	4a	0.78	0	0
2012	11	34	9.009	5820	Asphalt	4	3	4.37	0.11	2.20	100	4a	0.78	0	3
2014	11	34	9.009	5820	Asphalt	4	3	4.37	0.11	2.20	100	4a	0.78	0	7
2015	11	34	9.009	5940	Asphalt	5	3	4.37	0.11	2.20	100	4a	0.78	0	10
2016	11	34	9.009	6000	Asphalt	6	3	4.37	0.11	2.20	100	4a	0.78	0	3
2011	17	4	16.869	1220	Asphalt	14	3	4.32	0.32	3.45	90	3b	5.75	0	5
2012	17	4	16.869	1160	Asphalt	15	3	4.32	0.32	3.45	90	3b	5.75	0	2
2014	17	4	16.869	1160	Asphalt	17	3	4.32	0.32	3.45	90	3b	5.75	0	4
2015	17	4	16.869	1220	Asphalt	10	3	4.32	0.32	3.45	90	3b	5.75	0	8
2016	17	4	16.869	1220	Asphalt	7	3	4.32	0.32	3.45	90	3b	5.75	0	3
2011	17	8	12.783	1690	Asphalt	11	3	3.50	0.55	3.20	90	3b	0.08	0	7
2012	17	8	12.783	1660	Asphalt	12	3	3.50	0.55	3.20	90	3b	0.08	0	0
2014	17	8	12.783	1670	Asphalt	14	3	3.50	0.55	3.20	90	3b	0.08	0	6
2015	17	8	12.783	1760	Asphalt	14	3	3.50	0.55	3.20	90	3b	0.08	0	7
2016	17	8	12.783	1880	Asphalt	15	3	3.50	0.55	3.20	90	3b	0.08	0	4
2011	17	11	13.525	2020	Asphalt	9	3	3.64	0.34	2.42	90	4a	2.14	0	5
2012	17	11	13.525	1890	Asphalt	10	3	3.64	0.34	2.42	90	4a	2.14	0	8
2014	17	11	13.525	1890	Asphalt	12	3	3.64	0.34	2.42	90	4a	2.14	0	1
2015	17	11	13.525	2000	Asphalt	8	3	3.64	0.34	2.42	90	4a	2.14	0	7

Year	Route	Control Section	Segment Length (km)	Traffic volume (vehicles per day)	Surface type	Surface age (years)	Shoulder width (m)	Lane width (m)	Horizontal Curvature	Grade	Posted speed (km/h)	Plant Hardiness Zone	Driveway density (per km)	Wildlife fencing (percentage fenced)	Actual collisions
2016	17	11	13.525	2050	Asphalt	9	3	3.64	0.34	2.42	90	4a	2.14	0	9
2011	101	7	0.715	1020	Asphalt	12	2.85	3.50	0.50	0.50	80	5b	2.80	0	0
2012	101	7	0.715	1020	Asphalt	13	2.85	3.50	0.50	0.50	80	5b	2.80	0	0
2014	101	7	0.715	750	Asphalt	15	2.85	3.50	0.50	0.50	80	5b	2.80	0	0
2015	101	7	0.715	750	Asphalt	16	2.85	3.50	0.50	0.50	80	5b	2.80	0	0
2016	101	7	0.715	800	Asphalt	17	2.85	3.50	0.50	0.50	80	5b	2.80	0	0
2011	102	11	8.06	640	Asphalt	19	1.2	3.40	1.81	1.46	70	5b	13.65	0	1
2012	102	11	8.06	660	Asphalt	13	1.2	3.40	1.81	1.46	70	5b	13.65	0	0
2014	102	11	8.06	660	Asphalt	5	1.2	3.40	1.81	1.46	70	5b	13.65	0	0
2015	102	11	8.06	680	Asphalt	6	1.2	3.40	1.81	1.46	70	5b	13.65	0	0
2016	102	11	8.06	680	Asphalt	7	1.2	3.40	1.81	1.46	70	5b	13.65	0	2
2011	105	11	17.511	1010	Chip seal	14	1.5	3.47	1.50	5.11	80	5a	11.88	0	3
2012	105	11	17.511	910	Chip seal	15	1.5	3.47	1.50	5.11	80	5a	11.88	0	5
2014	105	11	17.511	890	Chip seal	17	1.5	3.47	1.50	5.11	80	5a	11.88	0	5
2015	105	11	17.511	880	Chip seal	2	1.5	3.47	1.50	5.11	80	5a	11.88	0	4
2016	105	11	17.511	880	Chip seal	1	1.5	3.47	1.50	5.11	80	5a	11.88	0	3
2011	107	2	9.65	520	Chip seal	10	1.9	3.46	0.04	3.80	80	4b	14.09	0	0
2012	107	2	9.65	520	Chip seal	11	1.9	3.46	0.04	3.80	80	4b	14.09	0	1
2014	107	2	9.65	470	Chip seal	13	1.9	3.46	0.04	3.80	80	4b	14.09	0	2
2015	107	2	9.65	470	Chip seal	14	1.9	3.46	0.04	3.80	80	4b	14.09	0	1
2016	107	2	9.65	470	Chip seal	15	1.9	3.46	0.04	3.80	80	4b	14.09	0	0
2011	107	4	15.834	870	Chip seal	14	1.7	3.50	0.57	4.19	80	4a	1.52	0	1

Year	Route	Control Section	Segment Length (km)	Traffic volume (vehicles per day)	Surface type	Surface age (years)	Shoulder width (m)	Lane width (m)	Horizontal Curvature	Grade	Posted speed (km/h)	Plant Hardiness Zone	Driveway density (per km)	Wildlife fencing (percentage fenced)	Actual collisions
2012	107	4	15.834	870	Chip seal	15	1.7	3.50	0.57	4.19	80	4a	1.52	0	0
2014	107	4	15.834	780	Chip seal	17	1.7	3.50	0.57	4.19	80	4a	1.52	0	0
2015	107	4	15.834	780	Chip seal	18	1.7	3.50	0.57	4.19	80	4a	1.52	0	0
2016	107	4	15.834	625	Chip seal	17	1.7	3.50	0.57	4.19	80	4a	1.52	0	1
2011	108	6	13.99	500	Chip seal	18	1.6	3.50	0.58	3.56	80	4a	0.79	0	0
2012	108	6	13.99	550	Chip seal	19	1.6	3.50	0.58	3.56	80	4a	0.79	0	1
2014	108	6	13.99	600	Chip seal	21	1.6	3.50	0.58	3.56	80	4a	0.79	0	1
2015	108	6	13.99	730	Chip seal	22	1.6	3.50	0.58	3.56	80	4a	0.79	0	1
2016	108	6	13.99	730	Chip seal	23	1.6	3.50	0.58	3.56	80	4a	0.79	0	2
2011	108	8	9.965	500	Chip seal	4	1.1	3.35	0.64	3.45	80	4a	0.40	0	1
2012	108	8	9.965	550	Chip seal	5	1.1	3.35	0.64	3.45	80	4a	0.40	0	1
2014	108	8	9.965	600	Chip seal	7	1.1	3.35	0.64	3.45	80	4a	0.40	0	1
2015	108	8	9.965	730	Chip seal	8	1.1	3.35	0.64	3.45	80	4a	0.40	0	0
2016	108	8	9.965	730	Chip seal	9	1.1	3.35	0.64	3.45	80	4a	0.40	0	0
2011	108	12	15.143	490	Chip seal	5	1.2	3.47	0.54	2.62	80	4a	1.32	0	2
2012	108	12	15.143	430	Chip seal	6	1.2	3.47	0.54	2.62	80	4a	1.32	0	2
2014	108	12	15.143	430	Chip seal	8	1.2	3.47	0.54	2.62	80	4a	1.32	0	0
2015	108	12	15.143	680	Chip seal	9	1.2	3.47	0.54	2.62	80	4a	1.32	0	1
2016	108	12	15.143	680	Chip seal	10	1.2	3.47	0.54	2.62	80	4a	1.32	0	0
2011	108	14	14.148	490	Chip seal	8	1.5	3.50	0.45	1.58	80	4b	2.47	0	1
2012	108	14	14.148	430	Chip seal	9	1.5	3.50	0.45	1.58	80	4b	2.47	0	2
2014	108	14	14.148	430	Chip seal	11	1.5	3.50	0.45	1.58	80	4b	2.47	0	3

Year	Route	Control Section	Segment Length (km)	Traffic volume (vehicles per day)	Surface type	Surface age (years)	Shoulder width (m)	Lane width (m)	Horizontal Curvature	Grade	Posted speed (km/h)	Plant Hardiness Zone	Driveway density (per km)	Wildlife fencing (percentage fenced)	Actual collisions
2015	108	14	14.148	680	Chip seal	12	1.5	3.50	0.45	1.58	80	4b	2.47	0	0
2016	108	14	14.148	680	Chip seal	2	1.5	3.50	0.45	1.58	80	4b	2.47	0	3
2011	109	2	9.626	1450	Asphalt	16	1.1	3.50	2.57	2.27	70	4b	12.15	0	3
2012	109	2	9.626	1440	Asphalt	17	1.1	3.50	2.57	2.27	70	4b	12.15	0	1
2014	109	2	9.626	1400	Asphalt	19	1.1	3.50	2.57	2.27	70	4b	12.15	0	2
2015	109	2	9.626	1400	Asphalt	20	1.1	3.50	2.57	2.27	70	4b	12.15	0	6
2016	109	2	9.626	1450	Asphalt	7	1.1	3.50	2.57	2.27	70	4b	12.15	0	5
2011	112	9	13.28	320	Asphalt	17	1.9	3.38	0.60	1.74	90	5a	4.97	0	1
2012	112	9	13.28	310	Asphalt	18	1.9	3.38	0.60	1.74	90	5a	4.97	0	0
2014	112	9	13.28	300	Asphalt	20	1.9	3.38	0.60	1.74	90	5a	4.97	0	0
2015	112	9	13.28	230	Asphalt	21	1.9	3.38	0.60	1.74	90	5a	4.97	0	1
2016	112	9	13.28	230	Asphalt	14	1.9	3.38	0.60	1.74	90	5a	4.97	0	1
2011	114	2	18.218	420	Chip seal	11	1.6	3.36	1.36	4.02	80	5a	4.28	0	3
2012	114	2	18.218	420	Chip seal	12	1.6	3.36	1.36	4.02	80	5a	4.28	0	0
2014	114	2	18.218	390	Chip seal	6	1.6	3.36	1.36	4.02	80	5a	4.28	0	3
2015	114	2	18.218	390	Chip seal	7	1.6	3.36	1.36	4.02	80	5a	4.28	0	1
2016	114	2	18.218	410	Chip seal	6	1.6	3.36	1.36	4.02	80	5a	4.28	0	3
2011	116	3	15.258	400	Chip seal	5	1.7	3.50	0.47	1.03	90	5a	3.80	0	0
2012	116	3	15.258	600	Chip seal	6	1.7	3.50	0.47	1.03	90	5a	3.80	0	1
2014	116	3	15.258	580	Chip seal	8	1.7	3.50	0.47	1.03	90	5a	3.80	0	0
2015	116	3	15.258	330	Chip seal	3	1.7	3.50	0.47	1.03	90	5a	3.80	0	0
2016	116	3	15.258	330	Chip seal	4	1.7	3.50	0.47	1.03	90	5a	3.80	0	1

Year	Route	Control Section	Segment Length (km)	Traffic volume (vehicles per day)	Surface type	Surface age (years)	Shoulder width (m)	Lane width (m)	Horizontal Curvature	Grade	Posted speed (km/h)	Plant Hardiness Zone	Driveway density (per km)	Wildlife fencing (percentage fenced)	Actual collisions
2011	116	5	10.19	350	Chip seal	6	1.6	3.50	0.11	0.70	90	4b	4.71	0	4
2012	116	5	10.19	350	Chip seal	7	1.6	3.50	0.11	0.70	90	4b	4.71	0	2
2014	116	5	10.19	330	Chip seal	9	1.6	3.50	0.11	0.70	90	4b	4.71	0	2
2015	116	5	10.19	290	Chip seal	10	1.6	3.50	0.11	0.70	90	4b	4.71	0	1
2016	116	5	10.19	290	Chip seal	0	1.6	3.50	0.11	0.70	90	4b	4.71	0	1
2011	117	2	12.689	740	Asphalt	27	1.3	3.50	1.35	0.63	80	5a	1.18	0	1
2012	117	2	12.689	740	Asphalt	28	1.3	3.50	1.35	0.63	80	5a	1.18	0	1
2014	117	2	12.689	840	Asphalt	30	1.3	3.50	1.35	0.63	80	5a	1.18	0	1
2015	117	2	12.689	840	Asphalt	31	1.3	3.50	1.35	0.63	80	5a	1.18	0	2
2016	117	2	12.689	900	Asphalt	32	1.3	3.50	1.35	0.63	80	5a	1.18	0	1
2011	118	2	14.593	900	Chip seal	6	1.9	3.50	0.37	1.64	80	4b	10.28	0	0
2012	118	2	14.593	900	Chip seal	7	1.9	3.50	0.37	1.64	80	4b	10.28	0	3
2014	118	2	14.593	860	Chip seal	6	1.9	3.50	0.37	1.64	80	4b	10.28	0	3
2015	118	2	14.593	860	Chip seal	7	1.9	3.50	0.37	1.64	80	4b	10.28	0	0
2016	118	2	14.593	970	Chip seal	8	1.9	3.50	0.37	1.64	80	4b	10.28	0	6
2011	121	3	7.753	1390	Asphalt	15	1.5	3.50	1.13	2.21	80	5b	11.74	0	1
2012	121	3	7.753	1390	Asphalt	16	1.5	3.50	1.13	2.21	80	5b	11.74	0	0
2014	121	3	7.753	1250	Asphalt	18	1.5	3.50	1.13	2.21	80	5b	11.74	0	2
2015	121	3	7.753	1250	Asphalt	19	1.5	3.50	1.13	2.21	80	5b	11.74	0	2
2016	121	3	7.753	1080	Asphalt	20	1.5	3.50	1.13	2.21	80	5b	11.74	0	4
2011	122	3	12.197	340	Chip seal	17	1.1	3.22	2.69	3.87	80	5a	1.64	0	0
2012	122	3	12.197	340	Chip seal	9	1.1	3.22	2.69	3.87	80	5a	1.64	0	2

Year	Route	Control Section	Segment Length (km)	Traffic volume (vehicles per day)	Surface type	Surface age (years)	Shoulder width (m)	Lane width (m)	Horizontal Curvature	Grade	Posted speed (km/h)	Plant Hardiness Zone	Driveway density (per km)	Wildlife fencing (percentage fenced)	Actual collisions
2014	122	3	12.197	410	Chip seal	1	1.1	3.22	2.69	3.87	80	5a	1.64	0	0
2015	122	3	12.197	410	Chip seal	2	1.1	3.22	2.69	3.87	80	5a	1.64	0	2
2016	122	3	12.197	400	Chip seal	3	1.1	3.22	2.69	3.87	80	5a	1.64	0	1
2011	126	2	14.547	2230	Asphalt	13	2.0	3.50	1.10	1.58	90	4b	3.99	0	7
2012	126	2	14.547	2270	Asphalt	14	2.0	3.50	1.10	1.58	90	4b	3.99	0	4
2014	126	2	14.547	2270	Asphalt	14	2.0	3.50	1.10	1.58	90	4b	3.99	0	3
2015	126	2	14.547	2280	Asphalt	15	2.0	3.50	1.10	1.58	90	4b	3.99	0	6
2016	126	2	14.547	2280	Asphalt	9	2.0	3.50	1.10	1.58	90	4b	3.99	0	10
2011	126	5	12.693	1880	Asphalt	3	2	3.50	0.22	1.57	90	4b	3.55	0	5
2012	126	5	12.693	1880	Asphalt	4	2	3.50	0.22	1.57	90	4b	3.55	0	7
2014	126	5	12.693	1890	Asphalt	6	2	3.50	0.22	1.57	90	4b	3.55	0	8
2015	126	5	12.693	1960	Asphalt	7	2	3.50	0.22	1.57	90	4b	3.55	0	7
2016	126	5	12.693	2280	Asphalt	8	2	3.50	0.22	1.57	90	4b	3.55	0	6
2011	134	22	14.087	330	Asphalt	9	1.8	3.50	0.15	0.45	90	4b	19.38	0	3
2012	134	22	14.087	330	Asphalt	10	1.8	3.50	0.15	0.45	90	4b	19.38	0	1
2014	134	22	14.087	300	Asphalt	12	1.8	3.50	0.15	0.45	90	4b	19.38	0	3
2015	134	22	14.087	300	Asphalt	6	1.8	3.50	0.15	0.45	90	4b	19.38	0	5
2016	134	22	14.087	380	Asphalt	7	1.8	3.50	0.15	0.45	90	4b	19.38	0	1
2011	160	3	6.622	2890	Asphalt	2	1.8	3.50	0.15	3.62	80	5a	25.07	0	5
2012	160	3	6.622	2690	Asphalt	3	1.8	3.50	0.15	3.62	80	5a	25.07	0	2
2014	160	3	6.622	2690	Asphalt	5	1.8	3.50	0.15	3.62	80	5a	25.07	0	1
2015	160	3	6.622	2690	Asphalt	6	1.8	3.50	0.15	3.62	80	5a	25.07	0	0

Year	Route	Control Section	Segment Length (km)	Traffic volume (vehicles per day)	Surface type	Surface age (years)	Shoulder width (m)	Lane width (m)	Horizontal Curvature	Grade	Posted speed (km/h)	Plant Hardiness Zone	Driveway density (per km)	Wildlife fencing (percentage fenced)	Actual collisions
2016	160	3	6.622	2690	Asphalt	7	1.8	3.50	0.15	3.62	80	5a	25.07	0	2
2011	180	2	15.411	870	Asphalt	7	2.0	3.50	0.31	1.55	80	4a	4.02	0	5
2012	180	2	15.411	720	Asphalt	8	2.0	3.50	0.31	1.55	80	4a	4.02	0	7
2014	180	2	15.411	720	Asphalt	10	2.0	3.50	0.31	1.55	80	4a	4.02	0	2
2015	180	2	15.411	720	Asphalt	11	2.0	3.50	0.31	1.55	80	4a	4.02	0	2
2016	180	2	15.411	740	Asphalt	12	2.0	3.50	0.31	1.55	80	4a	4.02	0	6
2011	180	4	6.646	870	Chip seal	13	2.0	3.50	0.40	3.76	80	4a	1.35	0	1
2012	180	4	6.646	720	Chip seal	14	2.0	3.50	0.40	3.76	80	4a	1.35	0	3
2014	180	4	6.646	720	Chip seal	16	2.0	3.50	0.40	3.76	80	4a	1.35	0	0
2015	180	4	6.646	720	Chip seal	17	2.0	3.50	0.40	3.76	80	4a	1.35	0	1
2016	180	4	6.646	740	Chip seal	0	2.0	3.50	0.40	3.76	80	4a	1.35	0	2
2011	180	6	13.725	500	Chip seal	7	2.0	3.50	0.50	3.10	80	3b	1.31	0	1
2012	180	6	13.725	470	Chip seal	8	2.0	3.50	0.50	3.10	80	3b	1.31	0	2
2014	180	6	13.725	470	Chip seal	10	2.0	3.50	0.50	3.10	80	3b	1.31	0	3
2015	180	6	13.725	470	Chip seal	11	2.0	3.50	0.50	3.10	80	3b	1.31	0	0
2016	180	6	13.725	450	Chip seal	12	2.0	3.50	0.50	3.10	80	3b	1.31	0	0
2011	180	9	18.476	490	Chip seal	7	2	3.50	0.49	3.20	80	3b	2.00	0	2
2012	180	9	18.476	520	Chip seal	7	2	3.50	0.49	3.20	80	3b	2.00	0	4
2014	180	9	18.476	520	Chip seal	1	2	3.50	0.49	3.20	80	3b	2.00	0	2
2015	180	9	18.476	520	Chip seal	2	2	3.50	0.49	3.20	80	3b	2.00	0	1
2016	180	9	18.476	500	Chip seal	3	2	3.50	0.49	3.20	80	3b	2.00	0	3

Table B.3: Variable Data for Case Study Segments

Year	Route	Control Section	Segment Length (km)	Traffic volume (vehicles per day)	Surface type	Surface age (years)	Shoulder width (m)	Lane width (m)	Horizontal Curvature	Grade	Posted speed (km/h)	Plant Hardiness Zone	Driveway density (per km)	Wildlife fencing (percentage fenced)	Actual collisions
2011	2	60	14.522	21300	Asphalt	5	2.7	3.75	0.57	0.91	110	5b	0	0	13
2012	2	60	14.522	21500	Asphalt	6	2.7	3.75	0.57	0.91	110	5b	0	0	9
2014	2	60	14.522	21400	Asphalt	8	2.7	3.75	0.57	0.91	110	5b	0	0	13
2015	2	60	14.522	22500	Asphalt	9	2.7	3.75	0.57	0.91	110	5b	0	0	12
2016	2	60	14.522	23200	Asphalt	10	2.7	3.75	0.57	0.91	110	5b	0	0	14
2011	11	27	14.4519	3860	Asphalt	6	3.0	3.75	0.02	1.33	100	4b	0.69	55.42	8
2012	11	27	14.4519	3840	Asphalt	7	3.0	3.75	0.02	1.33	100	4b	0.69	55.42	3
2014	11	27	14.4519	3760	Asphalt	9	3.0	3.75	0.02	1.33	100	4b	0.69	55.42	7
2015	11	27	14.4519	4010	Asphalt	10	3.0	3.75	0.02	1.33	100	4b	0.69	55.42	4
2016	11	27	14.4519	4150	Asphalt	11	3.0	3.75	0.02	1.33	100	4b	0.69	55.42	7
2011	105	14	10.514	2310	Chip seal	14	2.8	3.47	1.59	3.68	80	5a	11.51	0	2
2012	105	14	10.514	2290	Chip seal	15	2.8	3.47	1.59	3.68	80	5a	11.51	0	3
2014	105	14	10.514	2380	Chip seal	12	2.8	3.47	1.59	3.68	80	5a	11.51	0	0
2015	105	14	10.514	2380	Chip seal	13	2.8	3.47	1.59	3.68	80	5a	11.51	0	1
2016	105	14	10.514	2380	Chip seal	14	2.8	3.47	1.59	3.68	80	5a	11.51	0	2
2011	108	10	4.352	490	Chip seal	0	1.3	3.50	0.40	5.67	80	4a	0.92	0	1
2012	108	10	4.352	430	Chip seal	1	1.3	3.50	0.40	5.67	80	4a	0.92	0	0
2014	108	10	4.352	430	Chip seal	3	1.3	3.50	0.40	5.67	80	4a	0.92	0	0
2015	108	10	4.352	680	Chip seal	4	1.3	3.50	0.40	5.67	80	4a	0.92	0	0
2016	108	10	4.352	680	Chip seal	5	1.3	3.50	0.40	5.67	80	4a	0.92	0	0

Year	Route	Control Section	Segment Length (km)	Traffic volume (vehicles per day)	Surface type	Surface age (years)	Shoulder width (m)	Lane width (m)	Horizontal Curvature	Grade	Posted speed (km/h)	Plant Hardiness Zone	Driveway density (per km)	Wildlife fencing (percentage fenced)	Actual collisions
2011	123	2	4.889	350	Asphalt	13	2.0	3.50	0.28	2.75	90	4b	1.02	0	0
2012	123	2	4.889	350	Asphalt	14	2.0	3.50	0.28	2.75	90	4b	1.02	0	2
2014	123	2	4.889	510	Asphalt	16	2.0	3.50	0.28	2.75	90	4b	1.02	0	0
2015	123	2	4.889	510	Asphalt	17	2.0	3.50	0.28	2.75	90	4b	1.02	0	1
2016	123	2	4.889	500	Asphalt	18	2.0	3.50	0.28	2.75	90	4b	1.02	0	1

C. Appendix C: Correlational Study

Table C.1: Correlation Matrix of Potential Explanatory Variables

	Road classification	Directional divider	Segment Length	Traffic volume	Surface type	Surface age	Shoulder width	Lane width	Horizontal Curvature	Grade	Posted speed	Plant Hardiness Zone	Driveway density	Wildlife fencing
Road classification	1.000	0.553	-0.031	0.586	0.629	-0.085	0.830	0.630	-0.425	-0.224	0.841	0.188	-0.481	0.306
Directional divider	0.553	1.000	-0.025	0.747	0.348	-0.134	0.249	0.311	-0.226	-0.240	0.721	0.515	-0.365	0.129
Segment Length	-0.031	-0.025	1.000	-0.028	-0.238	-0.104	-0.116	0.079	0.098	0.158	0.000	-0.289	-0.078	-0.076
Traffic volume	0.586	0.747	-0.028	1.000	0.403	-0.199	0.382	0.243	-0.326	-0.330	0.752	0.469	-0.339	0.035
Surface type	0.629	0.348	-0.238	0.403	1.000	0.188	0.582	0.417	-0.179	-0.500	0.541	0.355	-0.039	0.193
Surface age	-0.085	-0.134	-0.104	-0.199	0.188	1.000	-0.058	-0.096	0.081	-0.075	-0.156	-0.015	-0.034	0.030
Shoulder width	0.830	0.249	-0.116	0.382	0.582	-0.058	1.000	0.644	-0.600	-0.245	0.680	-0.001	-0.408	0.269
Lane width	0.630	0.311	0.079	0.243	0.417	-0.096	0.644	1.000	-0.354	-0.191	0.519	-0.042	-0.285	0.137

	Road classification	Directional divider	Segment Length	Traffic volume	Surface type	Surface age	Shoulder width	Lane width	Horizontal Curvature	Grade	Posted speed	Plant Hardiness Zone	Driveway density	Wildlife fencing
Horizontal Curvature	-0.425	-0.226	0.098	-0.326	-0.179	0.081	-0.600	-0.354	1.000	0.265	-0.535	0.096	0.212	-0.143
Grade	-0.224	-0.240	0.158	-0.330	-0.500	-0.075	-0.245	-0.191	0.265	1.000	-0.419	-0.339	0.159	0.057
Posted speed	0.841	0.721	0.000	0.752	0.541	-0.156	0.680	0.519	-0.535	-0.419	1.000	0.402	-0.515	0.263
Plant Hardiness Zone	0.188	0.515	-0.289	0.469	0.355	-0.015	-0.001	-0.042	0.096	-0.339	0.402	1.000	0.031	0.250
Driveway density	-0.481	-0.365	-0.078	-0.339	-0.039	-0.034	-0.408	-0.285	0.212	0.159	-0.515	0.031	1.000	-0.200
Wildlife fencing	0.306	0.129	-0.076	0.035	0.193	0.030	0.269	0.137	-0.143	0.057	0.263	0.250	-0.200	1.000

D. Appendix D: SAS Generalized Linear Models Results

Table D.1: Generalized Linear Model Results for Divided Segments

Model Information		
Data Set	DATA.DIVIDED_SEGMENTS	
Distribution	Negative Binomial	
Link Function	Log	
Dependent Variable	Actual_collisions	Actual_collisions

Number of Observations Read	55
Number of Observations Used	55

Class Level Information		
Class	Levels	Values
Plant_Hardiness_Zone	3	4 5 6

Criteria For Assessing Goodness Of Fit			
Criterion	DF	Value	Value/DF
Deviance	46	74.8373	1.6269
Scaled Deviance	46	74.8373	1.6269
Pearson Chi-Square	46	62.4988	1.3587
Scaled Pearson X2	46	62.4988	1.3587
Log Likelihood		280.4892	
Full Log Likelihood		-122.7774	
AIC (smaller is better)		265.5549	
AICC (smaller is better)		270.5549	
BIC (smaller is better)		285.6282	

Algorithm converged.

Analysis Of Maximum Likelihood Parameter Estimates								
Parameter		DF	Estimate	Standard Error	Wald 90% Confidence Limits		Wald Chi-Square	Pr > ChiSq
Intercept		1	-12.7743	6.6835	-23.7677	-1.7809	3.65	0.0560
Plant_Hardiness_Zone	4	1	1.9980	0.4357	1.2814	2.7146	21.03	<.0001
Plant_Hardiness_Zone	5	1	0.7300	0.2096	0.3853	1.0746	12.13	0.0005
Plant_Hardiness_Zone	6	0	0.0000	0.0000	0.0000	0.0000	.	.
AADT_1000		1	0.0516	0.0191	0.0201	0.0830	7.28	0.0070
Length		1	0.1234	0.0285	0.0766	0.1703	18.76	<.0001
Grade		1	1.0131	0.3131	0.4981	1.5281	10.47	0.0012
Posted_speed		1	0.1342	0.0539	0.0455	0.2229	6.19	0.0129

Table D.1 (Continued): Generalized Linear Model Results for Divided Segments

Analysis Of Maximum Likelihood Parameter Estimates								
Parameter		DF	Estimate	Standard Error	Wald 90% Confidence Limits		Wald Chi-Square	Pr > ChiSq
Lane_width		1	-1.2016	0.4079	-1.8724	-0.5307	8.68	0.0032
Surface_age		1	0.0498	0.0239	0.0106	0.0891	4.36	0.0368
Dispersion		1	0.0391	0.0456	0.0057	0.2669		

Note: The negative binomial dispersion parameter was estimated by maximum likelihood.

LR Statistics For Type 3 Analysis			
Source	DF	Chi-Square	Pr > ChiSq
Plant_Hardiness_Zone	2	24.36	<.0001
AADT_1000	1	6.68	0.0098
Length	1	16.70	<.0001
Grade	1	9.84	0.0017
Posted_speed	1	6.79	0.0092
Lane_width	1	9.09	0.0026
Surface_age	1	4.10	0.0430

Table D.2: Generalized Linear Model Results for Undivided Segments

Model Information		
Data Set	DATA.UNDIVIDED_SEGMENTS	
Distribution	Negative Binomial	
Link Function	Log	
Dependent Variable	Actual_collisions	Actual_collisions

Number of Observations Read	195
Number of Observations Used	195

Class Level Information		
Class	Levels	Values
Plant_Hardiness_Zone	5	2 3 4 5 6
Surface_type___Aspha	2	0 1

Criteria For Assessing Goodness Of Fit			
Criterion	DF	Value	Value/DF
Deviance	184	225.7224	1.2268
Scaled Deviance	184	225.7224	1.2268
Pearson Chi-Square	184	191.2632	1.0395
Scaled Pearson X2	184	191.2632	1.0395
Log Likelihood		214.4732	
Full Log Likelihood		-365.0490	
AIC (smaller is better)		754.0980	
AICC (smaller is better)		755.8123	
BIC (smaller is better)		793.3740	

Algorithm converged.

Analysis Of Maximum Likelihood Parameter Estimates								
Parameter		DF	Estimate	Standard Error	Wald 90% Confidence Limits		Wald Chi-Square	Pr > ChiSq
Intercept		1	-0.7669	0.3379	-1.3227	-0.2111	5.15	0.0232
Plant_Hardiness_Zone	2	1	-0.0057	0.3173	-0.5276	0.5162	0.00	0.9857
Plant_Hardiness_Zone	3	1	-0.0522	0.2851	-0.5212	0.4167	0.03	0.8546
Plant_Hardiness_Zone	4	1	0.3401	0.2572	-0.0829	0.7632	1.75	0.1860
Plant_Hardiness_Zone	5	1	-0.1797	0.3009	-0.6746	0.3151	0.36	0.5503
Plant_Hardiness_Zone	6	0	0.0000	0.0000	0.0000	0.0000	.	.
Surface_type___Aspha	0	1	-0.8741	0.1583	-1.1344	-0.6138	30.51	<.0001

Table D.2 (Continued): Generalized Linear Model Results for Undivided Segments

Analysis Of Maximum Likelihood Parameter Estimates								
Parameter		DF	Estimate	Standard Error	Wald 90% Confidence Limits		Wald Chi-Square	Pr > ChiSq
Surface_type___Aspha	1	0	0.0000	0.0000	0.0000	0.0000	.	.
Grade		1	0.1377	0.0561	0.0454	0.2299	6.02	0.0141
Surface_age		1	-0.0226	0.0110	-0.0407	-0.0045	4.22	0.0399
AADT_1000		1	0.2435	0.0383	0.1804	0.3066	40.34	<.0001
Length		1	0.1160	0.0168	0.0884	0.1436	47.73	<.0001
Wildlife_fencing		1	-0.0086	0.0028	-0.0133	-0.0040	9.24	0.0024
Dispersion		1	0.0858	0.0437	0.0371	0.1982		

Note: The negative binomial dispersion parameter was estimated by maximum likelihood.

LR Statistics For Type 3 Analysis			
Source	DF	Chi-Square	Pr > ChiSq
Plant_Hardiness_Zone	4	14.15	0.0068
Surface_type___Aspha	1	29.08	<.0001
Grade	1	6.01	0.0142
Surface_age	1	4.24	0.0395
AADT_1000	1	37.50	<.0001
Length	1	44.98	<.0001
Wildlife_fencing	1	8.75	0.0031

E. Appendix E: SAS Linear Regression Results

Table E.1: Linear Regression Results for Divided Segments

Number of Observations Read	55
Number of Observations Used	55

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	776.00068	776.00068	111.95	<.0001
Error	53	367.38114	6.93172		
Corrected Total	54	1143.38182			

Root MSE	2.63282	R-Square	0.6787
Dependent Mean	5.41818	Adj R-Sq	0.6726
Coeff Var	48.59224		

Parameter Estimates

Variable	Label	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Standardized Estimate	90% Confidence Limits	
Intercept	Intercept	1	0.06035	0.61843	0.10	0.9226	0	-0.97497	1.09567
pred_	Predicted Value	1	0.98825	0.09340	10.58	<.0001	0.82383	0.83189	1.14462

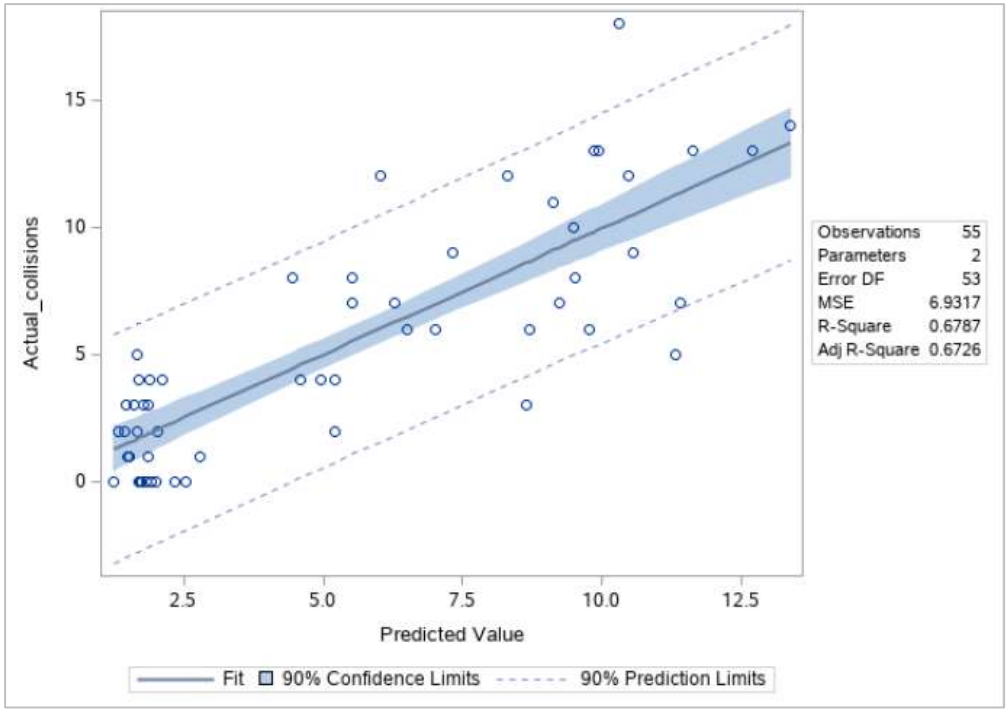

Figure E.1: Fit Plot for Actual vs Predicted Collisions (Divided Segments)

Table E.2: Linear Regression Results for Undivided Segments

Number of Observations Read	195
Number of Observations Used	195

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	934.07334	934.07334	229.43	<.0001
Error	193	785.74204	4.07120		
Corrected Total	194	1719.81538			

Root MSE	2.01772	R-Square	0.5431
Dependent Mean	3.03077	Adj R-Sq	0.5408
Coeff Var	66.57459		

Parameter Estimates									
Variable	Label	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Standardized Estimate	90% Confidence Limits	
Intercept	Intercept	1	0.17504	0.23753	0.74	0.4621	0	-0.21756	0.56763
pred_	Predicted Value	1	0.93833	0.06195	15.15	<.0001	0.73697	0.83594	1.04071

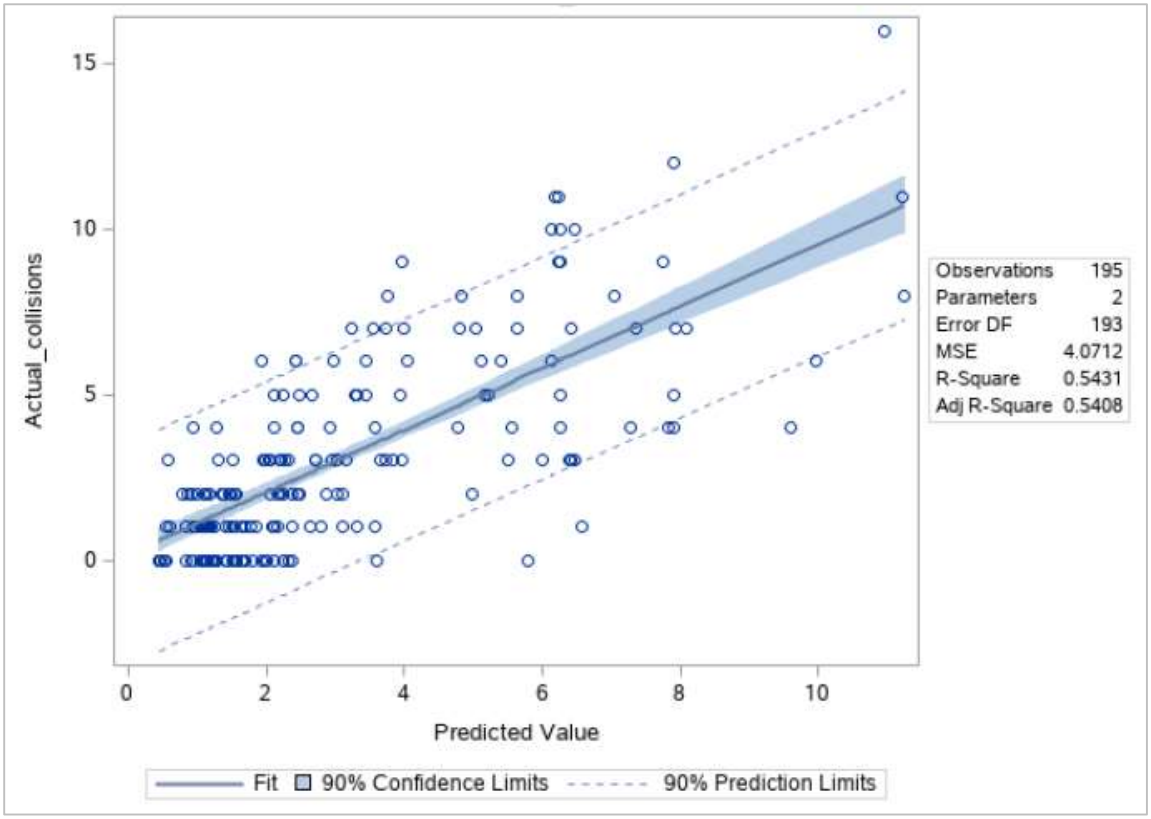

Figure E.2: Fit Plot for Actual vs Predicted Collisions (Undivided Segments)

F. Appendix F: Case Study Results

Table F.1: Case Study Results from the Collision Prediction Models

Year	Route	Control Section	Directional divider	Length	AADT	Surface type	Surface age	Lane width	Grade	Posted speed	Plant Hardiness Zone	Wildlife fencing	Actual collisions	Expected collisions	PFI
2011	2	60	1	14.522	21300	Asphalt	5	3.75	0.91	110	6	0	13	4.61	8.39
2014	2	60	1	14.522	21400	Asphalt	8	3.75	0.91	110	6	0	13	5.38	7.62
2016	2	60	1	14.522	23200	Asphalt	10	3.75	0.91	110	6	0	14	6.52	7.48
2015	2	60	1	14.522	22500	Asphalt	9	3.75	0.91	110	6	0	12	5.99	6.01
2012	2	60	1	14.522	21500	Asphalt	6	3.75	0.91	110	6	0	9	4.90	4.10
2011	11	27	0	14.4519	3860	Asphalt	6	3.75	1.33	100	4	55.42	8	5.79	2.21
2012	105	14	0	10.514	2290	Chip seal	15	3.47	3.68	80	5	0	3	1.13	1.87
2014	11	27	0	14.4519	3760	Asphalt	9	3.75	1.33	100	4	55.42	7	5.33	1.67
2016	11	27	0	14.4519	4150	Asphalt	11	3.75	1.33	100	4	55.42	7	5.60	1.40
2011	105	14	0	10.514	2310	Chip seal	14	3.47	3.68	80	5	0	2	1.16	0.84
2016	105	14	0	10.514	2380	Chip seal	14	3.47	3.68	80	5	0	2	1.17	0.83
2012	123	2	0	4.889	350	Asphalt	14	3.50	2.75	90	4	0	2	1.33	0.67
2011	108	10	0	4.352	490	Chip seal	0	3.50	5.67	80	3	0	1	0.75	0.25
2015	105	14	0	10.514	2380	Chip seal	13	3.47	3.68	80	5	0	1	1.20	-0.20
2016	123	2	0	4.889	500	Asphalt	18	3.50	2.75	90	4	0	1	1.26	-0.26
2015	123	2	0	4.889	510	Asphalt	17	3.50	2.75	90	4	0	1	1.30	-0.30
2014	108	10	0	4.352	430	Chip seal	3	3.50	5.67	80	3	0	0	0.69	-0.69
2016	108	10	0	4.352	680	Chip seal	5	3.50	5.67	80	3	0	0	0.70	-0.70
2015	108	10	0	4.352	680	Chip seal	4	3.50	5.67	80	3	0	0	0.72	-0.72

Year	Route	Control Section	Directional divider	Length	AADT	Surface type	Surface age	Lane width	Grade	Posted speed	Plant Hardiness Zone	Wildlife fencing	Actual collisions	Expected collisions	PFI
2012	108	10	0	4.352	430	Chip seal	1	3.50	5.67	80	3	0	0	0.72	-0.72
2014	105	14	0	10.514	2380	Chip seal	12	3.47	3.68	80	5	0	0	1.23	-1.23
2014	123	2	0	4.889	510	Asphalt	16	3.50	2.75	90	4	0	0	1.33	-1.33
2011	123	2	0	4.889	350	Asphalt	13	3.50	2.75	90	4	0	0	1.36	-1.36
2015	11	27	0	14.4519	4010	Asphalt	10	3.75	1.33	100	4	55.42	4	5.54	-1.54
2012	11	27	0	14.4519	3840	Asphalt	7	3.75	1.33	100	4	55.42	3	5.69	-2.69

Table F.2: Statistical Significance of PFI Values from the CPMs and the HSM SPFs

Model	Route	Control Section	PFI	t-stat	Statistical Significance
Divided CPM	2	60	6.72	8.84	Reject Ho, PFI statistically greater than expected value of 0 (performing worse)
Undivided CPM	105	14	0.42	0.80	Do not reject Ho, not significantly different
	11	27	0.21	0.22	Do not reject Ho, not significantly different
	123	2	-0.52	-1.36	Do not reject Ho, not significantly different
	108	10	-0.52	-2.69	Reject Ho, PFI statistically less than expected value of 0 (performing better)
Undivided HSM SPF	123	2	0.44	1.16	Do not reject Ho, not significantly different
	108	10	-0.22	-1.02	Do not reject Ho, not significantly different
	11	27	-1.67	-1.71	Do not reject Ho, not significantly different
	105	14	-6.17	-11.09	Reject Ho, PFI statistically less than expected value of 0 (performing better)
Divided HSM SPF	2	60	-10.78	-13.73	Reject Ho, PFI statistically less than expected value of 0 (performing better)

Table F.3: Example of t-Test for PFI from the Collision Prediction Models

t-Test: Paired Two Sample for Means

	Variable 1	*Variable 2*
Mean	**6.72**	0
Variance	2.88625	0
Observations	5	5
Pearson Correlation	#DIV/0!	
Hypothesized Mean Difference	0	
df	4	
t Stat	**8.844785443**	
P(T<=t) one-tail	0.000451066	
t Critical one-tail	**2.131846786**	
P(T<=t) two-tail	0.000902131	
t Critical two-tail	2.776445105	

Table F.4: Case Study Data for the HSM SPFs for 2-Lane Highways

Year	Route	Control Section	Length (miles)	AADT	Shoulder width (ft)	CMF_{2r}	Lane width	CMF_{1r}	Horizontal Curvature	CMF_{3r}	Grade	CMF_{5r}	Driveway density per mile	CMF_{6r}	Roadside Hazard Rating	CMF_{10r}
2012	123	2	3.038	350	6.6	0.9968	11.5	1.0029	0.28	1.086	2.75	1.00	1.65	0.84	4	1.07
2011	11	27	8.980	3860	9.8	0.9254	12.3	1.0000	0.02	1.008	1.33	1.00	1.11	0.91	2	0.94
2011	108	10	2.704	490	4.2	1.0141	11.5	1.0070	0.40	1.050	5.67	1.10	1.48	0.84	4	1.07
2016	123	2	3.038	500	6.6	0.9957	11.5	1.0037	0.28	1.086	2.75	1.00	1.65	0.85	4	1.07
2015	123	2	3.038	510	6.6	0.9956	11.5	1.0038	0.28	1.086	2.75	1.00	1.65	0.85	4	1.07
2014	11	27	8.980	3760	9.8	0.9254	12.3	1.0000	0.02	1.008	1.33	1.00	1.11	0.91	2	0.94
2011	123	2	3.038	350	6.6	0.9968	11.5	1.0029	0.28	1.086	2.75	1.00	1.65	0.84	4	1.07
2012	108	10	2.704	430	4.2	1.0116	11.5	1.0062	0.40	1.050	5.67	1.10	1.48	0.84	4	1.07
2014	108	10	2.704	430	4.2	1.0116	11.5	1.0062	0.40	1.050	5.67	1.10	1.48	0.84	4	1.07
2014	123	2	3.038	510	6.6	0.9956	11.5	1.0038	0.28	1.086	2.75	1.00	1.65	0.85	4	1.07
2015	108	10	2.704	680	4.2	1.0221	11.5	1.0098	0.40	1.050	5.67	1.10	1.48	0.85	4	1.07
2016	108	10	2.704	680	4.2	1.0221	11.5	1.0098	0.40	1.050	5.67	1.10	1.48	0.85	4	1.07
2016	11	27	8.980	4150	9.8	0.9254	12.3	1.0000	0.02	1.008	1.33	1.00	1.11	0.91	2	0.94
2015	11	27	8.980	4010	9.8	0.9254	12.3	1.0000	0.02	1.008	1.33	1.00	1.11	0.91	2	0.94
2012	11	27	8.980	3840	9.8	0.9254	12.3	1.0000	0.02	1.008	1.33	1.00	1.11	0.91	2	0.94
2012	105	14	6.533	2290	9.1	0.9254	11.4	1.0172	1.59	1.223	3.68	1.10	18.52	1.40	4	1.07
2011	105	14	6.533	2310	9.1	0.9254	11.4	1.0172	1.59	1.223	3.68	1.10	18.52	1.40	4	1.07
2016	105	14	6.533	2380	9.1	0.9254	11.4	1.0172	1.59	1.223	3.68	1.10	18.52	1.40	4	1.07
2015	105	14	6.533	2380	9.1	0.9254	11.4	1.0172	1.59	1.223	3.68	1.10	18.52	1.40	4	1.07
2014	105	14	6.533	2380	9.1	0.9254	11.4	1.0172	1.59	1.223	3.68	1.10	18.52	1.40	4	1.07

Table F.5: Case Study Data for HSM SPFs for Freeways (PDO Multiple Vehicle)

Year	Route	Control Section	Effective length	AADT	Inside Shoulder width (ft)	CMF_3	Median Width (ft)	CMF_4	Horizontal Curvature	CMF_1	Predicted PDO Multiple Vehicle
2011	2	60	7.383	21300	8.1	0.9683	90.0	0.9277	0.57	1.04	2.6345
2014	2	60	7.383	21400	8.1	0.9683	90.0	0.9277	0.57	1.04	2.6585
2016	2	60	7.383	23200	8.1	0.9683	90.0	0.9277	0.57	1.04	3.1084
2015	2	60	7.383	22500	8.1	0.9683	90.0	0.9277	0.57	1.04	2.9294
2012	2	60	7.383	21500	8.1	0.9683	90.0	0.9277	0.57	1.04	2.6826

Table F.6: Case Study Data for HSM SPFs for 4-Lane Freeways (FI Multiple Vehicle)

Year	Route	Control Section	CMF_3	CMF_4	Lane width (ft)	CMF_2	CMF_1	Predicted Fatal and Injury Multiple Vehicle
2011	2	60	0.9644	0.9251	12.3	0.9887	1.02	1.6164
2014	2	60	0.9644	0.9251	12.3	0.9887	1.02	1.6277
2016	2	60	0.9644	0.9251	12.3	0.9887	1.02	1.8362
2015	2	60	0.9644	0.9251	12.3	0.9887	1.02	1.7541
2012	2	60	0.9644	0.9251	12.3	0.9887	1.02	1.6391

Table F.7: Case Study Data for HSM SPFs for 4-Lane Freeways (FI Single Vehicle)

Year	Route	Control Section	Outside shoulder width (ft)	CMF_3	CMF_8	CMF_4	Outside rumble strips	CMF_9	Barrier	Clear Zone Width (ft)	CMF_{10}	CMF_{11}	CMF_2	CMF_1	Predicted Fatal and Injury Single Vehicle
2011	2	60	9.8	0.9644	1.0154	0.9740	0.6396	0.9649	0.1688	32.81	1.0041	1.0322	0.9887	1.08	6.4453
2014	2	60	9.8	0.9644	1.0154	0.9740	0.6396	0.9649	0.1688	32.81	1.0041	1.0322	0.9887	1.08	6.4648
2016	2	60	9.8	0.9644	1.0154	0.9740	0.6396	0.9649	0.1688	32.81	1.0041	1.0322	0.9887	1.08	6.8111
2015	2	60	9.8	0.9644	1.0154	0.9740	0.6396	0.9649	0.1688	32.81	1.0041	1.0322	0.9887	1.08	6.6776
2012	2	60	9.8	0.9644	1.0154	0.9740	0.6396	0.9649	0.1688	32.81	1.0041	1.0322	0.9887	1.08	6.4843

Table F.8: Case Study Data for HSM SPFs for Freeways (PDO Single Vehicle)

Year	Route	Control Section	CMF_3	CMF_8	CMF_4	CMF_{11}	CMF_1	Predicted PDO Single Vehicle
2011	2	60	0.9683	1.0072	0.9282	1.0427	1.07	11.5827
2014	2	60	0.9683	1.0072	0.9282	1.0427	1.07	11.6303
2016	2	60	0.9683	1.0072	0.9282	1.0427	1.07	12.483
2015	2	60	0.9683	1.0072	0.9282	1.0427	1.07	12.1524
2012	2	60	0.9683	1.0072	0.9282	1.0427	1.07	11.6779

Table F.9: Case Study Results for HSM SPFs for 4-Lane Freeways by Severity

Year	Route	Control Section	Actual collisions	Predicted Crash Frequency	Predicted PDO Multiple Vehicle	Predicted PDO Single Vehicle	Predicted Fatal and Injury Multiple Vehicle	Predicted Fatal and Injury Single Vehicle
2011	2	60	13	22.28	2.63446	11.5827	1.6164	6.4453
2014	2	60	13	22.38	2.65846	11.6303	1.6277	6.4648
2016	2	60	14	24.24	3.10838	12.483	1.8362	6.8111
2015	2	60	12	23.51	2.92937	12.1524	1.7541	6.6776
2012	2	60	9	22.48	2.68256	11.6779	1.6391	6.4843

Curriculum Vitae

Candidate's full name: Jacqueline Kelly Connors

Universities attended: University of New Brunswick
May 2018, Bachelor of Science in Engineering First Division (Civil Engineering)
May 2018, Diploma in Technology Management & Entrepreneurship

Conference Presentations: CSCE National Student Capstone Design Competition, Fredericton, NB, June 2018

Academic Awards:

2019-20	CN Scholarship from the Canadian Transportation Research Forum
2019-20	Esch Foundation scholarship on behalf of the TAC Foundation
2019-20	Albert M. Stevens Scholarship
2019-20	Esterbauer Civil Engineering Graduate Award
2018-19	New Brunswick Innovation Foundation STEM & Social Innovation Award
2018-19	Stevens Wilson Graduate Fellowship in Civil Engineering
2018-19	Esterbauer Civil Engineering Graduate Award
2017-18	Frank Wilson Scholarship
2016-17	Edwin Jacob Special University Scholarship
2015-16	Edwin Jacob Special University Scholarship
2014-15	Edwin Jacob Special University Scholarship
2013-14	UNB Scholarship
2013-14	William and Edward Akerley Memorial Scholarship