University of Nebraska - Lincoln
DigitalCommons@University of Nebraska - Lincoln

Civil and Environmental Engineering Theses, Dissertations, and Student Research

Civil and Environmental Engineering

Spring 5-4-2023

The Effects of Inaccurate and Missing Highway-Rail Grade Crossing Inventory Data on Crash and Severity Model Estimation and Prediction

Muhammad Umer Farooq
University of Nebraska-Lincoln, mfarooq2@huskers.unl.edu

Follow this and additional works at: https://digitalcommons.unl.edu/civilengdiss

 Part of the Transportation Engineering Commons

Farooq, Muhammad Umer, "The Effects of Inaccurate and Missing Highway-Rail Grade Crossing Inventory Data on Crash and Severity Model Estimation and Prediction" (2023). *Civil and Environmental Engineering Theses, Dissertations, and Student Research*. 191.
https://digitalcommons.unl.edu/civilengdiss/191

This Article is brought to you for free and open access by the Civil and Environmental Engineering at DigitalCommons@University of Nebraska - Lincoln. It has been accepted for inclusion in Civil and Environmental Engineering Theses, Dissertations, and Student Research by an authorized administrator of DigitalCommons@University of Nebraska - Lincoln.

THE EFFECTS OF INACCURATE AND MISSING HIGHWAY-RAIL GRADE CROSSING INVENTORY DATA ON CRASH AND SEVERITY MODEL ESTIMATION AND PREDICTION

by

Muhammad Umer Farooq

A DISSERTATION

Presented to the Faculty of

The Graduate College at the University of Nebraska

In Partial Fulfillment of Requirements

For the Degree of Doctor of Philosophy

Major: Civil Engineering

(Transportation Systems Engineering)

Under the Supervision of Professor Aemal J. Khattak

Lincoln, Nebraska

May, 2023

THE EFFECTS OF INACCURATE AND MISSING HIGHWAY-RAIL GRADE CROSSING INVENTORY DATA ON CRASH AND SEVERITY MODEL ESTIMATION AND PREDICTION

Muhammad Umer Farooq, Ph.D.
University of Nebraska, 2023

Advisor: Aemal J. Khattak

Highway-Rail Grade Crossings (HRGCs) present a significant safety risk to motorists, pedestrians, and train passengers as they are intersections where roads and railways intersect. Every year, HRGCs in the US experience a high number of crashes leading to injuries and fatalities. Estimations of crash and severity models for HRGCs provide insights into safety and mitigation of the risk posed by such incidents. The accuracy of these models plays a vital role in predicting future crashes at these crossings, enabling necessary safety measures to be taken proactively.

In the United States, most of these models rely on the Federal Railroad Administration's (FRA) HRGCs inventory database, which serves as the primary source of information for these models. However, errors or incomplete information in this database can significantly impact the accuracy of the estimated crash model parameters and subsequent crash predictions.

This study examined the potential differences in expected number of crashes and severity obtained from the Federal Railroad Administration's (FRA) 2020 Accident Prediction and Severity (APS) model when using two different input datasets for 560 HRGCs in Nebraska. The first dataset was the unaltered, original FRA HRGCs inventory

dataset, while the second was a field-validated inventory dataset, specifically for those 560 HRGCs. The results showed statistically significant differences in the expected number of crashes and severity predictions using the two different input datasets. Furthermore, to understand how data inaccuracy impacts model estimation for crash frequency and severity prediction, two new zero-inflated negative binomial models for crash prediction and two ordered probit models for crash severity, were estimated based on the two datasets. The analysis revealed significant differences in estimated parameters' coefficients values of the base and comparison models, and different crash-risk rankings were obtained based on the two datasets.

The results emphasize the importance of obtaining accurate and complete inventory data when developing HRGCs crash and severity models to improve their precision and enhance their ability to predict and prevent crashes.

DEDICATION

To Professor Mukhtiar Ahmed, my father, my guiding light, and my all-time hero….

ACKNOWLEDGMENTS

First and foremost, I would like to express my deep and sincere gratitude to my PhD advisor, Dr. Aemal Khattak, for giving me an opportunity to work under his kind auspices and providing me with invaluable guidance throughout the course of my doctoral degree. I consider myself extremely fortunate to have had the opportunity to know and be inspired by an individual like Dr. Khattak, whose sincerity and empathy towards his students is truly admirable. His patience, gentleness, and a great sense of humor are one of the best measures of his worth. I want to thank him for instilling in me confidence and for consistently showing appreciation towards my meager research contributions. Additionally, I would like to extend my heartfelt appreciation to his wife and family for their affection and support to my family during our stay in Lincoln, Nebraska.

Thenceforth, I would like to express my sincere appreciation to my committee members, Dr. Ronald Faller, Dr. Massoum Moussavi, and Dr. Yawen Guan, for their invaluable guidance and support throughout this research endeavor. I am deeply grateful to Dr. Faller for his encouragement and kind demeanor during our conversations. As instructors of my graduate courses, I have consistently observed Dr. Moussavi and Dr. Guan's steadfast commitment to the craft of diligent and effective teaching. In particular, the late-evening discussions with Dr. Moussavi after his classes remain some of the fondest memories of my doctoral journey. I also extend my sincere thanks to Dr. Guan for her gracious willingness to answer my research-related inquiries and her warm and helpful disposition.

I would like to convey my gratefulness to the Mid-America Transportation Center (MATC) for awarding me a four-year continuous graduate research assistantship and required funding for this research. I would also like to extend my gratitude to Dr. Yashu Kang and Dr. Huiyuan Liu for their contributions to the data collection process of this research and for their support as peers during their tenure as graduate students at MATC. Additionally, I would like to thank my colleagues and friends at MATC, including Dr. MM Shakiul Haque, Dr. Ernest Tufuor, Dr. Li Zhao, Abdul Farhan, Janet Renoe, and Larissa Sazama, for their assistance and encouragement throughout my doctoral program.

Above all, I would like to thank my dearest parents Professor Mukhtiar Ahmad and Professor Zarifa Rani, for they have been my true strength at times when I flourished or stumbled. I find myself the luckiest child for being fostered by such affectionate parents who have labored principally and uncomforted their lives for my well-being and contentment. Words, in my opinion, will fall short of adequately expressing their value. I also want to thank my wife Fasiha for standing by my side through thick and thin, for being my partner in life and in love, and for always being supportive and filling my days with joy and laughter. The role of my dear siblings, Muhammad Usman Ghani and Dur-e-Nayab Fatima is unquestionably, as well, truly obliging and commendable; since, they have been my true well-wishers and amicable cohorts for which they are and always will be highly credited. I am also thankful to my master's advisor, Dr. Anwaar Ahmad, for his mentorship, and for inspiring me to pursue my PhD in the United States. I am, and always will be, highly indebted to him for his goodwill and candor guidance.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF ABBREVIATIONS

FRA	Federal Railroads Administration
HRGCs	Highway Rail Grade Crossings
OPM	Ordered Probit Model
RPM	Random Parameter Model
WHO	World Health Organization
ZINB	Zero-inflated Negative Binomial Model

CHAPTER 1 INTRODUCTION

1.1 Background

Highway-rail grade crossings (HRGCs) are critical spatial locations of transportation safety because traffic crashes at highway-rail grade crossings are often catastrophic with profound consequences. According to the Federal Railroad Administration (FRA), more than 97% of highway-rail crossings across the US are at-grade, meaning both the tracks and the crossing highway are located at the same elevation (Federal Railroad Administration, 2020). While trains have the right-of-way at HRGCs, there are numerous recorded crashes each year when motorists and other highway users fail to yield the right-of-way to passing trains. Due to train involvement, crashes reported at HRGCs are invariably more injurious than crashes elsewhere on the surface transportation network. Crash outcomes may exacerbate if a train or involved vehicle is carrying hazardous materials (Khattak and Thompson, 2012; Landry et al., 2016; Khan et al., 2018; Farooq et al., 2016; Khattak and Farooq, 2023; Khattak et al., 2023).

For decades, the need to enhance HRGCs' safety has been a significant concern in the US. The Moving Ahead for Progress in the 21st Century (MAP-21) included a separate program that supported safety improvements to reduce the number of fatalities, injuries, and crashes at public HRGCs (MAP-21, 2014). Likewise, the Federal Railroad Administration (FRA) also aims to reduce highway-railroad crossing and trespasser incidents and has initiated multiple programs dedicated to railroad safety, such as the Railroad Safety Management Program (FRA, 2019) and the Risk Reduction Program

(FRA, 2020). Nonetheless, HRGC crash counts, injury severity, and associated safety concerns remain high. In 2021, there were reportedly 2,131 HRGCs crashes resulting in 237 fatalities and 653 injuries across the US (FRA, 2022). There is a continuous need to better understand crash mechanisms, recognize contributing factors to crash frequency and severity, develop countermeasures, and provide direction for policies aimed at improving HRGCs safety. There has been a general decline in the number of HRGCs crashes, injuries, and fatalities, though some years have seen increases when compared to the years immediately before them. However, these numbers continue to be alarming, and require attention (**Figure 1.1**).

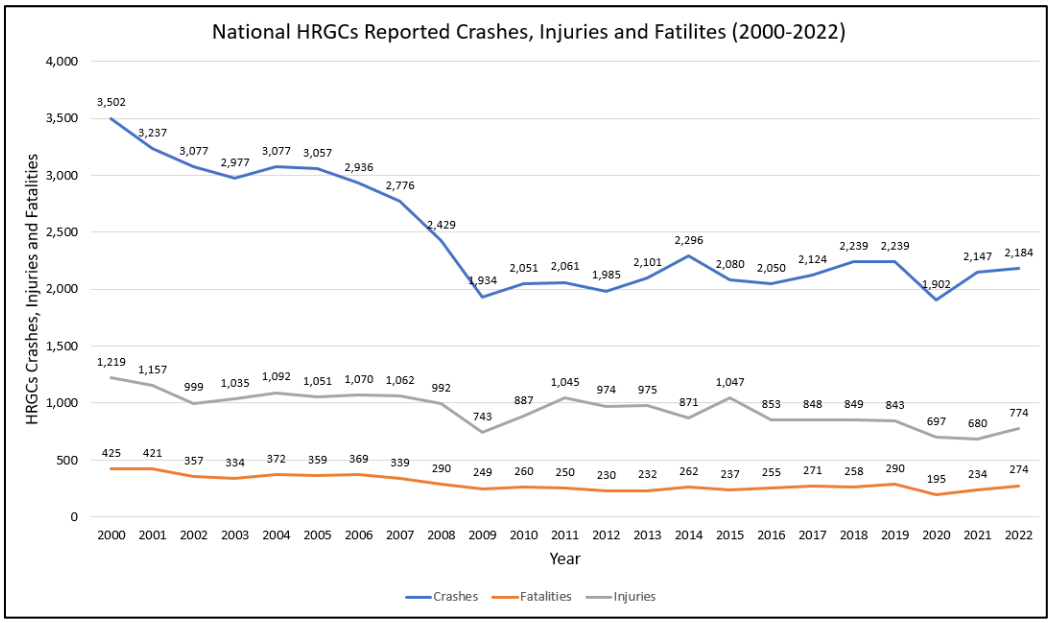

Figure 1.1 National HRGCs Crashes, Injuries and Fatalities from 2000 to 2022 (FRA)

The FRA Office of Safety Analysis manages HRGCs' inventory data and reports on its publicly accessible data portal. This data portal shares railroad safety information, including accident, incident, and inventory data with the public. The FRA regulations on reporting railroad accidents/incidents are included in Title 49 Code of Federal Regulations (CFR) Part 225. The purpose of the regulations in Part 225 is to provide FRA with accurate information concerning hazards and risks that exist on the Nation's railroads. FRA needs this information to efficiently conduct its regulatory and enforcement responsibilities under federal railroad safety laws. The FRA also uses these data to determine railroad safety trends and to establish hazard elimination and risk reduction programs aimed at preventing railroad-crash injuries and fatalities (FRA, 2011).

The FRA accident (also referred as crash) and inventory data on HRGCs include information on location of the crossings, functional classification, weather conditions, visibility, roadway conditions, warning devices, injury-severity levels, safety countermeasures, and many other factors (FRA, 2019). Analyzing the HRGCs' inventory and crash data provides insights and assists in the identification of cause-and-effect relationships about crash probabilities and outcomes. Previously, researchers have utilized FRA's inventory and crash data to process a variety of analytic methodologies to analyze factors that influence the risk of a crash and its severity occurrence at HRGCs. However, data accuracy is important because decisions are made on resource allocation and safety measures based on its analysis.

1.1.1 HRGCs Crash and Inventory Data Collection and Record Keeping

The US Congress passed the Accident Reports Act, Public Law No. 165, in 1910. The amended Accident Reports Act requires railroad carriers to file reports with the Secretary of Transportation on "all accidents and incidents resulting in injury or death to an individual or damage to equipment or a roadbed arising from the carrier's operations during the month." The Secretary of Transportation later delegated authority to the FRA to carry out the Accident Reports Act 103(c)(1) of the United States Code; 49 CFR 1.49(c) (11). The accident/incident reporting regulations at 49 CFR Part 225 were originally issued in response to the 1910 Accident Reports Act. Afterwards, Congress passed the Federal Railroad Safety Act in 1970. The FRA's accident/incident reporting requirements, 49 CFR Part 225, are currently issued under the dual statutory authority of the Accident Reports Act of 1910 and the Federal Railroad Safety Act of 1970 (FRA, 2011). Important crash-related factors on US DOT crossings are documented and kept every month to provide a monthly report, as per the legislation mentioned above.

Salient attributes of HRGCs accident/incident data are presented in **Table 1.1**. "Accident/Incident" is the term used by the FRA to describe the complete list of reportable events. These include collisions, derailments, and other events involving the operation of on-track equipment and causing reportable damage above an established threshold; impacts between railroad on-track equipment and highway users at crossings; and all other incidents or exposures that cause a fatality or injury to any person, or an occupational illness to a railroad employee (FRA, 2016).

Table 1.1 Key Features of FRA Crossing Accident/Incident Data (FRA, 2011)

Accident/incident data key features				
Railroad code	Day	Temperature	View obstruction	Maintenance railroad grouping
Railroad name	Hour	Visibility	Driver condition	Reporting railroad holding company
0Report year	Time	Weather conditions	Total injuries/ fatalities	Number of vehicle occupants
Incident number	County/city/state	Equipment type	Vehicle damage cost	Employees killed in reporting railroad
Incident year	Highway name	Track type	User struck by second train	Number of people on train
Incident month	Public/private	User age	Crossing illuminated	Video taken
Grade crossing id	Estimated vehicle speed	User gender	Driver in vehicle	Report key
Date	Vehicle direction	Highway user action	Passenger killed for reporting railroad	Reporting parent railroad name/ class
Month	Hazmat involvement	Driver passed vehicle	Narrative	

The FRA also maintains data on the current crossings' inventory. The FRA's "Guide for preparing U.S. DOT crossing inventory forms" is a comprehensive document intended to provide operating railroads and states with guidance on completing the US DOT crossing inventory form for inventory-recordkeeping of highway-rail and pathway crossings (FRA, 2016). These crossing inventory forms include valuable information on five key characteristics of HRGCs, which are listed in **Table 1.2**.

The FRA's database also records basic header information for all crossing types and includes information such as, date of revision, reporting agency, and reason for update. The primary operating railroad is required to update important data fields in the inventory database at least every three years as part of the periodic updating process (FRA, 2016; FRA, 2021).

Table 1.2 Data Features Recorded in U.S. DOT Crossing Inventory Forms (FRA, 2016)

Part I – location and classification information	**Part II – railroad information**	**Part III – highway or pathway traffic control device information**	**Part IV – physical characteristics and**	**Part V – public highway information**
State/county/city municipality	Total day and night through trains	Signs or signals	Traffic lanes crossing railroad	Highway system
Highway type and number	Maximum timetable speed	Type of passive traffic control devices associated with the crossing	paved roadway/pathway	Functional classification of road at crossing
Reporting agency	Total switching trains	Crossbucks assemblies	Illuminated crossing	Highway speed limit
Reasons for updating inventory	Typical speed over crossing	Pavement markings	Crossing surface	Annual average daily traffic (AADT)
Train traffic	Type and counts of tracks	Advance warning signs	If intersecting roadway within five hundred feet?	Estimated percent trucks
Quiet zones	Type of train detection system	Gate arms	Does the track run down a street?	Extensively used by school buses?
Crossing type and purpose	Monitoring devices	Gate configuration (2/3/4 quad)	Smallest crossing angle	Emergency services route
Type of trains using the crossing	Signaled track	Total county and types flashing lights	Availability of commercial power	Is crossing on state highway system?
Latitude/ longitude	Year of train count data	Bells		Linear referencing system
Average passenger train count per day	Total transit trains	Channelization devices		
Type of land use	Check if less than 1 movement per day?	Private crossing sign		
		Non-train active warning		
		Highway monitoring devices		

The quality of safety-related data including crashes and inventory is fundamental for the accuracy of analysis, appropriate resource allocation, and the design of effective countermeasures. Accuracy, completeness, consistency, integrity, reasonability, timeliness, uniqueness, validity, and accessibility are examples of popular data quality characteristics and dimensions (**Figure 1.2**). Since data accuracy is a key characteristic of high-quality data, a single incorrect data point can impact the analysis (Chapman, 2005). Decision makers cannot trust the data or make informed decisions unless it is accurate and reliable. This, in turn, can raise operational costs and cause problems for downstream users. Analysts end up relying on flawed reports and drawing incorrect conclusions based on their findings (Suer, 2021; Huh et al., 1990).

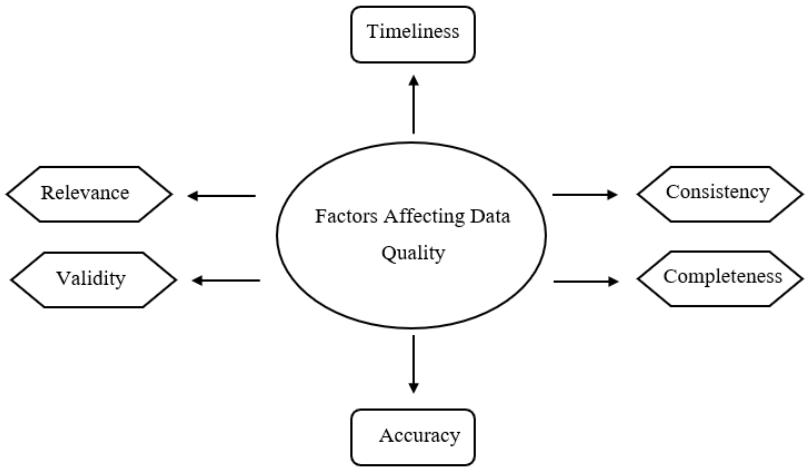

Figure 1.2 Factors Affecting Data Quality

According to the "garbage in, garbage out" (GIGO) principle, the quality of input data is closely related to the analysis' outputs (Oliveira et al., 2005). Past studies have already highlighted flaws in the veracity of U.S. DOTs crash data (Imprialou and Quddus, 2019; Abay, 2015; Alsop and Langley 2001; Amoros et al., 2006; Austin, 1995). However, data quality, on the other hand, is a subjective measure that relates to the degree to which data are appropriate for a given purpose (Imprialou and Quddus, 2019).

The two main problems of data relate to completeness and misreporting. When key parameters used to integrate accident datasets with other explanatory datasets (e.g., traffic volume related data) or variables that define crash outcomes and circumstances (e.g., severity, contributory factors) are incorrect or missing, the problem becomes more apparent, and potentially more serious (Watson et al., 2013).

The data gathering methods, as well as the contents and details of the included attributes, are determined by the intended usage. As a result, data quality may suffer when datasets are used for purposes other than those originally intended. Road crash reports recorded by public authorities (mostly local police) are one example of such datasets with various uses (Imprialou and Quddus, 2019). Crash reports normally contain enough information for their primary purposes of providing evidence in legal cases and developing regional and national safety performance statistics. Despite the restrictions, police crash reports are the primary source of data for road safety research due to a lack of alternatives (Amoros et al., 2006).

When analyzing transportation safety data, there are typically two goals: prediction — the ability to predict outcomes with future input variables, and inference — the extraction of information about how some key contributing factors are associated with the response variable. It should be noted that determining the essential crash contributing factors is an important task of highway safety analysis (Hezaveh and Cherry, 2018).

Accurate crash prediction at HRGCs is important for assisting with decisions related to safety improvements. The safety data recorded by FRA on crossings inventory and accidents/incidents is important as in previous studies, crash and inventory data on rail crossings were used to estimate crash models thereby providing an understanding of crash phenomena, identifying associated factors in an effort to improve safety, and estimating crash-risk ranking of HRGCs for safety improvement resource allocation (Khattak et al., 2020; Pasha et al., 2020; Fischhaber, 2014). These estimated models provide predictions of future accidents and the severity of those accidents.

The literature reviewed in the subsequent chapter delves into previous research studies that have examined the quality of data, with a specific focus on inventory data. Additionally, it will also discuss the various techniques employed in the past for analyzing crash frequency and crash severity and the importance of having accurate inventory data in order to rely on predictive models that are based on such data. This is of paramount importance as it ensures that the models are built on reliable data, and the predictions made by them can be trusted.

In conclusion, this research aims to systematically investigate errors in the Federal Railroad Administration's (FRA) Highway-Rail Grade Crossings (HRGCs) inventory data by conducting comparative studies based on crash prediction and severity modeling. The research used field-validated data and the latest FRA data on HRGCs inventory. Crash prediction and severity models were developed using five years of crash data from the FRA crash database. Through this research, the following questions were addressed: (1) Do the latest FRA data have any missing information or errors? (2) Which variables in the data have the highest percentage of errors and missing information? (3) Are the differences in crash prediction and severity estimates by utilizing FRA and field-validated data statistically significant? (4) Do Crash prediction and severity models based on FRA and field-verified data give statistically different estimates? By investigating these questions, the research aims to improve the accuracy of HRGCs inventory data and enhance the effectiveness of safety measures at these crossings.

1.2 Research Problem

Data modeling approaches are currently one of the most used tools for transportation safety analyses. The use of multiple modeling methodologies to predict crashes at HRGCs and estimate their injury-severity is gaining popularity. However, there is little evidence of data accuracy while undergoing safety analysis. In most cases, it is assumed that the data used for crash prediction and severity analysis is correct, which is not validated in research and requires examination. In addition, the value and need for error-free data for HRGCs' crash prediction and severity assessments are also not well-discussed in the literature.

These data are used to anticipate future crashes, and agencies utilize them to allocate resources and devise preventative measures. They are critical to creating cost-effective improvements in rail-crossings safety. If the data are incorrect to begin with, there is a greater probability of miscalculations in crash prediction, putting many lives at danger of serious injury and death. Furthermore, accurate, timely, and standardized data enable decision-makers to distinguish the primary factors that contribute to the cause of crashes and their outcomes, develop, and evaluate effective safety countermeasures, support traffic (both train and vehicular traffic) safety operations, track progress in reducing crashes and their severity, design effective vehicle safety regulations, and target safety funding.

As discussed previously, data quality is a subjective measure that relates to the degree to which data are appropriate for a given purpose (Imprialou and Quddus, 2019). Seeking 100 percent accuracy in safety data may be unrealistic, as the National Highway Traffic Safety Administration (NHTSA) in 2010 predicted that collecting and coding the

estimated 6.2 million police-reported crashes into a uniform format would cost about a billion dollars annually (NHTSA, 2010). Although the frequency of crashes at or near HRGCs is much lower than highway crashes, it nevertheless, necessitates a large financial budget, extreme precision in data collection, and is realistically, a time-intensive process. At the very least, authorities in charge of collecting data on HRGCs could devise a data recording plan in which they strive for the highest accuracy when recording those data variables that have previously been shown to have the greatest impact on crash prediction and severity at HRGCs (Saccomanno et al., 2003; Yau et al., 2003; Ries, 2007; Raub, 2009; Khattak et al., 2012; Khan et al., 2018; Yan at al., 2010; Gabree et al., 2014; Oh et al., 2006; Nam and Lee, 2006; Haleem and Gan 2015; Eluru et al., 2012; Hu et al., 2010; Salmon et al., 2013; Sharma and Pulugurtha, 2019; Young and Liesman, 2007; Hao et al., 2013; Fan et al., 2015; Das et al., 2021; Mathew and Benekohal, 2021; Mathew and Benekohal, 2020; Das et al., 2022). In addition, based on previous study trends, it is essential to ascertain which data variables require the most correction and updates and which factors are frequently inaccurate and have little intuitive impact on collision injury and severity estimation.

1.3 Research Objective and Hypothesis

This study was undertaken to investigate errors in the FRA HRGCs inventory data by conducting comparative studies based on crash prediction and severity modeling using field-validated data and the FRA data on HRGCs' inventory. For crash prediction and severity modelling, past crash data (2016-2020) from the FRA crash database were utilized.

The objectives of this research were: (1) to investigate accuracy and missing values in HRGCs inventory data and to examine if the missing values in the inventory data follow any pattern; (2) to investigate if there are any statistical differences in expected crashes estimated by using the 2020 FRA APS model by utilizing both FRA and field-validated data; (3) to investigate if there are any statistical differences in crash severity predictions obtained by employing the 2020 FRA APS model based on the FRA and field-validated data; (4) to assess the impact of data inaccuracy on crash frequency modelling, and lastly, (5) to assess the impact of data inaccuracy on crash severity modelling.

The following 5 hypotheses were tested in this research.

- Hypothesis 1: Missing values in FRA HRGCs inventory data do not follow a pattern (Alpha = 5%).
    - In this stage, various data visualization techniques and logical tests were employed to investigate whether the missing values in the inventory data adhered to a specific pattern or occurred randomly. The rationale behind this investigation was to understand the potential impact of missing data on the conclusions that could be drawn from the data and the methods that could be used for analysis. Given that the way in which data is missing can significantly influence the outcome of the analysis, it is crucial to properly understand the nature of missing data.

- Hypothesis 2: There is no difference in expected value of crashes estimated by utilizing FRA and field-validated data (Alpha = 5%).

    - For this stage, field-validated HRGCs inventory data were obtained from data-archives of a key prior Nebraska Department of Transportation (NDOT) study where inventory data from public rail crossings from nine counties (Lancaster, Cass, Douglas, Gage, Jefferson, Otoe, Saline, Sarpy and Saunders) were collected (Khattak et al., 2020). The selection of these nine counties was based on railroad network considerations, urban/rural nature of a county, proximity to the University of Nebraska-Lincoln, and availability of funds in the project. Furthermore, data descriptive analytics were performed to better understand the field-validated dataset. Later, based on the field-validated and FRA inventory datasets, difference in crash prediction (expected crashes) was compared by utilizing 2020 FRA Accident Prediction and Severity (APS) model.

- Hypothesis 3: There is no difference in severity prediction values estimated by utilizing FRA and field-validated data (Alpha = 5%).

    - For this stage, field-validated and FRA inventory datasets were utilized to estimate crash severity based on the 2020 FRA Accident Prediction and Severity (APS) model. Later, the difference in estimated crash severity was compared based on the two datasets.

- Hypothesis 4: There is no difference in estimated parameters' coefficients of new crash frequency prediction models from the two datasets (FRA Vs Field-Validated) (Alpha = 5%).
    - Following careful examination of the 2020 APS model, this portion of the research sought model development for crash frequency prediction to study the impact of data errors on prediction analysis. Two different Zero-inflated Negative Binomial (ZINB) models based on the FRA and field-validated data were estimated for this purpose, and differences in estimated parameter coefficients were analyzed. Furthermore, the impact of a change in models' estimated parameters on crash prediction was investigated by estimating average marginal effects. This aspect of the research helped to indicate how inaccuracies and misinformation within the inventory data of HRGCs may impede the task of developing reliable crash prediction models, underscoring the need for meticulous data validation.
- Hypothesis 5: There is no difference in estimated parameters' coefficients of new crash severity models from the two datasets (FRA Vs Field-Validated) (Alpha = 5%).
    - This portion of the research sought model development for crash severity prediction to study the impact of data errors on crash severity estimation. Two different Ordered Probit models based on the FRA and field-validated data were estimated for this purpose, and differences in estimated parameter coefficients were analyzed.

Furthermore, the impact of a change in crash severity models' estimated parameters on crash severity was investigated by estimating average marginal effects.

Figure 1.3 illustrates the design of the proposed research framework, highlighting the significance of this dissertation within the context of the existing literature on data gaps in Highway-Rail Grade Crossings inventory datasets and their impact on the prediction of crash frequency and severity.

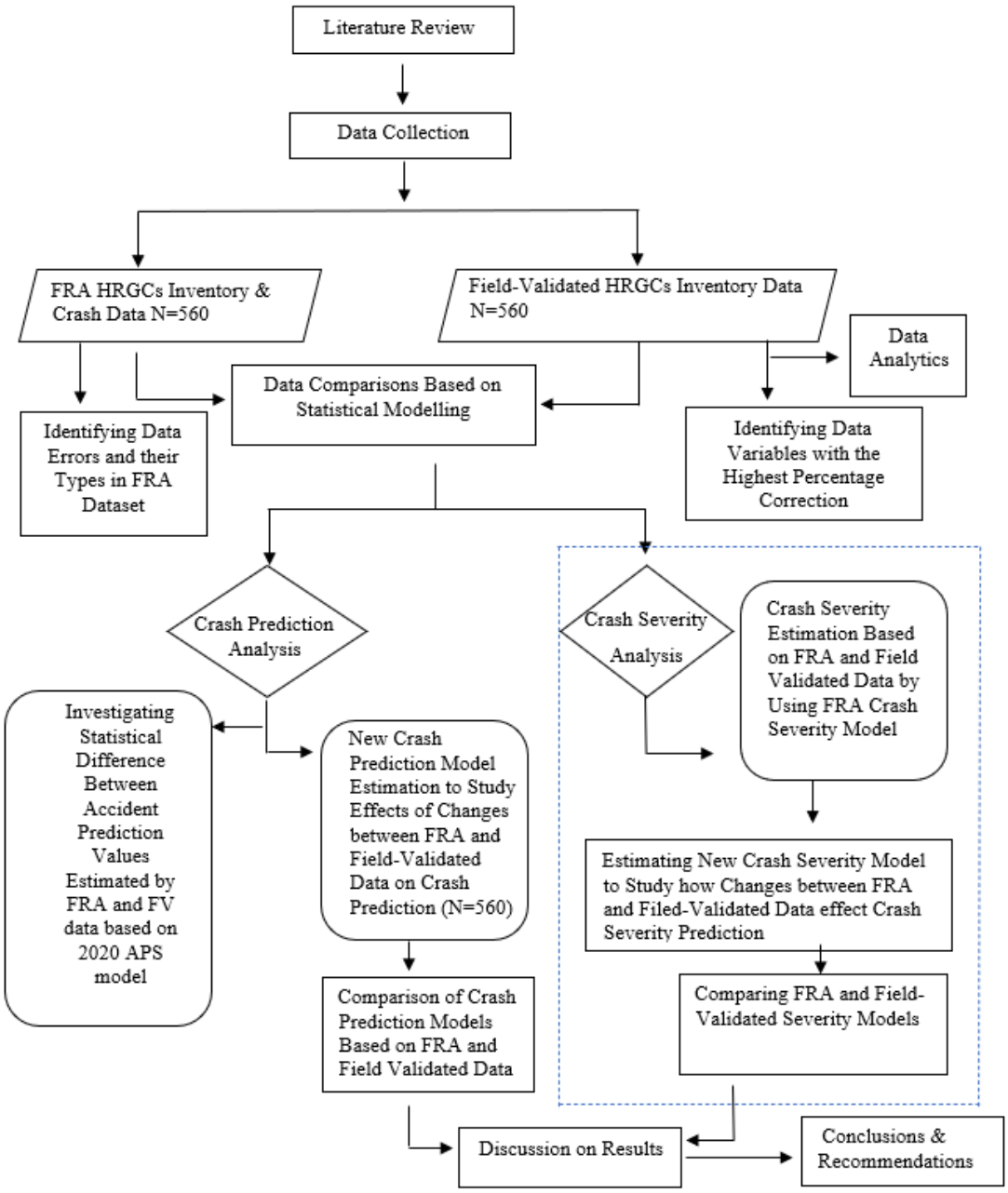

Figure 1.3 Proposed Research Framework

1.4 Dissertation Organization

This dissertation consists of six chapters. Chapter 1 introduces the study background, articulates the research problem, and presents the organization of the dissertation. It serves as an initiation to the work that follows and sets the stage for the subsequent chapters. Chapter 2 presents an in-depth examination of the published literature and open-accessed research reports. The literature review covers a range of topics, including studies on data quality, the impact of data inaccuracies on predictive modeling, previous crash prediction and crash severity studies on Highway-Rail Grade Crossings (HRGCs), and the various methodologies used to investigate crash prediction and crash severity. Additionally, the chapter also covers the Federal Rail Administration's (FRA) 2020 crash prediction and severity model for HRGCs. The chapter ends with identification of gaps in existing research.

Chapter 3 delves into the details of the data collection and field validation process of Highway-Rail Grade Crossing (HRGCs) inventory data. It explains the reasoning behind the selection of nine counties in Nebraska for inventory field validation, the limitations of data collection, and the geographic scope of data recording. The chapter also reports the percentages of total corrected and added missing values, identifies abandoned and non-existent HRGCs, and provides examples of data correction. Furthermore, it includes descriptive statistics for the Federal Rail Administration (FRA) and field-validated datasets, and highlights key insights gained from the data validation effort. Chapter 3 also explores the data errors and missing information within the dataset and presents statistical methodologies for data analysis and visualization. It investigates

logical errors in the dataset and utilizes heatmaps to illustrate the distribution of missing values in inventory data. Additionally, the chapter provides a detailed narrative of the inventory data, highlighting key insights from the Federal Rail Administration (FRA) and field-validated Highway-Rail Grade Crossing (HRGCs) inventory data.

Chapter 4 presents the use of 2020 FRA Accident prediction and severity model to see if there are any statistical differences between crash prediction and severity values by using FRA and field-validated datasets. Chapter 5 presents analysis of developing new comparative statistical models on crash prediction and crash severity to see if there are difference in parameters of the models based on two different datasets. This chapter also provides modelling results, modelling interpretations, visualization of the models and sensitivity analysis by estimating average marginal effects of estimated parameters.

Chapter 6 brings the dissertation to a close by summarizing the work that has been presented throughout the preceding chapters. It presents the key conclusions that have been drawn from the analysis, offering valuable insights into the state of data quality of FRA's Highway-Rail Grade Crossing (HRGCs) inventory data. Additionally, the chapter provides limitations of the research, recommendations for improving the data quality of HRGCs inventory, proposes safety improvements at HRGCs in relation to inventory data, and suggests potential areas for future research.

CHAPTER 2 LITERATURE REVIEW

2.1 Quality of Data and Reporting

Data accuracy, completeness, consistency, and reliability are all aspects of data quality. It is a crucial factor because poor data quality can lead to incorrect conclusions, bad decisions, resource waste, and missed opportunities. On the other hand, reliable data promotes wise judgment, increased effectiveness, and superior performance.

According to Wang et al. (2006) the field of data quality has witnessed significant advances in 21st century. Researchers have moved beyond establishing data quality as a field to resolving data quality problems, which range from data quality definition, measurement, analysis, and improvement to tools, methods, and processes. Furthermore, according to Tayi and Ballou (1998), the term “data quality” can best be defined as “fitness for use,” which implies that the concept of data quality is relative. Data that is deemed suitable for one purpose may not meet the quality standards for another purpose. The rising trend of utilizing data in various contexts, as evidenced by the popularity of data warehouses, has underscored the importance of addressing data quality issues. To ensure fitness for use, it is essential to move beyond the conventional focus on data accuracy and consider other factors that affect its usability.

In Veregin’s (1999) view, the meaning of 'quality' varies depending on the context in which it is applied. Defining quality for data can be more challenging than for manufactured goods, as data lack tangible characteristics that enable straightforward quality assessments. Instead, quality in data is determined by intangible properties like

completeness and consistency. However, upon closer examination, these distinctions may not be as substantial as they seem initially. Data are the outcome of a production process, and the approach in which this process is executed has a significant impact on the reliability of the data.

2.1.1 Types of Errors in Data

There are several types of errors that can occur in the data such as measurement errors, data entry errors, processing errors, non-response errors, selection errors, missing information, observational errors, execution errors, outliers, and others (**Figure 2.1**). The usual reasons for these errors include human errors, limitations in technology or measurement tools, and limitations in the sampling or data collection methods. For example, human errors can lead to mistakes in data entry or survey administration. Limitations in technology or measurement tools can lead to inaccurate or imprecise data. And limitations in sampling or data collection methods can lead to a biased sample (Westerlund, 2007; Helmreich, 2000; Klein et al., 1997).

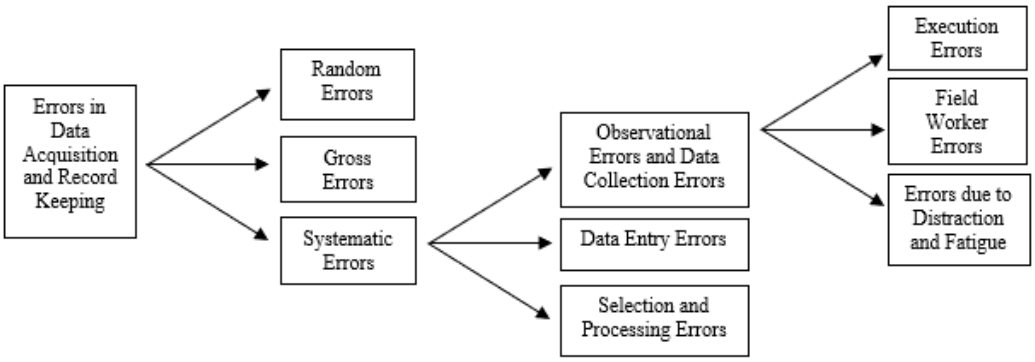

Figure 2.1 Types of Errors

Imperialou and Quddus (2019) conducted a study on the quality of crash data for road safety research. The study found that crash data used in safety analyses often contain inaccuracies or missing information. The authors identified the most significant data quality issues as inaccuracies in crash location and time, difficulties in linking data with traffic data due to inconsistencies in databases, misclassification of crash severity, inaccuracies, and incompleteness of information on involved users' demographics, and inaccurate identification of crash contributing factors. According to the study, there is variability in the scope and severity of data quality problems across attributes, and the degree to which they affect road safety analyses is not entirely clear.

Senders et al. (2020) discussed in great detail, how human errors delve into the complexities of errors in the fields of psychology, engineering, and philosophy. Their work offers a detailed examination of fundamental and significant issues pertaining to the nature and causes of human errors, and it also touches on the factors that cause humans to

commit errors in relation to data collection and information collection. Additionally, Jackle (2008) used historical data to examine measurement and data collection errors. He indicated that inaccurate reporting of event history data is common, either as a result of respondents forgetting to report events or providing incorrect dates. The study discovered that the estimates from the event history data were significantly skewed by measurement errors.

Barchard and Pace (2011) studied the impact of various data entry methods on the accuracy of data and statistical results. They conducted an experiment where 195 undergraduate students were randomly assigned to one of three data entry methods: double-entry, visual checking, and single-entry. After receiving training in their assigned method, participants entered 30 data sheets, each containing six types of data. The results showed that visual checking resulted in 2,958% more errors than double-entry and was not significantly better than single-entry. These data entry errors could have severe consequences on coefficients alphas, correlations, and t-tests. For example, 66% of the visual checking participants produced incorrect values for coefficient alpha, which was sometimes wrong by more than 40%. Furthermore, these data entry errors were difficult to detect, as only 0.06% of the errors were blank or outside of the allowable range for the variables. The authors suggested that researchers should replace single entry and visual checking with more effective data entry methods, such as double entry.

Notably, these errors have a tendency to appear simultaneously and can accumulate, leading to more consequential inaccuracies in the data. Therefore, it is critical to detect and rectify errors as soon as possible during the data collection, processing, and analysis phases to reduce their influence on the outcomes.

2.1.2 Significance of Data Validation

Data validation is the process of ensuring that the data entered into a system is accurate, complete, and consistent. It is an important step in the data management process as it helps to ensure data integrity and fitness for its intended purpose. There are several ways to validate data; however, the common methods are data range validation, data type validation, data format validation, data consistency validation, and data completeness validation (**Figure 2.2**).

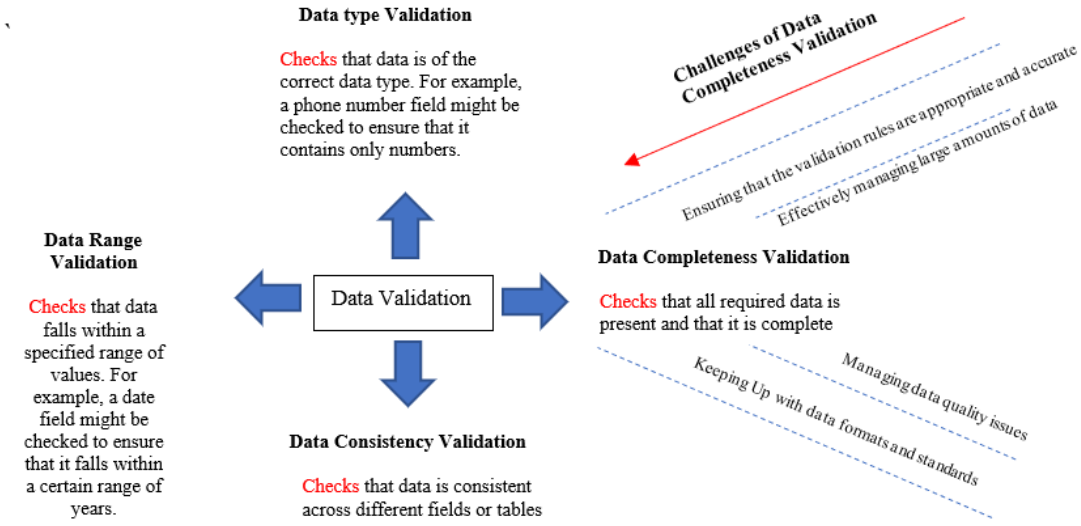

Figure 2.2 Types of Data Validation

Loo conducted a study in 2006 to validate the spatial variables in Hong Kong's crash data from 1993 to 2004. The validation process involved three data sources: crash data, road network data, and district board database. To minimize the need for human

resources and manual validation, a GIS-based system was utilized. The results showed that between 65-80% of police crash records from 1993 to 2004 contained correct road names and district board information. However, for the year 2004, the police crash database had an error rate of 12.7% for road names and 9.7% for district boards. These findings indicate that caution should be exercised by traffic safety researchers when analyzing crash databases, and thorough validation of spatial data should be conducted prior to any scientific analysis.

In a recent study by Breck et al. (2019), a comprehensive data validation process was undertaken to enhance the accuracy and efficiency of machine learning research. The study resulted in the development of a data validation system capable of detecting anomalies. While the limitations and challenges of the new system were acknowledged, the study emphasized the potential benefits of early error detection, including improved quality of predictive models, reduced engineering hours spent on debugging, and a shift towards data-centric workflows in model development.

Furthermore, Souleyrette et al. (2007) conducted validation of crash data from Iowa. The authors sought to examine the discrepancies between database records and crash narratives in the Iowa DOT's Office of Traffic and Safety crash database, as well as the implications of these differences. According to the study, the regular statistical editing of data, combined with the provision of working databases to institutions such as the Center for Transportation Research and Education (CTRE)/Iowa Traffic Safety Data Service, the University of Iowa, and the Iowa Department of Transportation (Iowa DOT), had led to databases that, while considered incomplete by traditional standards, was still considered "public use" due to the dynamic nature of the central DOT database.

Furthermore, the study noted that the final analysis of serious crashes can be delayed, and the crash numbers can continue to change long after the incident year. To address these issues, the authors suggested that the Iowa DOT, its Office of Driver Services, and institutional data users/distributors must establish protocols for data use, distribution, and labeling. In order to do so, data must be collected to determine the extent of the difference between database records and crash narratives and diagrams.

2.1.3 Data Errors in Inventory and Crash Data

It is highly unlikely that the information collected and recorded in transportation related inventory and crash datasets would not contain inaccuracies or inconsistencies. These errors, which can happen at different points during data collection, entry, and analysis, can significantly affect the accuracy and validity of the data.

Past research has shown that in most cases in crash datasets, there exist errors in crash locations. Crash location is reported with multiple different systems around the world such as linear referencing, offset from junction, coordinates and address. Considerable inaccuracies in crash locations have been reported for all systems (Burns et al., 2014, Brown et al., 2015). For instance, Miler et al. (2016) found that 33.5% of the crashes of a relatively large database (8,550 observations) had inaccurate crash location attributes. The inaccuracies may be due to human error, equipment failure (when GPS is used), limited training of personnel or can be inherent to the reporting method (Brown et al., 2015, Imprialou et al., 2015).

Furthermore, past studies have also revealed inaccurate data recording for crash times, crash involved users, crash contributory factors and crash injury severity (Stanton and Salmon, 2009; Cummings, 2002; Beanland et al., 2013; Erkut et al., 2007; Couto et al., 2016; Hao and Kamga, 2017). Some studies have also investigated how changes in operational and safety data affect the operational and safety outcomes of highway work zones and rail network (Haque and Sangster, 2018; Haque, 2022; Haque et al., 2023a; Haque et al., 2023b). Misclassification of crash injury severity is not random, as it has been linked to specific crash or user characteristics; for example, sensitive user injuries seem to be over-classified (Amoros et al., 2006). In addition to the selection bias caused by crash under-reporting, classification bias may also affect analyses that use crash severity to explain crash occurrences, such as severity modeling or multivariate count regression models. To address this issue, some researchers have proposed linking crash data with hospital data prior to analyses (e.g., Watson et al., 2015), which could be quite effective but time and data intensive.

Previous studies have also discussed data under-reporting in highway and bridge inventory databases; however, in these studies, errors in inventory data were not identified per se, but investigations were conducted to reduce the percentage of data errors in inventory data collection practices (Jalayer et al., 2014; Caddell et al., 2007; Bolukbasi et al., 2004). In the literature, however, no previous research has examined under-reporting issues in the FRA crash or inventory data, indicating the need to investigate the current state of data inaccuracy in this database.

2.2 Modeling Approaches and Methodology for Crash and Severity Prediction

As previously stated, there are two goals in analyzing transportation safety data: prediction, which is the ability to predict outcomes with future input variables and, inference, which is the extraction of information about how key contributing factors are associated with the response variable (**Figure 2.3**). This section mentions some key studies on crash prediction and crash severity. Most of these studies have been done in highway safety analysis, and their methodologies were later used in rail-crossing safety research; a review of some of the salient aspects of the modeling approaches is presented next.

Crash frequency is defined as the number of crashes per unit time (e.g., per year or per five years). It serves as the primary indicator of highway safety. Several variables such as driver behavior, road geometry, weather, vehicle characteristics, and roadway environment, can influence crash frequency. The impact of such variables on crash occurrence can vary from case to case, but previous research has shown that both behavioral factors related to driver errors and non-behavioral factors related to road geometry, vehicle, and environment can significantly affect traffic crashes. Researchers typically retrieve only a small number of factors from each class to be used as independent variables in the modeling process (Anderson et al., 2020; Lao et al., 2011).

Lord and Mannering (2010) summarized the data and methodological issues in accident frequency analyses that should be addressed or taken into account in model development and data analyses in the following eleven aspects: over-dispersion, under-dispersion, time-varying explanatory variables, temporal and spatial correlation, low

sample-mean and small sample size, injury severity and accident-type correlation, under-reported accidents, omitted-variables bias, endogenous variables, functional form, and fixed parameters.

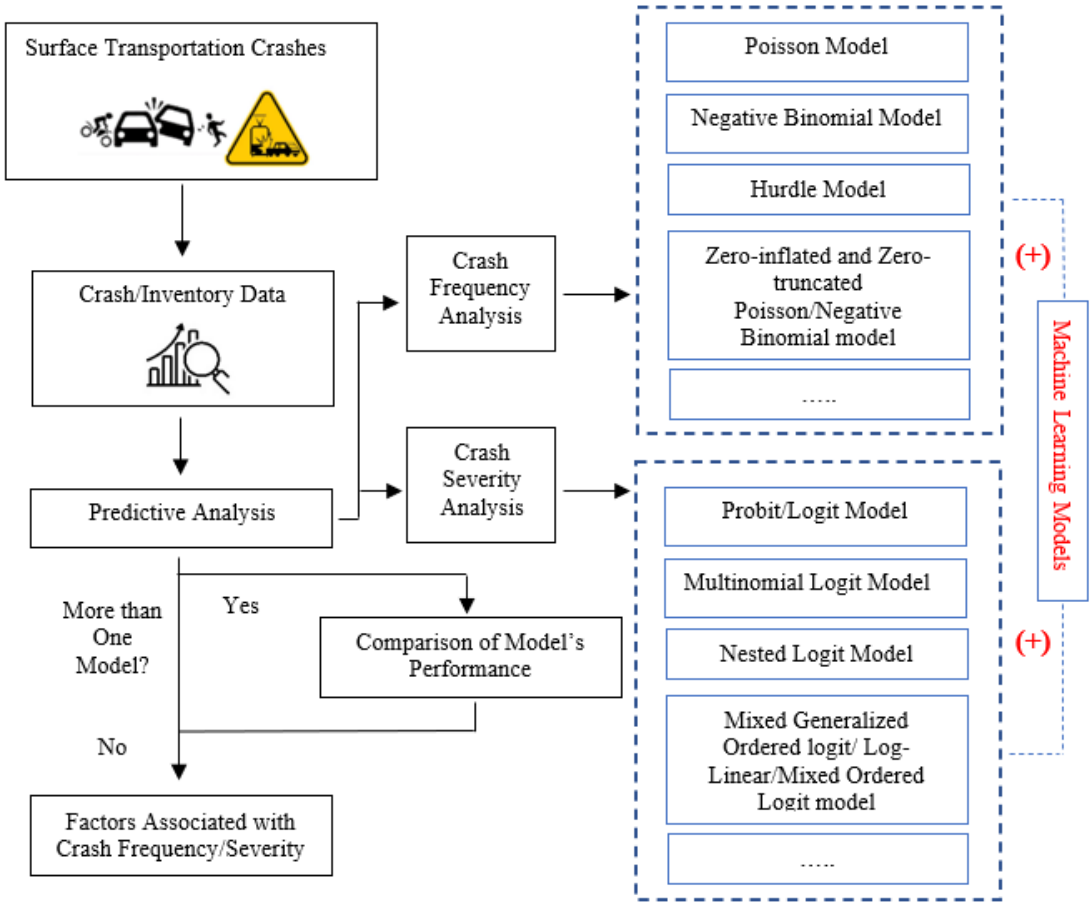

Figure 2.3 Crash Prediction and Severity Models Used in the Past Research

A wide range of methods have been utilized over the years to deal with the data and methodological issues associated with crash frequency data — many of which could jeopardize the statistical validity of an analysis if not properly addressed. **Table 2.1** lists the major existing models used in accident frequency analysis, along with peer research for each model.

Table 2.1 Accident Frequency Literature

Preliminary Modeling Approaches	Past accident frequency research
Poisson regression model	Miaou (1994)
Negative binomial/Poisson-Gamma model	Malyshkina and Mannering (2010)
Zero-inflated model	Lord et al. (2007)
Random-effects model	Wang et al. (2009)
Random-parameter model	Anastasopoulos and Mannering (2009)
Finite mixture/Markov switching	Park and Lord (2009)
Hierarchical/multilevel model	Kim et al. (2007)
Generalized additive model	Xie and Zhang (2008)

The severity of trauma caused by crashes is typically evaluated using the term "crash injury severity". The severity of injury caused to road users is typically used to determine crash injury severity. Typically, five ordinal categories of crash injury severity are modeled, but this can vary depending on the type of research. These are: no apparent injury (i.e., property damage only), possible injury, suspected minor injury, suspected serious injury, and fatal injury (Wang and Abdel-Aty, 2008b; Lord, 2006; Abdel-Aty et al., 2005).

Savolainen et al. (2011) summarized data and methodological issues in crash-injury severity analyses from eight perspectives, some of which are like those used in crash frequency analyses: under-reported crashes, ordinal nature of crash and injury

severity data, fixed parameters, omitted variable bias, small sample size, endogeneity, within-crash correlation, and spatial and temporal correlations. Crash severity analysis can be performed in a variety of ways depending on the purpose.

Some researchers investigated how geometric, traffic, and environmental factors affect accident severity at specific traffic sites associated with different severity levels (e.g., fatal, seriously injured, injured). While these studies typically use each crash as a unit, analysis can also be done based on the driver vehicle units involved in crashes to assess individual severity. Statistical techniques such as the multinomial logit model, ordered logit or probit model, nested logit model, mixed logit models, mixed ordered logit model and others have been used to study crash-injury severities. **Table 2.2** lists the primary models used for crash-injury severity analysis, along with case studies for each method.

Table 2.2 Crash Injury Severity Literature

Preliminary Modeling Approaches	Past crash injury outcome research
Binary logit/probit model	Haleem and Abdel-Aty (2010)
Classification and regression tree	Chang and Wang (2006)
Multinomial logit/probit model	Islam and Mannering (2006)
Nested logit model	Savolainen et al. (2011)
Ordered logit/probit model	Wang and Abdel-Aty (2008a)
Mixed logit model	Anastasopoulos and Mannering (2012)
Mixed ordered logit model	Srinivasan (2002)
Log-linear model	Chen and Jovanis (2000)
Mixed generalized ordered logit model	Eluru et al. (2008)

2.3 Crash Frequency-Based Analysis for HRGCs

Several studies are available dealing with crashes on HRGCs; **Table 2.3** presents a brief overview of these on estimating crash prediction at HRGCs, including the methodology used, data resources utilized, and explanatory variables considered in the

modeling process. Most of the studies have relied on generalized linear modelling where Poisson family models were used.

Table 2.3 Past Relevant Studies on HRGCs Crash Prediction Modelling

Year	Authors	No of observations	Data type	Location	Context	Method used	Types of explanatory variables
2010	Yan at al.	6244 train–vehicle crashes	27 years of FRA HRGCs database (1980-2006)	United States	Using hierarchical tree-based regression model to predict train–vehicle crashes at passive highway-rail grade crossings	Hierarchical tree-based regression model	Crossbucks only and crossbucks combined with stop signs, and stop-sign treatments
2006	Oh et al.	Crash data of 162 HRGCs	1998-2002 Korean national railroad accident database	South Korea	Accident prediction model for railway-highway interfaces	Poisson model, gamma model and zero-inflated Poisson model	Traffic volume, average daily train volumes, the proximity of crossings to commercial areas, time duration between the activation of warning signals and gates, and the distance of the train detector from crossings
2006	Nam and Lee	100 highway–rail grade crossings	Korean national railroad accident database	South Korea	Accident frequency model using zero probability process	Zero-inflated models	Roadway characteristics, guardrails, number of tracks, control device indicator, warning time
2020	Lu et al.	Past 19 years crashes on 5,713 HRGCS	FRA HRGCs safety and inventory data	North Dakota, United States	A gradient boosting crash prediction approach for highway-rail grade crossing crash analysis	Gradient boosting (GB) model	Traffic exposure factors: highway traffic volume, railway traffic volume, and train travel speed
2016	Lu and Tolliver	344 HRGCS data from 1996- 2014	FRA HRGCs safety and inventory data	North Dakota, United States	Accident prediction model for public highway-rail grade crossings	Conway–maxwell–Poisson model Bernoulli model The hurdle poison model	Warning devices, highway pavement condition, appearance of pavement markings, appearance of interconnection/pre-emption, smallest crossing angle, appearance of pullout lane, functional classifications of highway, train traffic density, highway user types, weather conditions, track conditions, highway traffic density,

							maximum train speed, location
2019	Zheng et al.	Past 19 years data on 354 crashes on 5,713 HRGCs	FRA HRGCs safety and inventory data	North Dakota, United States	Predicting highway–rail grade crossing collision risk by neural network systems	Neural network (NN) model	AADT, presence of flashing lights, highway stop signs, presence of cross buck
2020	Keramati et al.	3,310 Crossings, including 475 crash records	FRA HRGCs safety and inventory data	North Dakota, United States	A simultaneous safety analysis of crash frequency and severity for highway-rail grade crossings: the competing risks method	Competing risks method	Crash information, type of train service, train detection, availability of commercial power, distance to nearby roadway intersection
2020	Brod et al.	Data from 9,870 at grade crossings	FRA HRGCs safety and inventory data	United States	New model for highway-rail grade crossing accident prediction and severity	Zero-inflated negative binomial model	Exposure, AADT, maximum timetable speed, total trains, type of surface, warning lights and gates

2.4 Crash Injury Severity-Based Analysis for HRGCs

Understanding the significant factors that influence crash injury severity at public HRGCs is critical for developing countermeasures to reduce deaths and injuries at these locations as part of the MAP-21 mission. Investigation of crash injury severity risk factors at HRGCs is more crucial compared to traditional roadways due to the additional complex interaction between highway users and the HRGCs environment. **Table 2.4** presents salient studies on HRGCs crash severity analysis.

Table 2.4 Past Relevant Studies on Injury Severity Analysis at HRGCs

Year	Authors	No of Observations	Data Type	Location	Context	Method Used	Explanatory Variables
2012	Eluru et al.	14,532 observations	HRGCs collision data from 1997 to 2006	United States	A latent class modeling approach for identifying vehicle driver injury severity factors at highway-railway crossings	Latent segmentation based ordered logit model	Driver age, time of the accident, presence of snow and/or rain, vehicle role in the crash and motorist action prior to the crash
2015	Haleem and Gan	5,528 public HRGCs	2009 through 2013 FRA data on HRGCs	United States	Contributing factors of crash injury severity at public highway-railroad grade crossings in the U.S.	Mixed logit model	Female highway users, young and middle-aged drivers, non-passing of standing vehicles, presence of warning bells
2015	Fan et al.	7,414 crashes at public HRGCs	2005 to 2012 FRA data on HRGCs	United States	Analyzing severity of vehicle crashes at Highway-Rail Grade Crossings with multinomial logit modeling	Multinomial logit model	Truck-trailer vehicles in snow and foggy weather conditions, development area types (residential, commercial, industrial, and institutional), and higher daily traffic volumes
2013	Hao and Daniel	15,881 highway–rail grade crossing crashes	2002 and 2011 FRA data on HRGCs	United States	Severity of injuries to motor vehicle drivers at highway–rail grade crossings in the United States	Probit model	Peak hour, weather, visibility, vehicle type, vehicle speed, annual average daily traffic, train speed, driver age and gender, area type, and type of highway pavement
2010	Hu et al.	410 HRGCs crashes	2001 to 2005 Public rail crossing data	Taiwan	Investigation of key factors for accident severity at railroad grade crossings by using a logit model	Logit model	Number of daily trains, trains speed, highway features, crossing features, and traffic controls
2019	Zhao et al.	1,409 train-pedestrian crashes	2007 to 2016 pedestrian crashes at HRGCs	United States	A clustering approach to injury severity in pedestrian-train crashes at highway-rail grade crossings	Latent class clustering (LCC) and Binary logit models	Absence of flashing lights, advance warnings, rural areas, lower visibility conditions, and older pedestrians

2018	Zhao and Khattak	303 crashes at or near HRGCs	2002 to 2013 data on reported motor vehicle crashes at or near HRGCs.	Nebraska, United States	Injury severity in crashes reported in proximity of rail crossings: The role of driver inattention	Random parameters binary logit regression model	Seatbelt usage, presence of passengers, driver's age, gender, weather, train involvement, highway speed limit, road surface condition, and lighting condition

2.5 2020 Accident Prediction and Severity Model (APS) by FRA

State agencies have also contributed to crash prediction research at HRGCs. Since the late 1980s, the United States Department of Transportation's Accident Prediction and Severity (APS) model has been utilized by federal, state, and municipal governments to analyze the likelihood of crashes at highway-rail grade crossings. The FRA Office of Research and Development was informed by state and local government agencies that the old APS gives similar outcomes for most crossings in their jurisdictions, making it impossible to distinguish among hazardous HRGCs. Most crossings with no incidents in the previous 5 years, as well as similar-site specific factors (such traffic volumes and warning systems), account for the low variance among APS-generated ratings. There was a need to address these difficulties by new consensus methods of analysis.

In addition, the old APS model had three independent accident prediction models, one for each of the three principal types of grade crossing warning devices: passive (signage only), flashing lights, and gates. According to Brod et al. (2020), there's no compelling need to break crash prediction into three different models when the warning device type might be treated as a grade crossing characteristic in a single model for all crossings. Furthermore, the results from the various models may be incongruent. For example, the APS predicts a higher risk for crossings with the identical features but for a more protective warning device for some combinations of grade crossing characteristics.

It's simple to see how a grade crossing risk analysis in a corridor or region could produce extremely dubious assessments of relative risk between similar crossings with different warning device types (Brod et al., 2020).

Another shortcoming in the APS model was that it didn't provide a way to see if risk measures at different crossings differed statistically. For instance, consider two crossings with annual accident rates of 0.21 and 0.23, respectively. There is no factual basis for treating two crossings differently if the difference in estimated risk is not statistically significant. These were compelling reasons to establish a new grade crossing safety model, an alternative to the APS, is to effect evidence-based safety management of grade crossings. As a result, the FRA funded research into the development of a new model that considered current consensus analysis methods and data patterns. The new model also attempted to address several weaknesses in the old APS model, providing analysts with a more dependable tool (Brod et al., 2020).

Below is a functioning version of the two-part 2020 accident prediction model. The first part is a count model, and the second part is a zero inflated model. Before considering the possibility of excess zeroes, the count model is for predicted crashes. The zero-inflation model is used to calculate the likelihood of an inflated zero. (An "inflated zero" is a zero-crash count that does not result from the characteristics of a grade crossing; rather, it is zero because the crossing accident risk is effectively zero.) The total number of trains is the explanatory variable for the zero-inflated model; that is, the fewer trains at a grade crossing, the greater the probability of an excess zero.

$$N_{count\ predicted} = e^{[\beta_0+\beta_1*IExpo+\beta_2*D_2+\beta_3*D_3+\beta_4*RurUrb+\beta_5*XsurfID2s+\beta_6*IAADT+\beta_7*IMaxTnSpeed]} \quad \text{eq (1)}$$

$$P_{InflatedZero} = \frac{z}{1+z} \quad \text{eq (2)}$$

$$z = e^{[\gamma_0 + \gamma_1 * ITotalTrains]} \quad \text{eq (3)}$$

$$N_{predicted} = N_{Countpredicted} * (1 - P_{InflatedZero}) \quad \text{eq (4)}$$

Where $N_{Countpredicted}$ are predicted accidents of count model (data for left-hand side of regression are counts of accidents at crossings in 5-year period, $P_{InflatedZero}$ is the probability that the grade crossing is an "excess zero", $N_{predicted}$ are predicted accidents after accounting for excess zeroes, $IExpo$ is the exposure, equal to average annual daily traffic times daily trains, D_2 and D_3 show indicator variable for warning device type lights and gates, $RurUrb$ shows rural or urban classification of road leading to HRGCs, $XsurfID2s$ shows type of surface used (and can be timber, asphalt concrete, rubber and there combination), $IMaxTnSpeed$ indicate maximum timetable speed (integer value between 0 and 99), $IAADT$ shows average annual daily traffic, and $ITotalTrains$ show total number of daily trains.

Following are the estimated coefficients (Table 4.1 in Brod et al., 2020):

ZINB regression count model coefficients (negative binomial with log link)

Variable	Estimate	Std. Error	Z value	Pr(>|z|) (p-value)	Significance Code
(Intercept)	-8.35922	0.32079	-26.059	< 2e-16	***
lExpo	0.19023	0.02866	6.638	3.18e-11	***
D2	-0.28478	0.04806	-5.926	3.10e-09	***
D3	-0.85770	0.04089	-20.976	< 2e-16	***
RurUrb	0.39346	0.03162	12.444	< 2e-16	***
XSurfaceID2s	0.13182	0.01715	7.686	1.52e-14	***
lMaxTtSpd	0.68760	0.68760	22.702	< 2e-16	***
lAadt	0.10626	0.10626	3.511	0.000446	***
Log(theta)	-0.25934	.08867	-2.925	.003447	**

ZINB regression zero-inflation coefficients (binomial with logit link)

Variable	Estimate	Std. Error	z-value	Pr(>|z|) (p-value)	Significance Code
(Intercept)	1.17084	0.19001	6.162	7.19e-10	***
lTotalTr	-1.01088	0.08452	-11.961	< 2e-16	***

Significance codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

The FRA also developed a crash injury severity model that estimated the probabilities of three injury types: fatal, injury, or PDO. These estimates' explanatory variables are grade crossing characteristics. As a result, the study attempted to model three variables:

Probabilities to estimate – Fatal

$$P(\text{acctype} = \text{fatal} \mid A) \qquad \text{eq (5)}$$

Probabilities to estimate - Injury

$$P(\text{acctype} = \text{injury} \mid A) \qquad \text{eq (6)}$$

Probabilities to estimate-PDO

$$P(\text{acctype} = PDO \mid A) \qquad \text{eq (7)}$$

keeping in mind that severity probabilities Sum to 1

$$P(\text{ fatal } \mid A) + P(\text{ injury } \mid A) + P(PDO \mid A) = 1 \qquad \text{eq (8)}$$

An ordered logit model was estimated to estimate crash injury severity of HRGCs crashes.

Following are the estimated coefficients according (Table 5.1 Brod et al., 2020):

Variable	Coeff.	Estimate	Std. Error	z-Value	Pr(>|z|) (p-value)	Confidence Level
(PDO | Injury)	κ_1	–3.05946	0.19728	–15.5082	< 1E-16	> 99.9
(Injury | Fatal)	κ_2	–4.60832	0.20025	–23.0127	< 1E-16	> 99.9
lMaxTtSpdSq	β_1	–0.29043	0.02368	–12.2637	< 1E-16	> 99.9
lThru	β_2	–0.10696	0.02408	–4.44116	< 9E-06	> 99.9
lSwitch	β_3	0.13847	0.04140	3.34481	< 9E-04	> 99.9
lAadt	β_4	–0.03317	0.01354	–2.45074	< 2E-02	> 99.0
Rural Urban	β_5	–0.14500	0.05106	–2.83989	< 5E-03	> 99.5
D1	β_6	–0.20471	0.06004	–3.40951	< 7E-04	> 99.9
Summary Statistics						
Residual Deviance				**AIC**		
18224.88				18224.88		

where, lMaxTtSpdSq is a variable based on the square of maximum timetable speed (mtts) at a grade crossing, lThru is the number of daily through trains at the crossing, lSwitch is the number of daily switch trains at the crossing, and l Aadt is the average annual daily highway traffic at the crossing. These variables were transformed by using the equation:

$$L(X) = \log\left[1 + \frac{X(\bar{X}-1)}{X}\right] \qquad \text{eq (9)}$$

Where $\bar{X}$ is the mean value of the variable X. The transformation achieves two objectives. The transformed variable is calculable at 0, and the value of the transformed variable is equal to the log of the untransformed variable at its mean value.

Accident Severity Forecast Formulas

$$Z_i = \sum_{k=1}^{6} \beta_k X_{ki} = \beta_1 \cdot l\,\text{MaxTtSpdSq}\; q_i + \beta_2 \cdot l\,\text{Thru}_i + \beta_3 \cdot \text{lSwitch}_i + \beta_4 \cdot \text{lAadt}_i + \beta_5 \cdot \text{RuralUrban}_i + \beta_6 \cdot \text{D}_i \quad \text{eq (10)}$$

$$P(Y_i = PDO) = \frac{1}{1+\exp(Z_i - \kappa_1)} \quad \text{eq (11)}$$

$$P(Y_i = \text{Injury}) = \frac{1}{1+\exp(Z_i - \kappa_2)} - \frac{1}{1+\exp(Z_i - \kappa_1)} \quad \text{eq (12)}$$

$$P(Y_i = \text{Fatal}) = 1 - \frac{1}{1+\exp(Z_i - \kappa_2)} \quad \text{eq (13)}$$

Where, the subscript i indicates a grade crossing, Y_i is the variable indicating accident type (fatal, injury or PDO). And κ_1 is a coefficient of the threshold separating PDO from injury accident, and κ_2 is a coefficient of the threshold separating injury from fatal accident.

2.6 Gaps in the Literature

There have been studies where prediction models have been used based on crash and inventory data from different surface transportation modes. But the credibility of these models is questionable because there is a higher likelihood of errors in the dataset that has been used to estimate those models. For rail-crossings, much of the associated research relied on data from the FRA's public highway-rail grade crossings database,

ignoring the fact that the data became erroneous over time due to not being updated on a regular basis. This problem may lead to inaccurate crash forecasts and injury severity estimation at HRGCs, which are critical for railroad agencies' resource allocation. This issue necessitated the development of a study that can use verified and up-to-date inventory and crash data to estimate predictors of crash frequency and severity at HRGCs, allowing for a more comprehensive understanding of crash hazards and better policy decisions on highway-rail grade crossings.

CHAPTER 3 RESEARCH DATA

This chapter discusses in detail the process of data collection, data filtration, data assessment and analytics.

3.1 Data Collection

The present study has utilized three datasets: (a) the FRA crash database on HRGCs, (b) the FRA inventory database on HRGCs, and (c) the field-validated inventory database on HRGCs. Crash history and inventory records of 560 HRGCs across nine counties in Nebraska were extracted from the publicly available FRA website for the analysis. Furthermore, inventory data of these 560 HRGCs were field validated for comparative analysis.

3.1.1 Data Description of Crash Data

The FRA collects and analyzes data on rail crossing crashes. The annual crash data for HRGCs can be downloaded from the FRA safety data website. Following each grade crossing crash, railroads submit Form 6180.57 to the FRA, which contains crash information on HRGCs. Each crash is represented as a single row in Form 6180.57 and data is downloaded as a single table (in Excel or Access formats). Crash data from the five-year period between 2016 and 2020 were examined for the analysis.

The crash database includes a wide range of useful information, such as the type of crossing, location of the crash, cause and outcome of each crash, the presence of safety devices, crash data summaries, vehicle and train information, injured occupant information, environmental and weather conditions at the time of the crash, hazmat involvement, time of the crash, visibility and temperature, driver condition, lightning conditions of the crossing, reporting railroad holding company, number of people on train, total fatalities and injuries etc. The available crash/accident data form and data fields are presented in **Appendix B**.

3.1.2 Data Description of Inventory Data

The Federal-Aid Policy Guide (FAPG 924.9(a)(1)) mandates that each state must maintain a system for collecting and preserving records of crash, traffic, and highway data, including the characteristics of both highway and train traffic at railway-highway grade crossings (U.S. Department of Transportation, 1991). The National Highway-Rail Crossing Inventory Reporting Requirements also require states and railroads to exchange information and promptly update crossing data records as changes occur, to ensure the crossing inventory serves as an effective database. Consequently, the FRA records information from each state and maintains a comprehensive database of HRGCs across the United States.

Furthermore, the data recording of inventory features of HRGCs is governed by the “FRA Instructions for Electronic Submission of U.S. DOT Crossing Inventory Data, Grade Crossing Inventory System (GCIS), v2.9.0.0, Released: 7/2/2019.” Railroads, transit authorities, and States are responsible for electronically submitting updates for

grade crossing data through FRA Form 6180.71. Each grade crossing in the inventory is represented by its own row in the data, reflecting the latest information provided by the submitting agency. This inventory recordkeeping encompasses data on the location, design, functional classification, type of land use, quiet zones, channelization configuration, posted speed limits and other important safety features of each crossing. It also includes information on the ownership of the crossing, the number of tracks, and the average daily train and vehicle traffic.

The FRA estimates that there are about 250,000 public and private highway-rail grade crossings in the United States (FRA, 2022). There are approximately 136,000 of these HRGCs, that are thought to be active, with at least one train passing through on a daily average. Approximately 85% of these active crossings have active warning mechanisms, like flashing lights and gates. The most common type of crossing is the "warning-device-only" crossing, which has active warning devices but no physical barrier to block vehicles. These crossings make up about 53% of the total inventory. The second most common type is the "passive" crossing, which has no active warning devices and relies on signage and pavement markings to alert drivers. These crossings make up about 25% of the total inventory. The remaining 22% of crossings are protected by a combination of active warning devices and physical barriers (FRA, 2019).

Local coordinators typically submit updates to the HRGCs inventory data using FRA-approved guidelines for recording various field names and value assignments. Accordingly, updated values for the designated field names must be submitted by authorized users through field in-person data collection. The field names, field descriptions, and values used in this study are listed in **Appendix A**, based on the FRA

HRGCs inventory database. However, reporting updates for the inventory database does not always require verification from other agencies. As a result, some fields, like traffic and train volumes, and timetable speed are not routinely updated, potentially resulting in inaccurate or out-of-date data. This raises concerns for states and railroad companies and affects the reliability of crash prediction models based on the database. Additionally, FRA offers geospatial information to the public on rail networks, including the location of HRGCs and Amtrak stations, identified by latitude and longitude in the database (FRA, 2022; Khattak et al., 2020).

In this research, inventory data and crash data were integrated to understand effects of data inaccuracies on crash prediction and severity models. For this purpose, various important fields are integrated in the data set such as crossing ID, State, Country, nearest city name, functional classification of roads, rural or urban nature of the area where HRGCs are located etc. The integrated data also consisted of other important aspects from HRGCs inventory data, including details on train traffic frequency at rail crossings, such as the number of daylight and nighttime thru trains, transit trains, and passenger trains, as well as the number of main, siding, yard, and transit tracks. The data also included important inventory characteristics related to crossing safety, such as the presence of signs and signals, the number of crossbuck assemblies, stop and yield signs, bells, flashing lights, channelization devices, and gate configurations. Additionally, the data encompassed information on the crossing highway, including the number of traffic lanes, pavement type, highway functional classification, street name, and posted speed limit, etc.

3.1.3 Data Description of Field-Validated Inventory Data

The objective of the field-validation was to highlight inaccuracies in the Federal Rail Administration's inventory data and to assess the extent of errors present. This study involved the manual inspection and verification of 560 HRGCs in Nebraska as part of the 2020 rail-crossing safety project by the Nebraska Department of Transportation (Khattak et al., 2020). A substantial number of database variables, including roadway speed limit, pavement type, warning devices at crossings, signs, signals, and land use, were thoroughly examined and confirmed.

The study considered public, at-grade, and operational grade crossing locations. A manual inspection of public rail crossings in Lancaster County was conducted, and field conditions were compared to the database, leading to correction of any inaccuracies and addition of missing information as found. Based on elements like the railroad network, the county’s urban/rural characteristics, proximity to the University of Nebraska-Lincoln, and the availability of project funding, this validation process was later expanded to eight additional counties, including Cass, Douglas, Gage, Jefferson, Otoe, Saline, Sarpy, and Saunders.

3.2 Data Assessment

This section covers in detail the process of verification of inventory information for 560 HRGCs in Nebraska. Additionally, it discusses the identification of missing values and logical errors in the inventory data. Data validation checks are performed, and missing values are highlighted through data visualization techniques such as heat maps, Upset plots, density plots and MCAR test. Finally, the section concludes with a

descriptive analysis of the field-validated data for understanding and summarizing the main characteristics of the dataset.

3.2.1 Data Filtration

In order to comprehend the geographic distribution of HRGCs in Nebraska, a filter was applied to the inventory database to select only the HRGCs that were operational, public, and at-grade crossings throughout the state. This helped visualize the pattern of HRGCs land use and was essential in formulating a data acquisition plan. The result of this filtering process revealed 2,853 public, operational, and at-grade crossings in Nebraska, as illustrated in **Figure 3.1**.

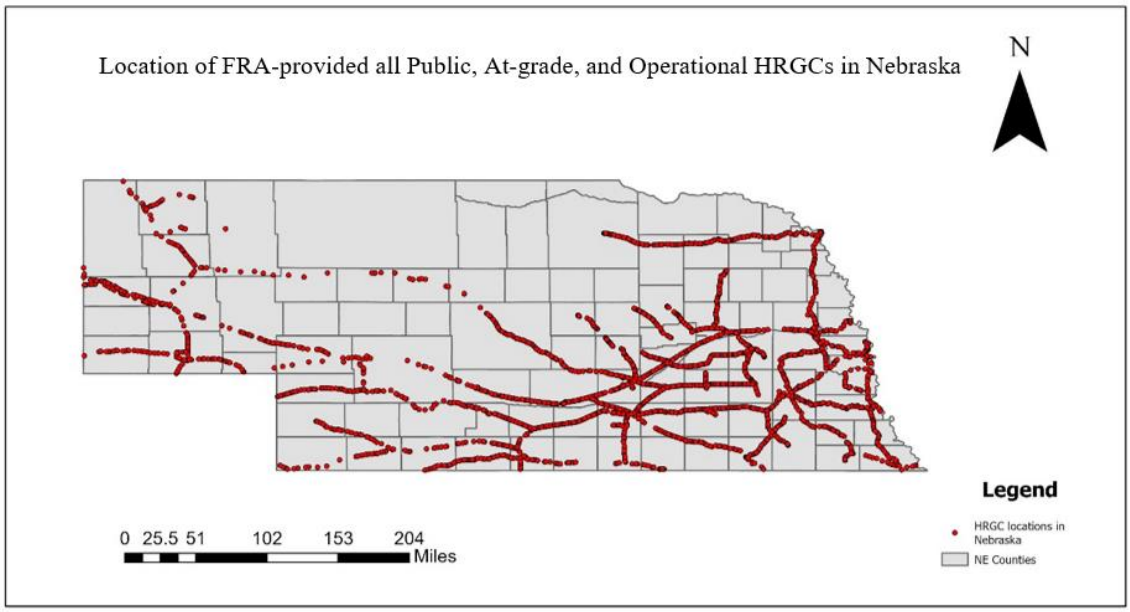

Figure 3.1 Locations of FRA-provided all Public, At-grade, and Operational HRGCs in Nebraska

Nine counties were chosen for the study, as previously mentioned, based on factors such as population, railroad network considerations, a county's urban/rural makeup, accessibility, and the project's financial resources. The locations of 560 public, at-grade, and operational HRGCs in a subset of nine Nebraska counties (Lancaster, Cass, Douglas, Gage, Jefferson, Otoe, Saline, Sarpy, and Saunders) are depicted on a map in **Figure 3.2**.

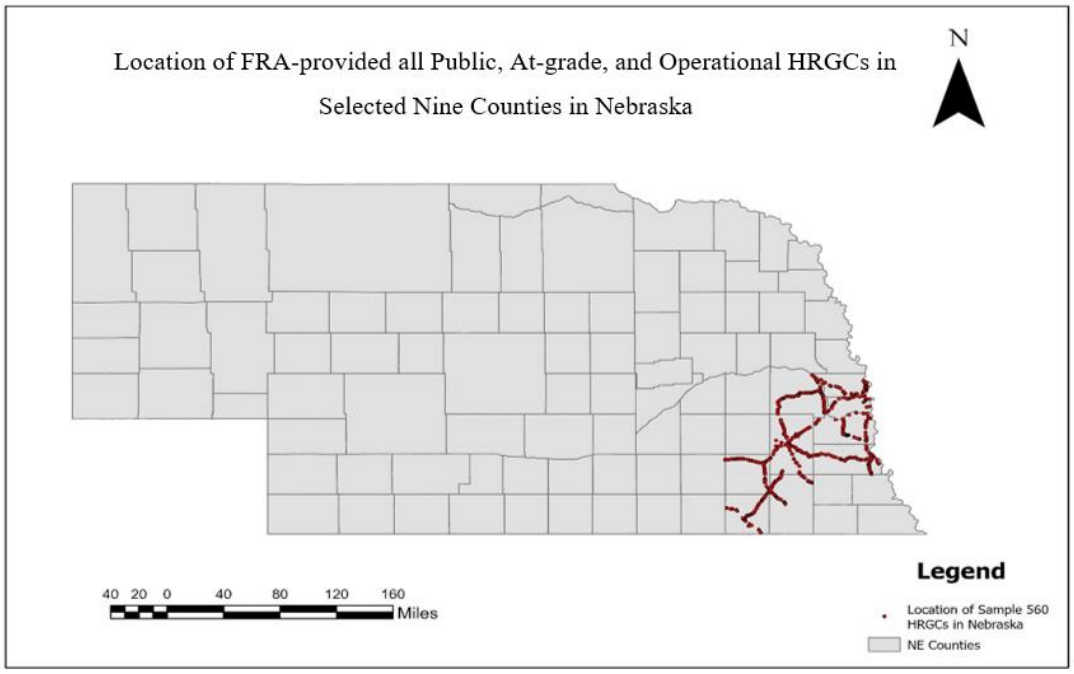

Figure 3.2 Locations of FRA-provided Public, At-grade, and operational HRGCs in selected nine counties in Nebraska

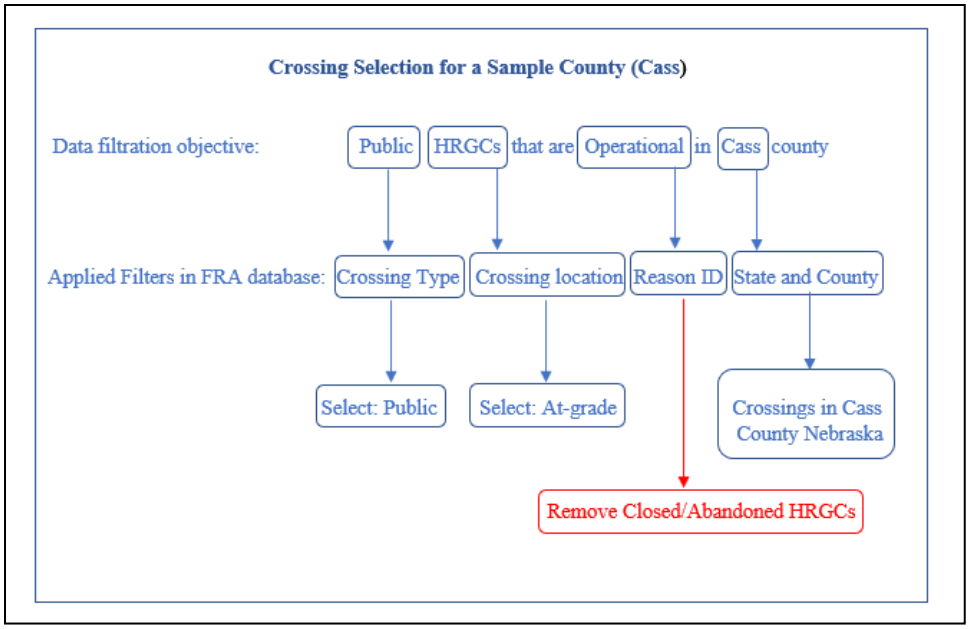

Figure 3.3 FRA HRGCs Data Filtration Process for a Sample County (Cass)

The data filtration process and inventory database variables used for the county of "CASS" are depicted in **Figure 3.3**. Out of the 58 rail crossings present in the FRA HRGCs database, only 55 HRGCs were selected after exclusion of private, elevated (grade-separated), and closed crossings. Similar data filtration was conducted for the remaining eight counties as well.

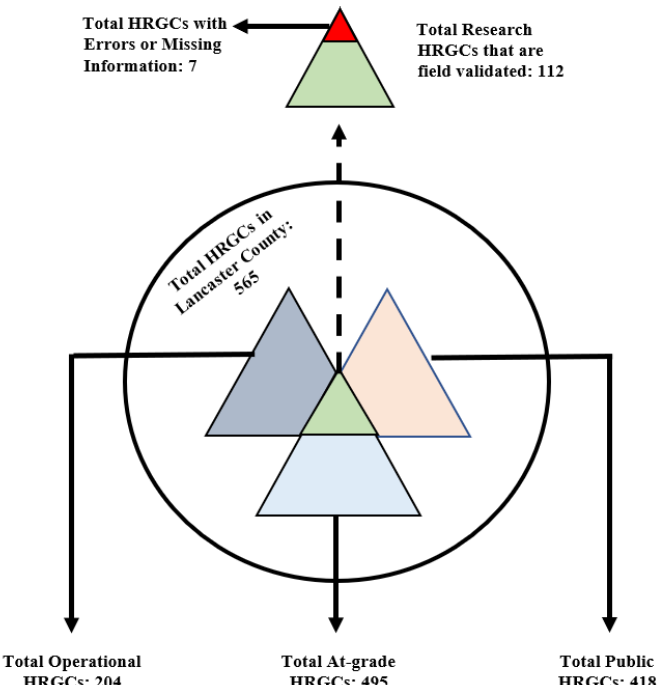

Figure 3.4 Results for Filtration Process for Lancaster County

Figure 3.4 demonstrated how HRGCs were chosen for field validation through a filtering procedure by screening rail crossings that were operational, private, and public. In order to better comprehend a data validation case, HRGCs from Lancaster County were considered; there were initially 565 counties in the county, of which 204 were operational, 495 were at-grade, and 418 were public. 112 crossings were collected with all three attributes present, so a total of 112 crossings were field validated. Of these 112 crossings, field validation revealed that 7 crossings inventory-data in the FRA dataset had missing information or data errors.

Figure 3.5 Data Correction Example, Crossing 064112B (Khattak et al., 2020)

For every HRGCs visited, a comprehensive examination of 53 database variables was conducted and digital images were acquired. Any inaccuracies in the database were rectified in accordance with field conditions, and missing values were added if those attributes were found on the field. The disparity between the recorded information in the FRA HRGCs inventory database and the actual conditions at crossing 064112B is vividly exemplified in **Figure 3.5**, which highlights a discrepancy in the presence of yield signs, pavement type, approach surface type, and pavement markings. Additionally, crossing storage distances were validated. For example, **Figure 3.6** illustrates how the storage distance in the FRA data is measured as 64 feet, but field validation changed the value to 40 feet. The safety of HRGCs has been strongly correlated with storage distance (Keramati et al., 2020).

Orthophoto of Crossing 072946C

3D Street View of Crossing 072946C

Field Validation of Crossing 072946C (Crossing Surface, Advanced Warning Devices, Storage Distance, Crossbuck Assemblies etc.)

Field Validation of Road Conditions near Crossing 072946C (Pavement Markings, Traffic Lanes, Lane Type, Crossing Angle etc.)

Crossing Numbe	DATA SOURCE	Latitude	Longitude	City	Crash Severity	AADT	Storage Distance	ITDTRAIN	MTTSPD	IMTTSPD	WDTLIT	Mast-Mou	Total Cour	Gates		Roadway	WDTGATE	Bells
072946C	FRA	40.7746	-96.705	LINCOLN	No Crash	8900	64	0	10	2.3979	1	Incandesc	5	2		2	1	
072946C	FIELD VALID	40.7746	-96.705	LINCOLN	No Crash	8900	40	0	10	2.3979	1	Incandesc	5	2		2	1	

Correction in Storage Distance of Crossing 072946C based on Field Validation

Figure 3.6 Data Correction Example, Crossing 072946C

Orthophoto of Crossing 064130Y

Field Validation of Flashing Light Pairs of Crossing 064130Y

Crossing Num	DATA SOU	Latitude	Longitude	City	AADT	TDTRAINS	ITDTRAINS	MTTSPD	WDTLIT	Mast-Mounted Flas	Total Count of Flashing Light Pairs	Gates	R
064130Y	FRA	40.7991	-96.7247	LINCOLN	8600	0	0	75	1	Incandescent	2	2	
064130Y	FIELD VALI	40.7991	-96.7247	LINCOLN	8600	0	0	75	1	Incandescent	4	2	

Correction in Total Count of Flashing Light Pairs of Crossing 064130Y based on Field Validation

Figure 3.7 Data Correction Example, Crossing 064130Y

The illustration in **Figure 3.7** is about data validation for a specific location (crossing 064130Y) where a discrepancy was found between the actual number of flashing light pairs (4) and the number recorded in the Federal Railroad Administration's (FRA) inventory data set (2). The validation involved a field check and revealed the difference, indicating that the FRA inventory data set was incorrect.

Figure 3.8 An Example of an Abandoned Crossing 083524P (Khattak et al., 2020)

Furthermore, **Figure 3.8** depicts a case example of an abandoned crossing (identified by ID 083524P) that remains listed in the Federal Railroad Administration's High Rail Grade Crossing inventory database.

Furthermore, to fulfill the objectives of the study, crash data from the past 5 years were extracted from the FRA crash database. The fields (discussed in **Table 1.1**) available in the crash database consist of a series of categories, such as crash information, crossing information, train information, environmental factors, highway characteristics, etc. For instance, the crash information includes time of crash, AM or PM, injury severity outcome, number of injuries or fatalities of roadway users, number of injuries or fatalities of railroad employees, number of injuries or fatalities of train passengers, etc. Environmental factors at the time of crash consist of temperature, weather conditions, lighting conditions, etc. Train information includes number of cars, number of locomotives, type of train, train speed, etc. Additionally, other important factors such as release of hazardous materials are also included.

To fulfill the objectives of the study, crash data from the past 5 years were also extracted from the FRA crash database. The available fields in the crash database, outlined in **Table 1.1** (Chapter 1), encompass various categories such as crash information, crossing information, train information, environmental factors, highway characteristics, and more. For example, the crash information category includes details such as the time of the crash, whether it occurred in the AM or PM, the severity of injuries, and the number of injuries or fatalities for roadway users, railroad employees, and train passengers. The environmental factors at the time of the crash, such as temperature and weather conditions, as well as lighting conditions, were also taken into consideration. Train information, such as the number of cars, locomotives, type of train, and train speed, were also incorporated in the integrated dataset.

The crash data on our selected 560 HRGCs revealed that for a 5-year period (2016-2020), a total of 34 crashes were recorded on 28 HRGCs. However, for all public and open crossings in Nebraska, 171 crashes were recorded. This illustrates that our sample HRGCs accounted for approximately 17% of the total crashes recorded in Nebraska in the past 5 years.

3.2.2 Identification of Logical Errors and Missing Values

Identifying inaccuracies in HRGCs inventory data is critical for the objective of this study. Inaccuracies in data can take the form of missing values, transcription errors, or errors caused by illogical values in the data. Although 560 crossings were chosen for field validation due to their proximity to the University of Nebraska-Lincoln, the availability of funds in the project, and other factors described above, an initial in-office data validation of the complete set of public and operational crossings in Nebraska was

performed to determine the magnitude of logical errors in the overall dataset. Certain rules were established for initial validation purposes, and if a variable was discovered to contradict the rule, that specific crossing from the inventory dataset would fail the validation test. Crossings that showed no-failing based on validation rules, on the other hand, are considered “Passed” crossings. **Figure 3.9** and **Table 3.1** highlight key findings from the initial data validation of the FRA inventory dataset, which used validation rules to evaluate logical errors in the data.

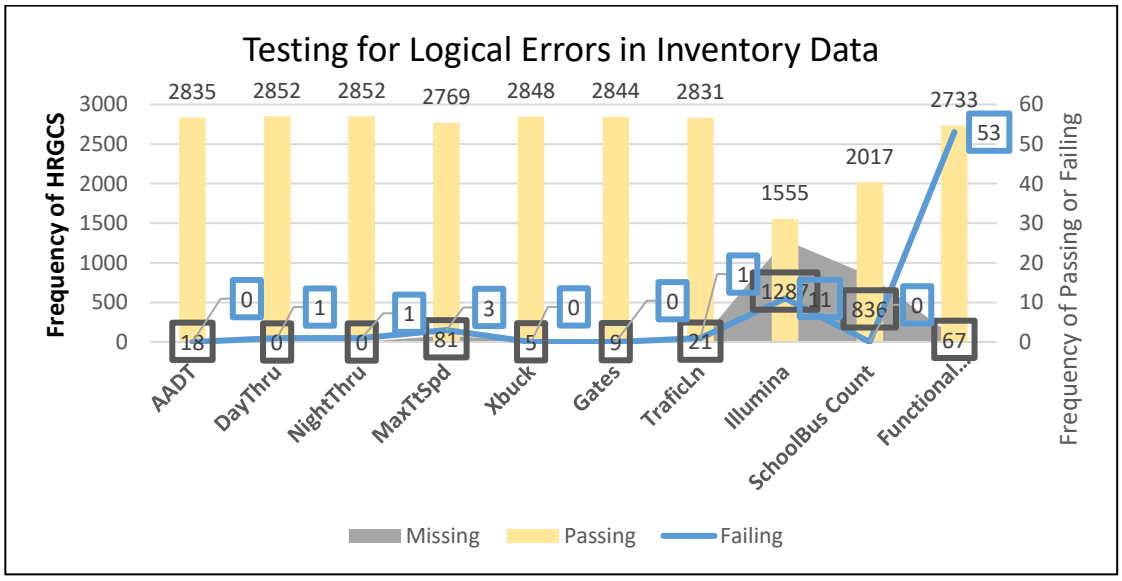

Figure 3.9 Initial Data-Validation to Check for Logical Errors in Inventory Data (N=2853)

Initial data validation of the entire Nebraska operational and public crossings shown in **Figure 3.9** revealed that the functional classification variables in the data contained the most logical errors. According to the "Functional Classification (Development)" variable, 53 crossings were considered to be in rural areas according to

the FRA data, when they were actually in cities with respect to "Nearest" variable. In the same manner, 11 crossings were indicated to be illuminated in inventory data which were not illuminated in case of crashes (in crash data). This step demonstrated that most crossings have logical values in the data and passed the initial data-validation.

Table 3.1 Validation Rules for Testing for Logical Errors in Inventory Data

Variables	Validation Rule	Validation Warnings
AADT	AADT=0 For crashes where reportedly vehicles were involved	0
Day & Night Thru Trains	DayThru=0 or NightThru=0 but a crash was reported	2
MaxTtSpd	MAXTtsp (Maximum trains speed) > (Recorded train speed during crash +10 mph)	3
Xbuck	Crossbuck=6 for a one lane highway	0
Gates	Number of Gates=0, with Gate Configuration	0
TraficLn	Traffic Lanes=0 with AADT>0	1
Illuminated	Not illuminated crossings where crash data shows crossing illumination	11
School-Bus Count	School bus Count >0, for crashes involving "School Buses" in "Highway User" Category	0
Functional Classification (Development)	Functional Classification "Rural" for "In City" category in "Nearest" Variable	53

It was also determined that missing values account for most errors (**Table 3.2**) in inventory data. Furthermore, the missing values are analyzed additionally to discover if trends may be found in the missing values of the FRA inventory dataset. This would aid in determining whether missing data in important candidate variables for crash prediction models were random or followed a trend. The analysis of missing data can be used to make suggestions on which variables should be targeted to ensure prevention of missing values. Data on all public crossings in Nebraska were kept for analysis before examining 560 sample HRGCs to assess the nature and degree of missing values. Following the

same process as the initial in-office data-validation, the complete inventory data of Nebraska public crossings were checked for missing values before analyzing the missing values in the selected inventory data of 560 HRGCs.

Table 3.2 Missing Values in Candidate Variables in Nebraska HRGCs Inventory Data (N=2853)

Variable	Number of Missing Values	Variable	Number of Missing Values
AADT	18	Cross Bucks	5
Percentage of Trucks	23	Max Train Speed	81
Highway Speed	194	Total Trains	326
Crossing Illuminated	1287	E-Monitor Device	1706
Traffic Lanes	21	Health Monitor	1782
Bells	40	Pavement Marking	16
Gate Config	238	Monitoring Device	2312
Gates	9	Crossing Surface IDs	16

The heat map in **Figure 3.10** shows that the highest percentage of missing values are in E-monitoring device, Health-monitoring (health state awareness monitoring of the entire wheel-track system), LED indicator, illuminated crossings, monitoring device indicator and highway distance from crossings. However total trains have 11.43% of missing values, number of bells have 1.4 % of missing values. The least percentage of missing values were observed in AADT, maximum train speed, gates, and total daily trains. Following the analysis of the entire Nebraska crossing dataset, the missing values for sample 560 HRGCs were investigated further. In addition to heatmaps, various plots were created to display different aspects of missing values in the data. Additionally,

variables that were either candidate variables or had been previously used for crash and severity predictions were evaluated.

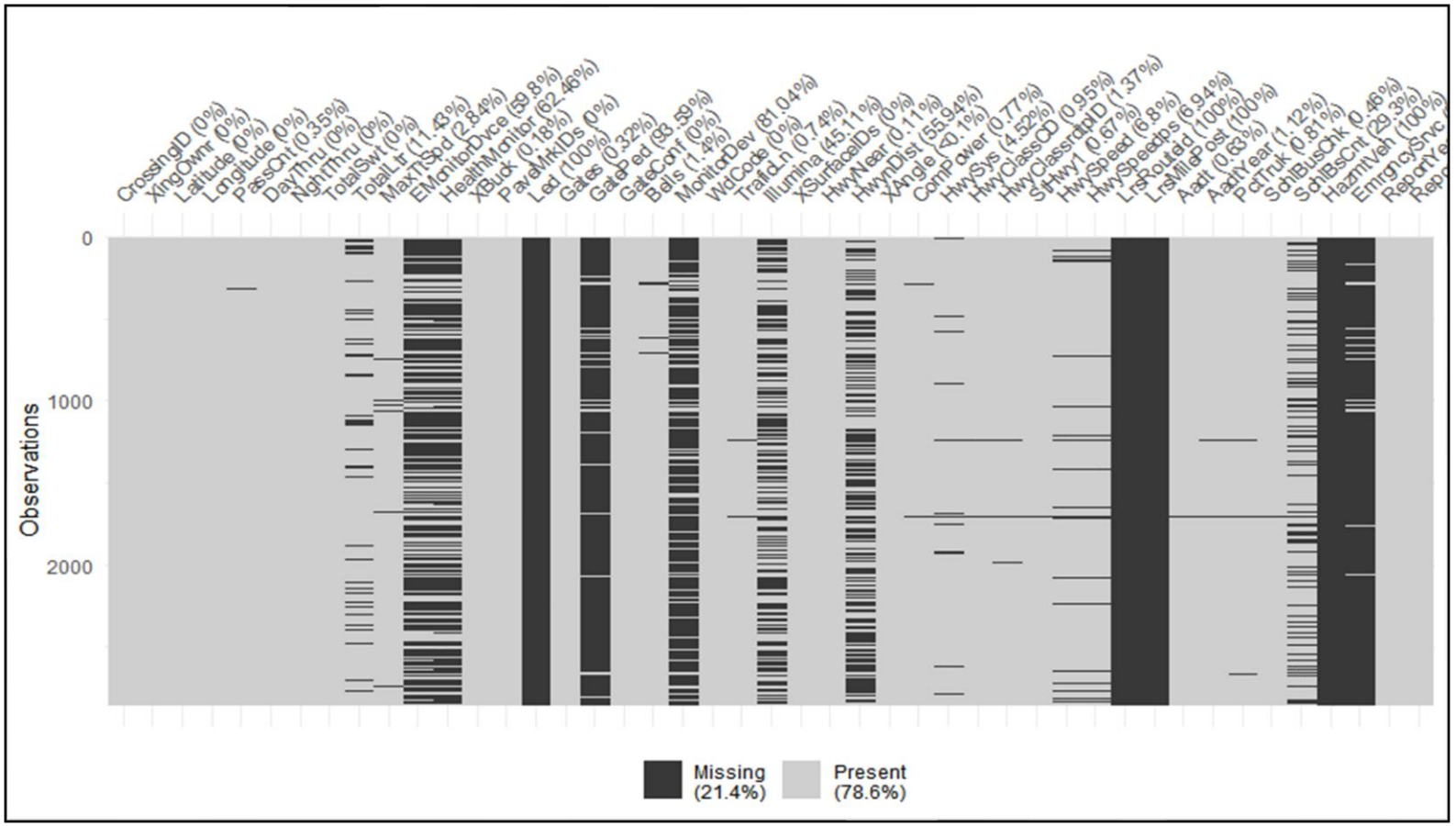

Figure 3.10 Heat Map of Missing Values in Candidate Variables of FRA HRGCs Inventory Data for Nebraska (N=2853)

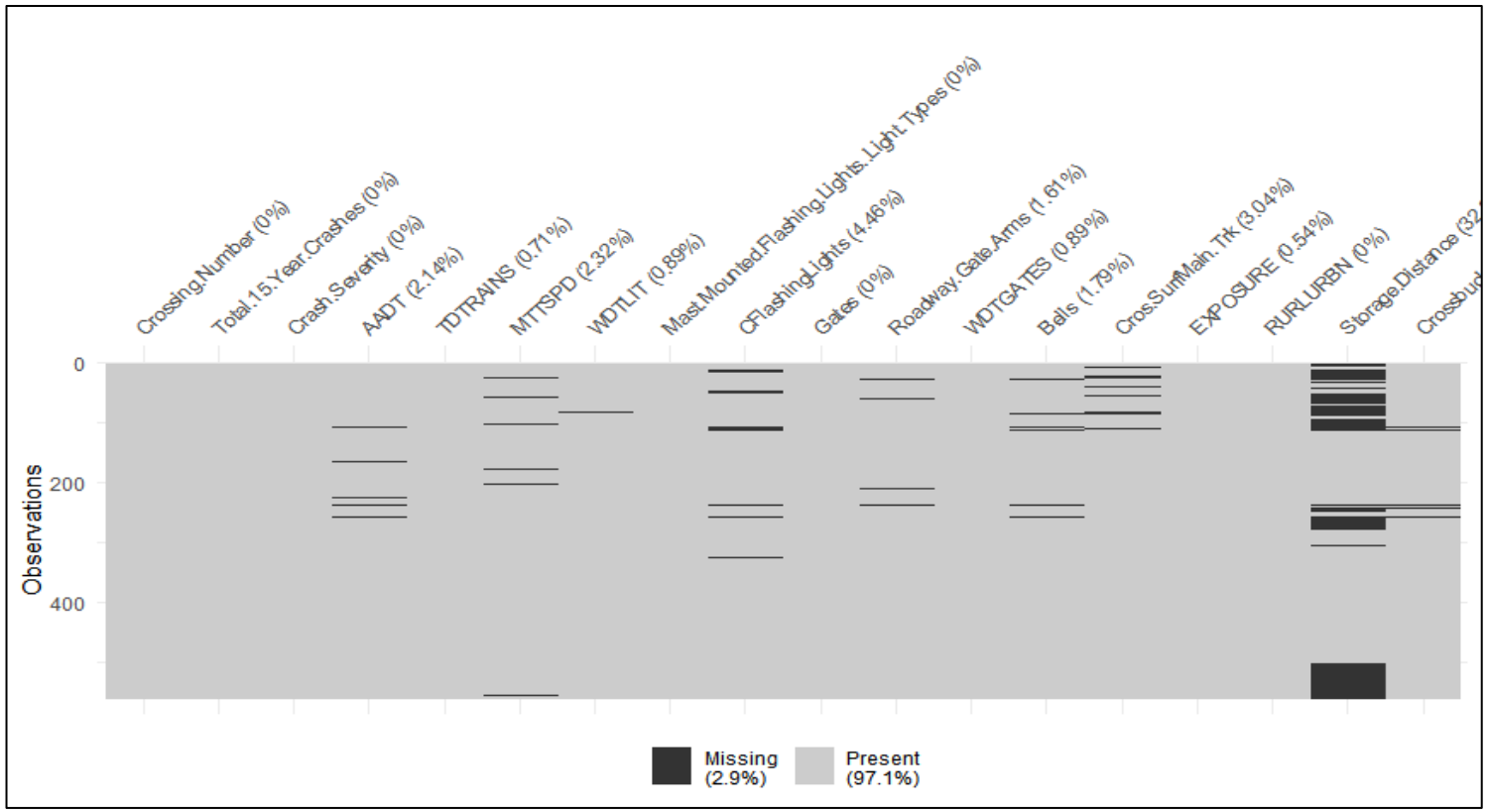

Figure 3.11 Heat Map of Missing Values in Candidate Variables of FRA HRGCs Inventory (N=560)

Individual heat map of 560 HRGCs inventory data showed that important candidate variables such as AADT, maximum timetable speed, warning device type light, count of flashing lights, gate arms, bells, surface type of main track, cross bucks and storage distance all had missing values. The highest percentage of missing values was observed in storage distance, bells, and crossing surface type of main track (**Figure 3.11**)

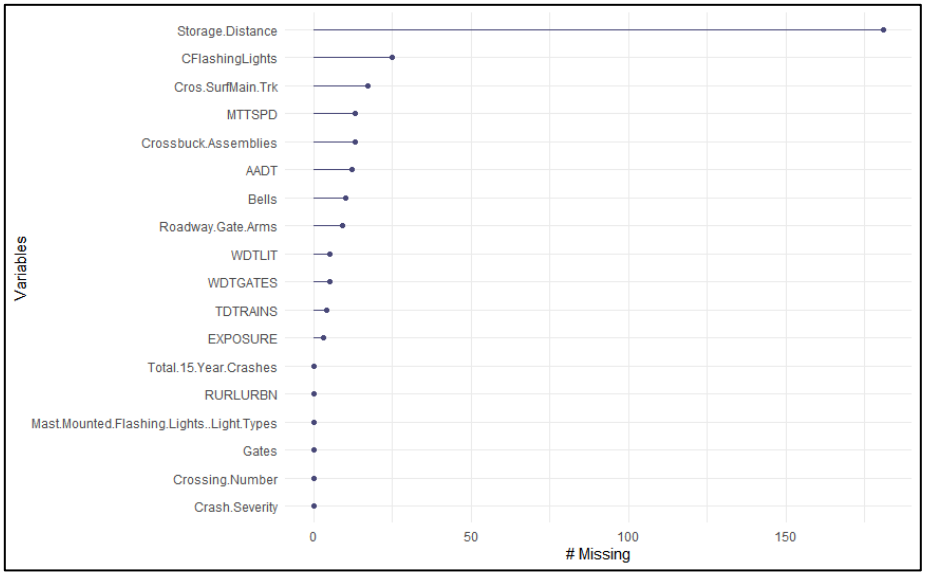

Figure 3.12 Number of Missing Values for each variable (N=560)

Another depiction of missing values in crash and inventory data for 560 HRGCs in **Figure 3.12** showed that the highest percentage of missing values were observed in storage distance, crossing surface type for main track and crossbuck assemblies. It is seen in the data that though some variables didn't have many missing values such as AADT, highway speed, etc., but they have rather impractical values that are deemed incorrect. For example, observing an AADT value of 1, 2, or 3 in inventory data appears

impractical. Furthermore, Up-Set plot; an alternative to Venn diagram was used to determine whether missing values were consistent among variables of specific rows in the dataset. The Up-Set plot in **Figure 3.13** depicts that there were no consistencies in the missing values among variables of certain rows. As for missing values in "storage distance" indicator, there was no intersection with other sets of missing values for 150 rows(cases). However, the plot showed that in only 4 cases in dataset, there was an intersection in the missing values of identified variables.

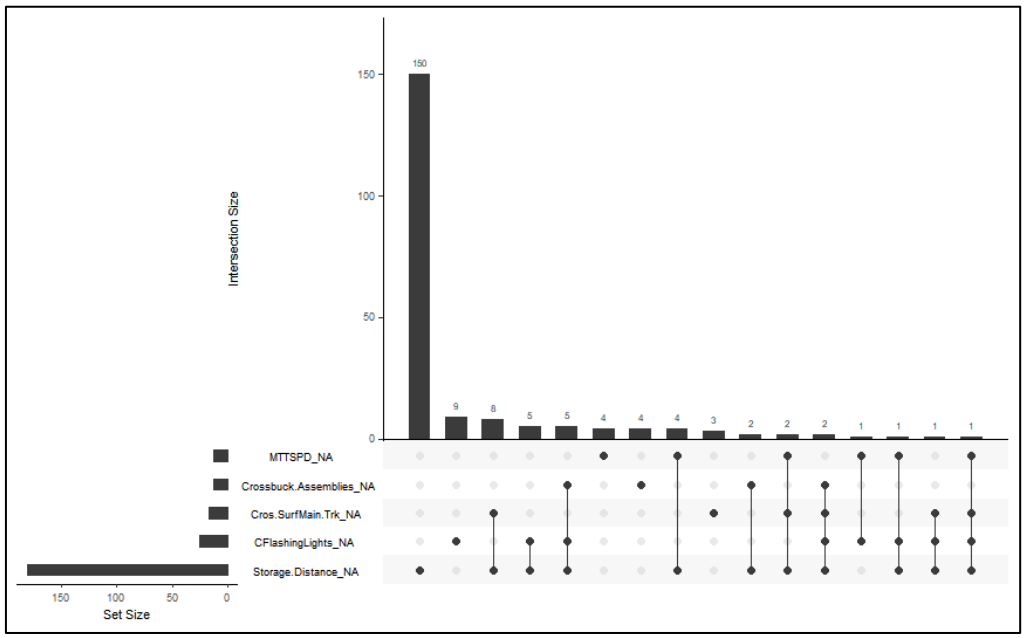

Figure 3.13 UpSet Plot for Missing Values of Variables with the Highest Percentage of Missing Values (N=560)

Furthermore, **Figure 3.14** illustrates density plot for missing values in number of bells, Storage distance and crossbuck assemblies. The figure shows that for most cases of missing values in three variables, there were no reported crashes (2016-2020) at HRGCs.

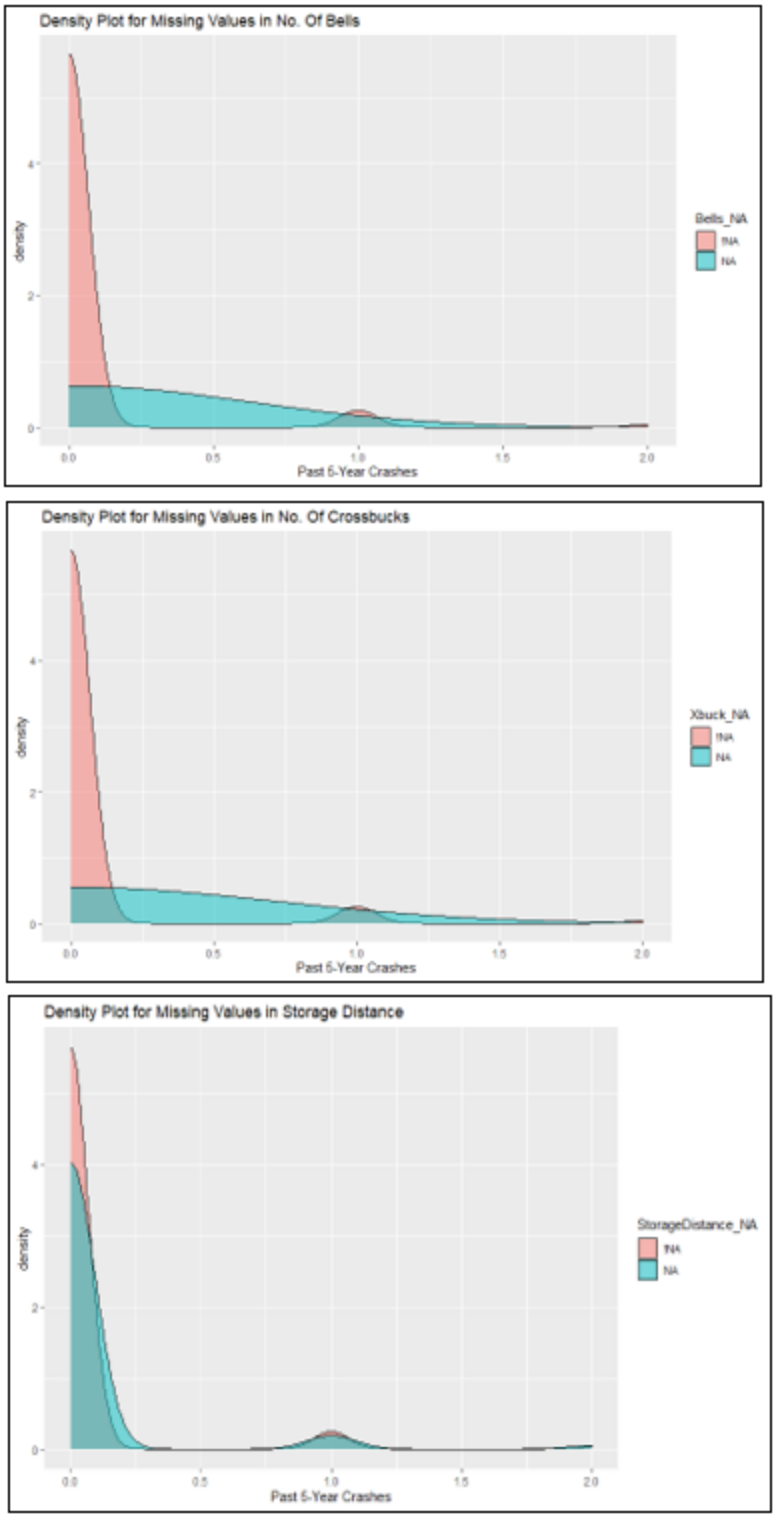

Figure 3.14 Density Plots for Missing Values for Number of Bells, Crossbuck Assemblies, and Storage Distance

The way that data is missing can significantly affect the inferences that can be made about the data and the methods that can be used to analyze it, so it is crucial to determine whether missing values in data occur in a predictable pattern or randomly. If missing values are discovered to be missing at random, it means that neither the missing value itself nor any other variable, observed or unobserved, is related to the probability of a value being missing. Methods like “listwise deletion” or “mean imputation” can be applied in this situation to handle the missing data without introducing bias (Kaiser, 2014).

On the other hand, if missing values are found to be non-random, it means that the probability of a value being missing is related to the missing value or some other variable. This is known as “non-ignorable missingness”. In this case, more advanced methods such as “multiple imputation” or “inverse probability weighting” may be needed to manage the missing data without introducing bias. Checking if missing data is missing at random or not helps to determine the appropriate methods to oversee missing data and avoid potential bias in the analysis and conclusion (Kaiser, 2014).

Based on Little’s missing completely at random (MCAR) test (Little, 1988), the missing data in important candidate variables were evaluated for randomness. The test’s null hypothesis was that the data with missing values were missing completely at random and the test statistic was a chi-squared value. Based on lower p-value, it was concluded that the data was not missing completely at random and followed patterns. The MCAR test (**Table 3.3**) and plots shown previously have helped to justify the need for field validation of the HRGCs inventory data to eliminate missing values.

Table 3.3 MCAR Test for Missing Values in HRGCs Inventory Data

Statistic	Degree of Freedom (df)	P-Value	Missing Patterns
227	97	0.00000	22

The MCAR test performed here also addressed Hypothesis 1, which was initially formulated in this research to determine whether the missing values in the FRA inventory data were missing at random or following a specific pattern.

3.2.3 Descriptive Statistics of Field-validated Data

A total of 53 inventory-related database variables were checked for each field-visited HRGCs, and digital pictures of the HRGCs were obtained. Any incorrect values in the database were corrected according to field conditions, and missing values were added if they were available in the field. **Table 3.4** presents a summary of the corrections and missing value additions for the nine Nebraska counties from field visits. In aggregate, 560 HRGCs were field-investigated and 5 (Approx. 1%) were found to be either abandoned or non-operational, or altogether non-existent. This effort resulted in 2,241 values to be corrected and 1,732 missing values to be added, giving an average of 7.4% of the database values that were changed at each HRGCs.

Table 3.4 Summary of Corrections and Added Missing Values from Field Validation

County	Number of Corrected Values	Number of Missing Values Added	HRGCs Visited	Abandoned/Non-existent	Percent Corrected and Added Missing Values
Lancaster	376	657	112	0 (Private HRGCs removed)	9.2
Cass	307	83	56	0 (Private HRGCs removed)	7.1
Douglas	286	108	67	4	5.9
Gage	115	347	41	0 (Private HRGCs removed)	11.3
Jefferson	174	25	45	0 (Private HRGCs removed)	4.3
Otoe	285	46	77	1	4.2
Saline	119	37	61	0 (Private HRGCs removed)	4.1
Sarpy	144	59	25	0 (Private HRGCs removed)	8.1
Saunders	435	370	76	0 (Private HRGCs removed)	10.6
Total	2241	1732	560	0 (Private HRGCs removed)	7.4

According to **Table 3.5**, 74.11% of crossings in our sample HRGCs data had a single track. Likewise, two-tracked crossings accounted for 18.93% of total crossings. However, there was only one crossing in the data with 6 and 7 total tracks. The table also reveals that most HRGCs were in public spaces (60.71%), had two crossbuck assemblies (90.71%), had no gates (71.07%), and had two lanes crossing the railway track (91.25%).

Table 3.5 Key Features of Field-Validated Inventory Data (N=560)

Variable	Categories	Frequency of Total Tracks	Percentage
Total Tracks	1	415	74.11%
	2	106	18.93%
	3	27	4.82%
	4	4	0.71%
	5	1	0.18%
	7	1	0.18%
	6	1	0.18%
	Not Operational/Not Accessible	5	0.89%
Land Use Type	Open Space	340	60.71%
	Commercial	84	15.00%
	Industrial	70	12.50%
	Residential	58	10.36%
	Institutional	3	0.54%
	Not Operational/Not Accessible	5	0.89%
Crossbuck Assemblies	No Crossbuck Assemblies	7	1.25%
	1	8	1.43%
	2	508	90.71%
	3	11	1.96%
	4	17	3.04%
	6	2	0.36%
	8	1	0.18%
Roadway Gate Arms	0	398	71.07%
	2	150	26.79%
	3	1	0.18%
	4	4	0.71%
	5	1	0.18%
	8	1	0.18%
	Not Operational/Not Accessible	5	0.89%
Number of Lanes	1	13	2.32%
	2	512	91.42%
	3	6	1.07%
	4	16	2.86%
	5	4	0.71%
	6	2	0.36%
	8	2	0.36%
	Not Operational/Not Accessible	5	0.89%

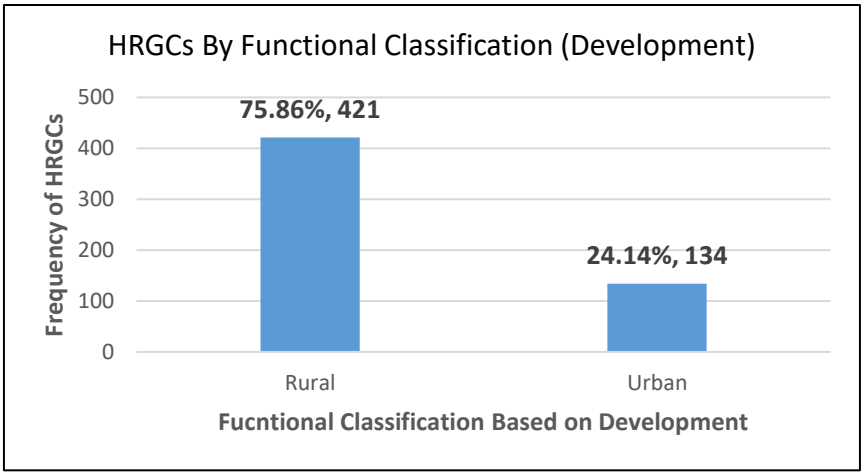

Figure 3.15 Distribution of HRGCs by Functional Classification

Data analytics of the field validated data also revealed that 75.86% of HRGCs were in rural areas however, 24.14% of HRGCs were in urban areas (**Figure 3.15**). Furthermore, functional classification based on road function indicated that majority of the crossings were at local access (67.93%). However, major, and minor collectors accounted for 19.03% and 4.32% of HRGCs in the field-validate data (**Figure 3.16**).

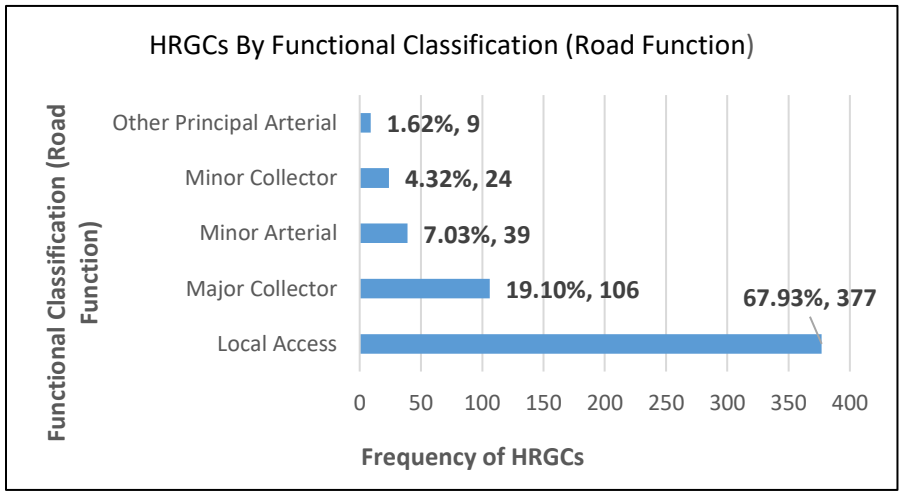

Figure 3.16 Distribution of HRGCs by Functional Classification (Road Function)

Figure 3.17 presents the distribution of the HRGCs by posted speed limits. The figure indicates that the highest percentage of HRGCs had posted speed limit of 50 mph on the roads intersecting the subject rail crossings. However, 55 mph and 25 mph speed limits were posted on roads intersecting 3.24% and 27.93% of the total HRGCs in field-validate inventory data.

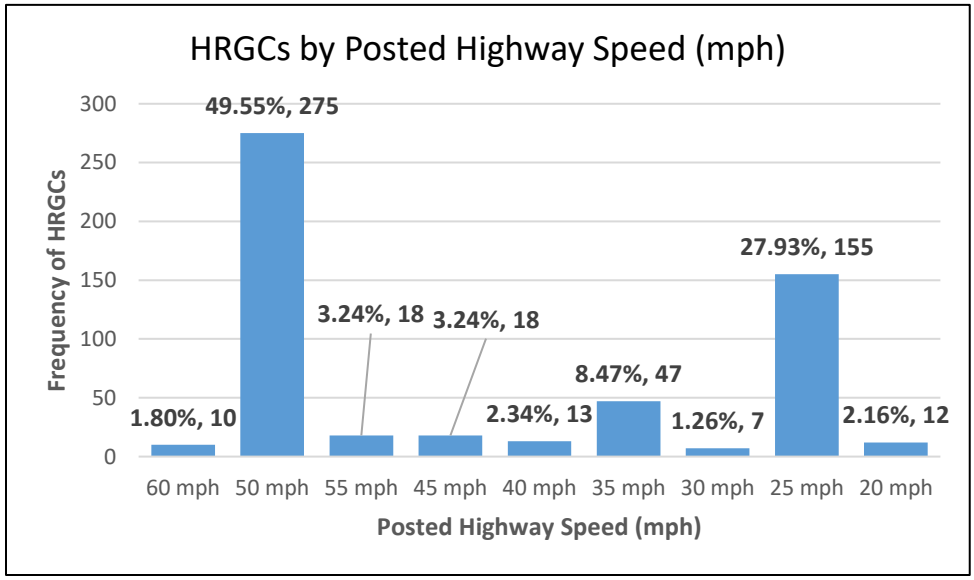

Figure 3.17 Distribution of HRGCs by Highway Speed (mph)

Figure 3.18 indicates that majority of the HRGCs had gravel as an approach surface type (49.55%) However, brick approach surface type was only in 3 cases of HRGCs in the data.

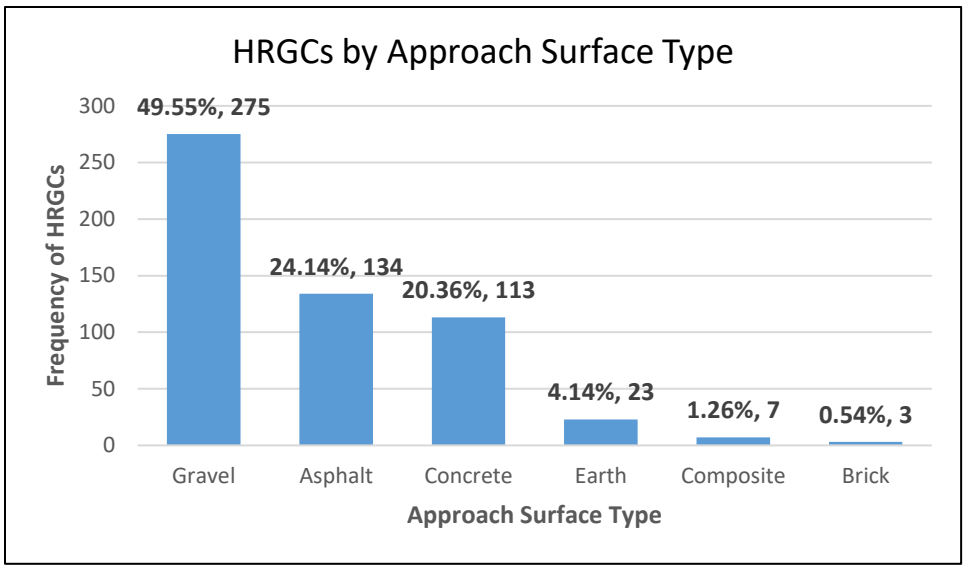

Figure 3.18 Distribution of HRGCs by Approach Surface Types

Figure 3.19 shows a histogram plot demonstrating the distribution of the studied HRGCs by natural logarithmic values of AADT. Based on the figure, it can be observed that the maximum and minimum values of AADT are around 39,000 vehicles per day and one vehicle per day, respectively. The average AADT for all considering crossings is approximately 1352 vehicles per day.

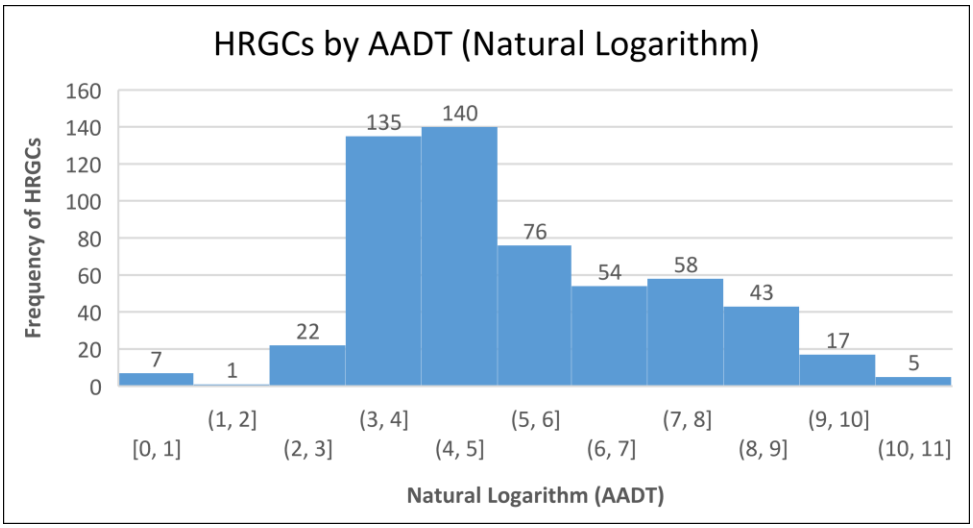

Figure 3.19 Distribution of HRGCs by AADT (Natural Logarithm)

3.3 Chapter Summary

This chapter presented the data collection and assessment process performed in this research. Three datasets were utilized: the FRA crash database on HRGCs, the FRA inventory database on HRGCs, and the field-validated inventory database on HRGCs. To conduct the analysis, crash history, and inventory records of 560 HRGCs across nine counties in Nebraska were extracted from the publicly available FRA website. Additionally, the inventory data of these 560 HRGCs were field validated for comparative analysis. Furthermore, the chapter discussed in detail the FRA Form 6180.57, which contained crash information on HRGCs. In addition, the inventory data collection process by the FRA was also described in detail, with a discussion on the aspects of inventory covered in Form 6180.71. Furthermore, various aspects of field validation and manual inspection of HRGCs inventory data were discussed. It is noteworthy that for this research, public, at-grade, and operational grade crossing were considered. The chapter also explained the process involved in crash and inventory data integration to complete the planned tasks of the study.

After giving insights about data collection and data integration, this chapter delved into thorough assessment of data, the process of verification of HRGCs inventory, identification of missing information, and data validation checks by using data visualization techniques such as heat maps, density and Upset plots etc. The initial data validation of the inventory aspect of all public and at-grade crossings in Nebraska highlighted several illogical errors. For instance, the functional classification of 53 HRGCs were reported as “Rural,” but at the same time, they were marked as “In city” in

the “Nearest” variable. Density plots for missing values for the number of bells, Crossbuck assemblies, and storage distance provided further information on the missing details in the inventory dataset.

Further investigation was conducted to determine whether the missing values in the data occurred in a predictable pattern or randomly (Hypothesis 1 of this research). The MCAR test showed that the data were not missing at random and followed a set pattern. In addition, the descriptive statistics for validated inventory revealed that in the dataset, 5 (roughly 1%) HRGCs were found abandoned, non-operational, or altogether non-existent. In the field-validation process, a total of 2,241 values were corrected and 1,732 missing values were added, resulting in an average of 7.4% of the database values being changed at each HRGC. Field-validation of the inventory also revealed that 74.11% of crossings in the study sample had a single track, 60.71% HRGCs were in public spaces, 71.07% had no gates, and 49.55% of the HRGCs had gravel as an approach surface type.

The data description and preliminary data assessment undertaken in this chapter have proven to be beneficial, providing a comprehensive understanding of the subject data and laying the groundwork for devising effective strategies for crash and severity predictions, as well as model estimation based on the data at hand. Through this process, important insights into the nature of the data were gained, enabling more informed decisions about the appropriate analytical techniques and tools to employ in future analyses.

CHAPTER 4 PREDICTING CRASH FREQUENCY AND SEVERITY WITH 2020 APS MODEL: FIELD-VALIDATED VS FRA INVENTORY DATA

4.1 Background

Accurate inventory data of highway-rail grade crossings (HRGCs) is critical for the correct prediction of crashes and their severity by using the 2020 Federal Railroad Administration (FRA) crash prediction model. The model is utilized for resource allocation and strategical planning to limit potential causes of crashes at grade crossings. The FRA relies on this information to make informed decisions and implement effective measures to improve safety at HRGCs. However, if the inventory data are not accurately verified to reflect the actual physical attributes of the crossings, the crash and severity prediction models will produce incorrect results.

4.2 Criteria for Comparison

In this pursuit, the original FRA HRGCs inventory data and the field-validated inventory data (both merged with the FRA HRGCs Accident/Incident database) for the 560 HRGCs were used to assess the effects of erroneous or missing information on crash prediction utilizing the FRA's 2020 APS model (Equation 1-14).

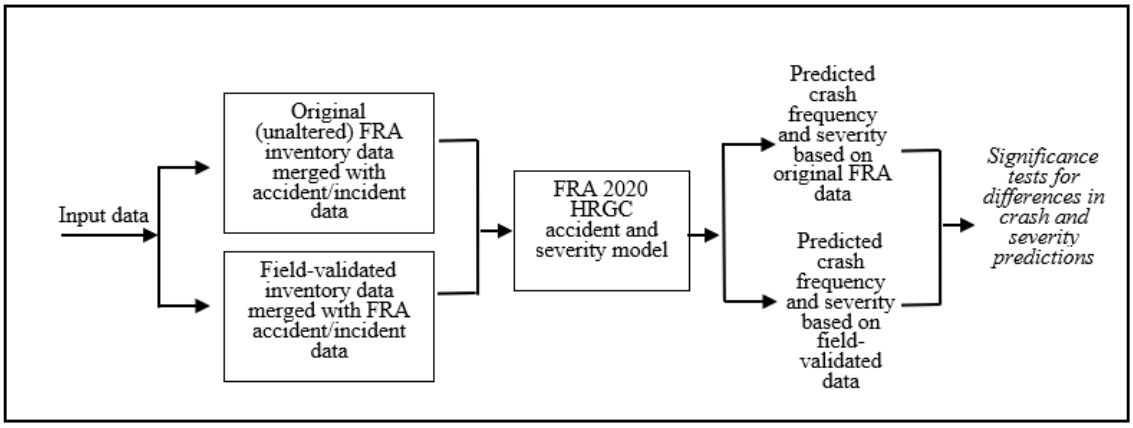

Figure 4.1 Approach for Crash Prediction and Severity Comparison Utilizing the FRA's 2020 APS Model

The two sets of crash predictions (expected crashes) and crash severity were then statistically evaluated for differences. **Figure 4.1** illustrates the approach utilized for crash count and severity prediction comparison.

4.3 Results and Interpretation of the Comparison

To fulfill the study's objectives, crash data were retrieved from the FRA crash database for the past five years. Furthermore, FRA data were incorporated into the FRA's 2020 crash prediction model and plotting the results. The same process was carried out using field validated data. The plots indicated a noticeable difference in crash predictions based on the two datasets.

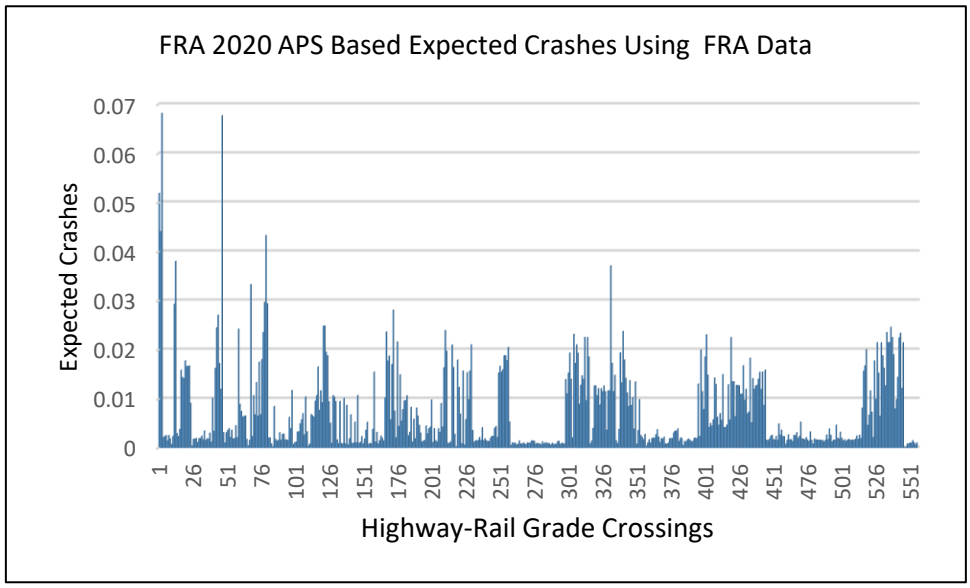

Figure 4.2 Expected Crashes Based on 2020 APS Model (FRA Data)

To better understand the extent of these differences, the percentage differences were calculated and plotted. As seen in **Figure 4.2-4.3**, there is a visible difference between expected crash values derived from the two inventory datasets.

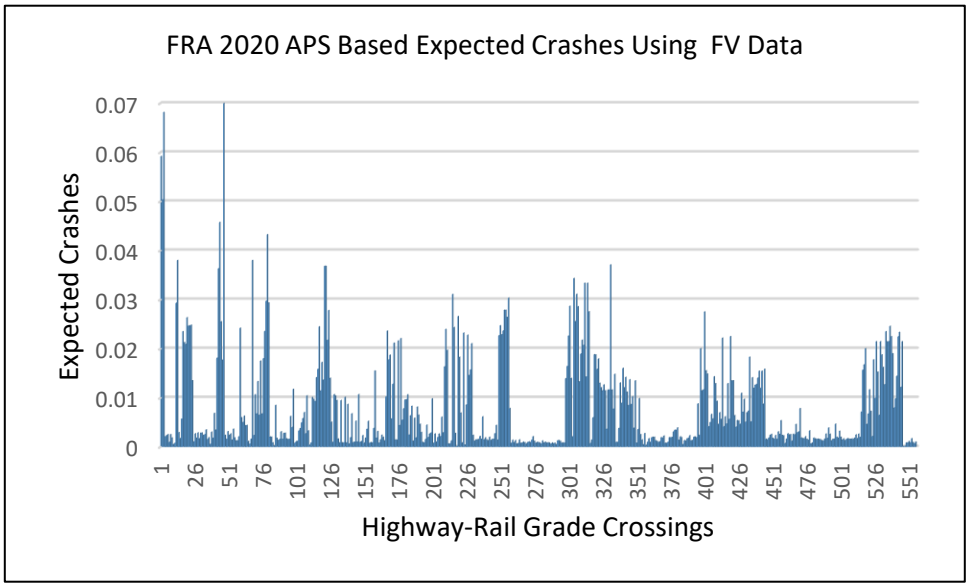

Figure 4.3 Expected Crashes Based on 2020 APS Model (FV Data)

Figure 4.4 displays the percentage differences between the FRA and Field Validated (FV) data crash predictions. It is evident from the data that the range of percentage differences varies from 0 to 120%, indicating that the expected crashes could be either approximately double or half the actual crash count, highlighting the issue of data inaccuracies. These findings reveal a significant discrepancy between the crash predictions estimated using FRA and field-validated datasets.

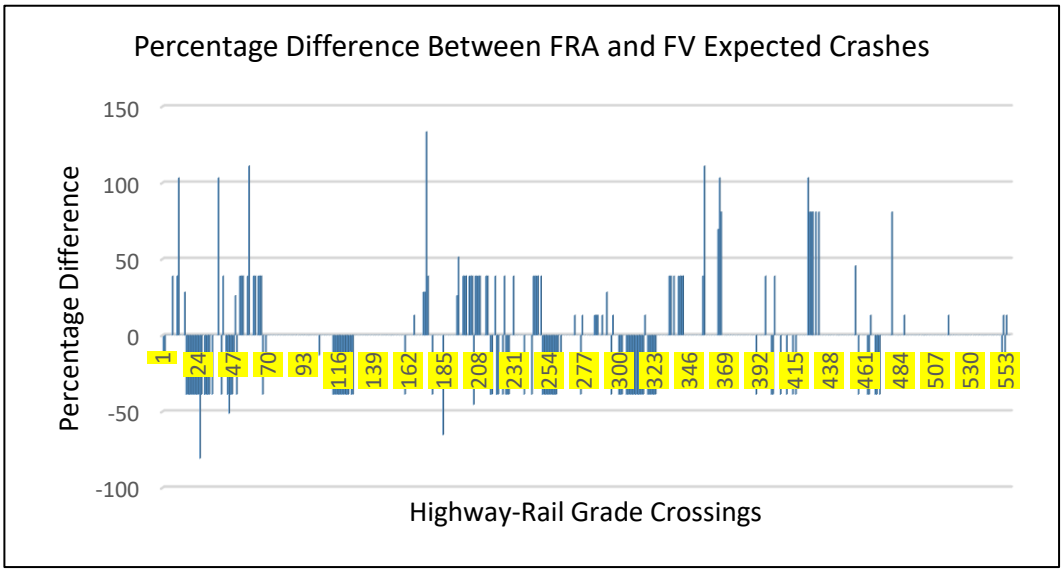

Figure 4.4 Percentage Difference Between FRA and FV Expected Crashes

Table 4.1 presents the absolute values of percentage difference between FRA and Field-validated expected crashes. It can be observed that, based on the two datasets, the majority of HRGCs exhibited expected crashes within a 0-20% percentage difference. However, a higher percentage difference in expected crashes was only observed in seven HRGCs.

Table 4.1 Percentage Difference Between FRA and FV Data Based Expected Crashes

Percentage Difference in Expected Crashes	Number of HRGCs
0-20%	390
20-40%	144
40-60%	4
60-80%	2
80-100 %	8
100-120 %	7
Total	555

To determine the normality of both estimated predictions (based on FRA and field-validated data), a Q-Q plot was utilized. This plot is a graphical representation that compares the quantiles of the expected crashes with the quantiles of a standard normal distribution. When the data are normally distributed, the points on the Q-Q plot should form a roughly straight line. As depicted in **Figure 4.5**, the data appear to weakly align with a straight line, indicating some normality in the distribution.

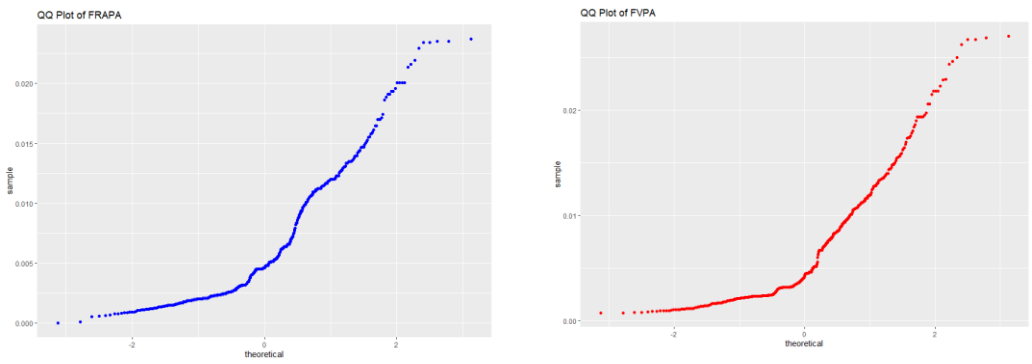

Figure 4.5 Q-Q plots for checking normality of expected crashes (FRA vs FV)

The Q-Q plots used to visualize the distribution of predicted crashes based on two datasets were found to be unclear. Therefore, histograms were developed to better

illustrate the distribution of these data. **Figure 4.6** provides a clear representation of the fact that both datasets do not exhibit a normal distribution.

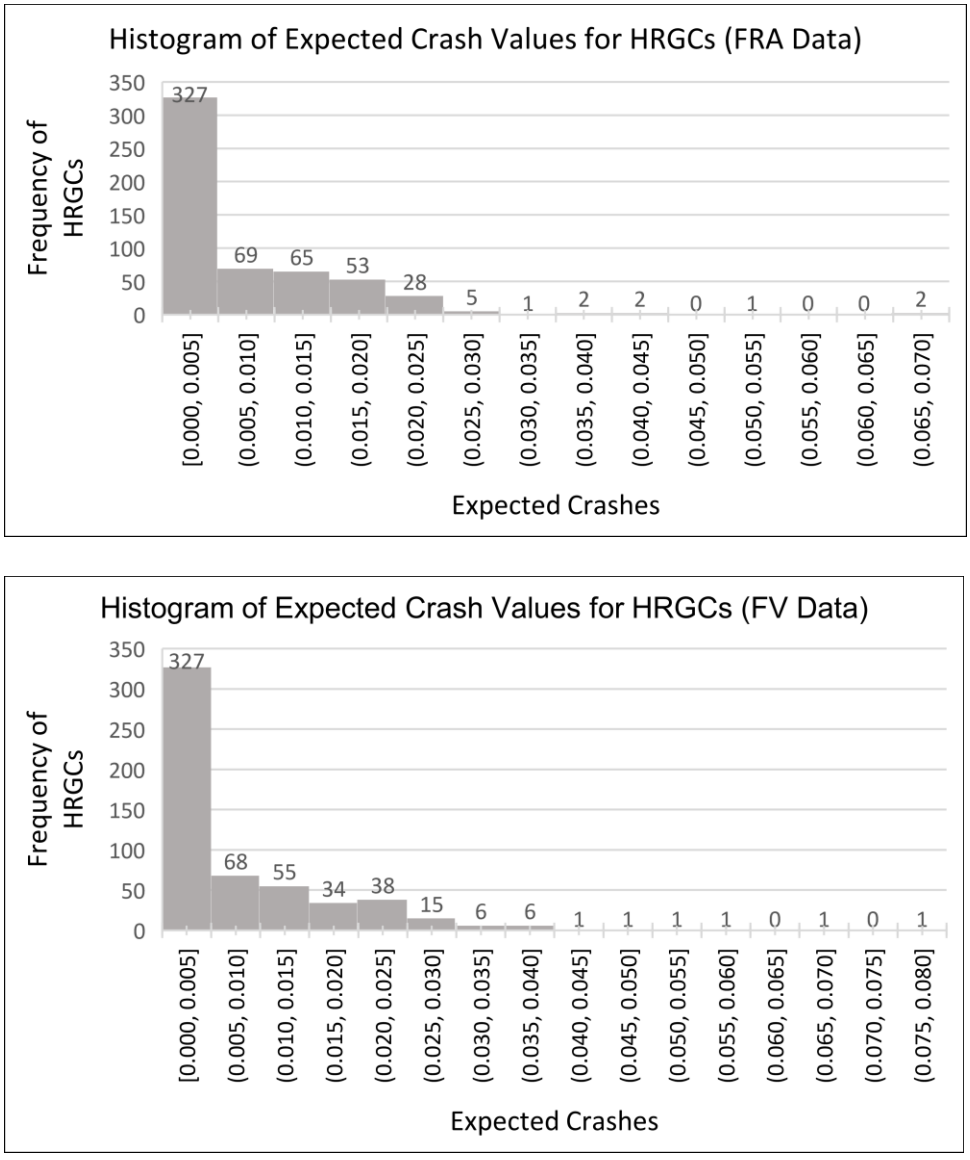

Figure 4.6 Histograms for Expected Crashes (FRA vs FV)

However, the predicted crashes based on FRA data showed a slight inclination toward a normal distribution, whereas the field-validated inventory data did not appear to be normally distributed. To further analyze the normality of the predicted crashes based

on FRA and field validated data, the Shapiro-Wilk test was utilized. This test provides a clear and formal way to determine whether or not the data are normally distributed (Yap and Sim, 2011). When conducting a Shapiro-Wilk test, a p-value less than 0.05 indicates that the data in at least one of the data columns (expected crashes) is not normally distributed. In such cases, a non-parametric statistical test is a good option to compare differences between the two datasets, such as the Wilcoxon rank-sum test (also known as the Mann-Whitney U test) or the Kruskal-Wallis test (Yap and Sim, 2011). Unlike t-tests or ANOVA, these tests do not assume normality in the data and are often used as alternatives when normality assumptions are not met.

Table 4.2 Shapiro-Wilk Test Normality Test for Expected Crashes

Variable	**W**	**P-value**
Original FRA data crash prediction	0.87519	2.2e-16
Field-validated data crash prediction	0.84692	2.2e-16

The null hypothesis of the Shapiro-Wilk normality test (**Table 4.2**) was that the data were normally distributed. The test output showed the test statistic “W” and the corresponding “p-value.” For expected crashes based on two datasets, the “p-value” was less than 0.05, indicating convincing evidence against the null hypothesis. This meant that the data were unlikely to be normally distributed. The values of the test statistic “W” for both columns (expected crashes) were between 0 and 1. The closer “W” was to 1, the more normal the data were. Both columns had “W” values less than 1, which suggested that the data were not very normal. In summary, the results of the Shapiro-Wilk normality

test suggested that expected crashes based on FRA and field-validated inventory data were not normally distributed.

Due to the absence of normality in the predicted crashes, it was determined that a non-parametric statistical test would be the most suitable for comparing the differences between the two datasets. Accordingly, the Wilcoxon rank-sum test, also known as the Mann-Whitney U test, was employed in the study, and the corresponding findings are presented in **Table 4.3**. The Wilcoxon Rank Sum Test is often described as the non-parametric version of the two-sample t-test. This test addressed Hypothesis 2 of this research.

Table 4.3 Wilcoxon Rank Sum Test for Differences in Expected Crashes

Variables	**W**	**P-Value**
Original FRA Vs FV data crash predictions	122640	3.622e-10
Alternative hypothesis	*True location shift is not equal to 0*	

The results of the test revealed a “p-value” of less than 0.05 and test statistic “W” of 122640; saying that there is convincing evidence to suggest that the medians of the two distributions differ. The alternative hypothesis, which was stated as "the true location shift is not equal to 0," indicates that one population's distribution is shifted either to the left or right of the other, thus implying different medians. Since this was a non-parametric test, parameters such as the mean were not estimated. Instead, the test was performed solely to find evidence that one distribution was shifted to the left or right of the other.

Furthermore, inventory and crash data (2016-2020) for 560 HRGCs were utilized, revealing that only 28 of the selected grade crossings experienced at least one crash during this period, resulting in a total of 34 crashes. To predict the probability of a crash resulting in fatalities, injuries, or property damage, the 2020 APS model was employed, using Equation 1-14 to calculate accident severity prediction probabilities, and Hypothesis 3 of the research was tested. The observed ordinal variable representing the type of crash severity (PDO, injury, or fatality) was utilized as the dependent variable for the model.

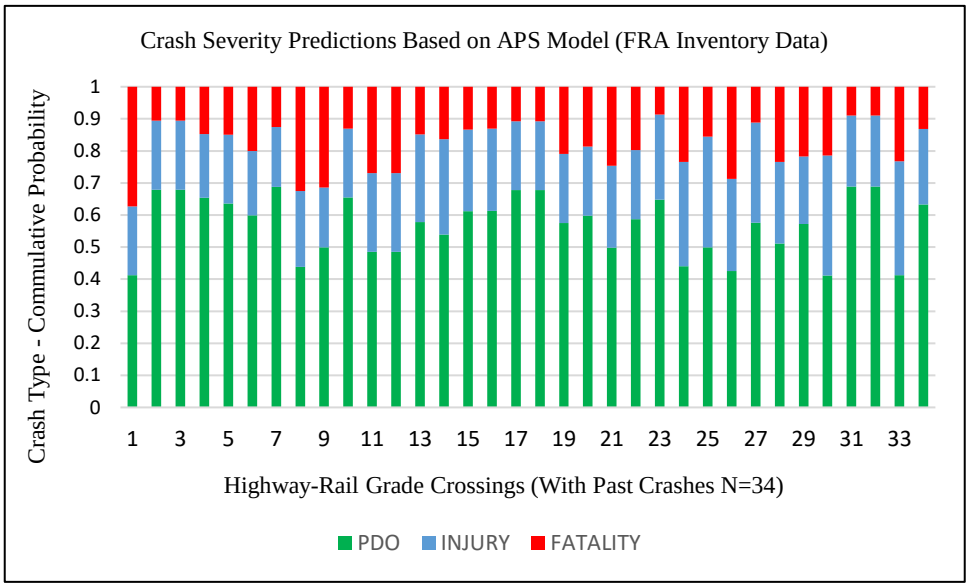

Figure 4.7 Crash Severity Predictions Based on 2020 APS Model (FRA Data)

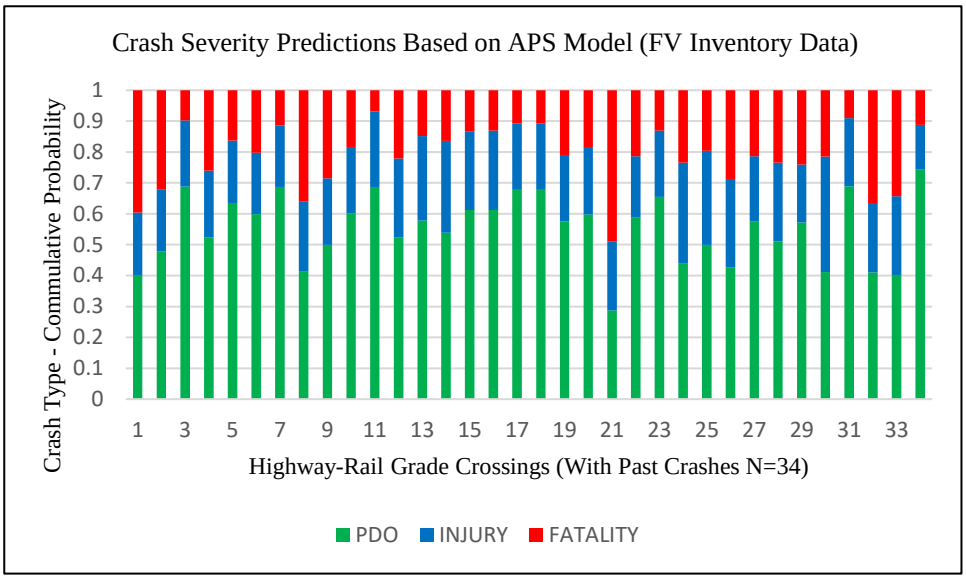

Figure 4.8 Crash Severity Predictions Based on 2020 APS Model (FV Data)

Two datasets were employed to run the model, and the findings indicated that erroneous data on HRGCs' inventory can impact crash severity prediction. **Figures 4.7-4.8** present evidence of the observed disparities in the cumulative probabilities of three types of crash severity (i.e., PDO, injury, and fatality) when using the two datasets.

Additionally, a comparative analysis revealed that while the disparity between FRA and field-validated crash severity predictions is not drastic, it is noticeable and implies that using incorrect HRGCs inventory datasets can lead to errors in severity prediction modeling.

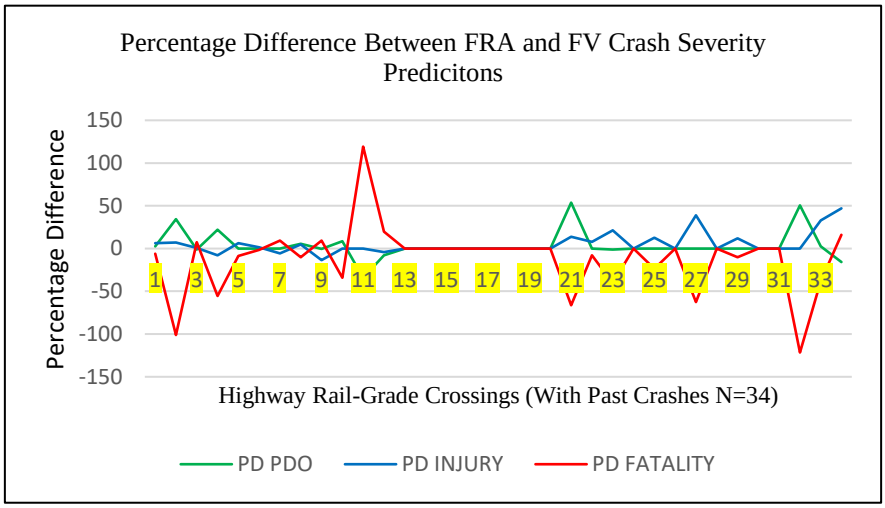

Figure 4.9 Percentage Difference Between FRA and FV Crash Severity Predictions

Figure 4.9 illustrates percentage differences between crash severity predictions based on FRA and field-validated data.

To investigate if probabilities of PDO, injury and fatal crashes estimated by using the FRA and field validated data on 2020 FRA Accident Severity Model were statistically significantly different, Wilcoxon Rank Sum Tests were performed which indicated statistically significant differences between the probabilities of PDO, injury and fatal crashes. This analysis addressed Hypothesis 3 of this research (**Table 4.4**).

Table 4.4 Wilcoxon Rank Sum Test for Differences in Crash Severity Prediction

Variables	**W**	**P-Value**
Prob. of PDO Crash	623	0.00050
Prob. of Injury Crash	675.5	0.00231
Prob. of Fatal Crash	488.5	0.00274
Alternative hypothesis	*True location shift is not equal to 0*	

4.4 Chapter Summary

This chapter focused on analyzing how data inaccuracies and missing information in inventory data of HRGCs can affect crash frequency and crash severity prediction. For this part of the research, two datasets (i.e., FRA unaltered HRGCs inventory data and field-validated HRGCs inventory data) were employed on the 2020 Accident Prediction and Severity Model (APS). In addition, the two datasets used for the analysis were also merged with the FRA HRGCs Accident/Incident database. The two sets of crash predictions (expected value of crashes) and crash severity were then statistically evaluated for differences. Furthermore, FRA data were incorporated into the FRA's 2020 crash prediction model and results were plotted. The same process was carried out using field-validated data. The plots indicated a noticeable difference in expected crashes based on the two datasets. The analysis showed that there was a significant discrepancy between the expected crash values derived from the two inventory datasets. Furthermore, the percentage differences between the FRA and Field Validated (FV) data's expected crashes were estimated. From the data it was revealed that the range of percentage differences varies from 0 to 120%, indicating that the predicted crash risk could be either about double or half the actual crash counts. To understand, if the differences in crash predictions were statistically significant, the normality of both estimated predictions (based on FRA and field-validated data) were estimated by developing Q-Q plots. Based on data visualization, histograms were developed and Shapiro-Wilk test for normality was estimated. Furthermore, Wilcoxon rank-sum test was performed which is a non-parametric statistical test, that statistically compared differences between the two

expected crashes. The results of the test revealed robust evidence to suggest that the medians of the two distributions of crash predictions differ, implying differences in predictions, which addressed Hypothesis 2 of this research.

Furthermore, crash severity predictions based on the FRA's 2020 crash and severity model were also estimated by using the FRA and field-validated dataset. As, 2020 APS was based on past 5-year crashes, the study utilized inventory data and crash data from 2016 to 2020 for 560 HRGCs, revealing that only 28 of the selected grade crossings experienced at least one crash during this period, resulting in a total of 34 crashes. The findings indicated that erroneous data on HRGCs' inventory can substantially impact crash severity prediction. Additionally, a comparative analysis revealed that while the disparity between FRA and field-validated crash severity predictions was not drastic, it was noticeable and implied that using incorrect HRGCs inventory datasets can lead to errors in severity prediction modeling. For crash severity predictions, Wilcoxon Rank Sum Tests were performed that revealed statistically significantly different probabilities for PDO, injury and fatal crashes when using FRA and filed validated datasets.

CHAPTER 5 CHAPTER 5 DATA ERRORS AND THEIR IMPACTS ON CRASH/SEVERITY MODEL ESTIMATION

This chapter is dedicated to examining the impact of errors and missing information in the FRA's HRGCs inventory data on estimated crash frequency and severity model parameters. For this purpose, the original (unaltered) FRA inventory dataset and field verified dataset for 560 HRGCs were used. Identical parameters were utilized during model estimation of comparative crash frequency models in order to conduct a more comprehensive analysis of how one parameter's coefficient differentiated from others and to what extent.

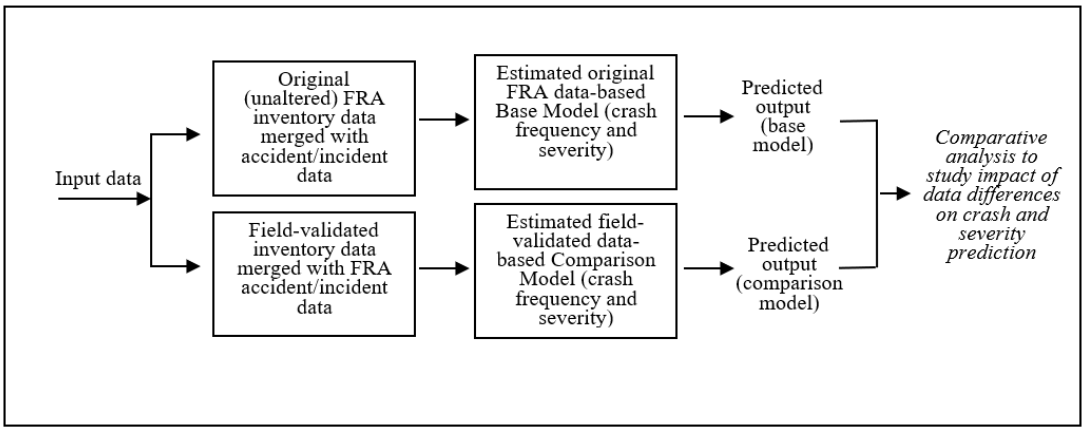

Figure 5.1 Approach Followed for Crash Prediction Models Estimation

Furthermore, two crash severity models were developed with identical parameters. Crash data from the past 15 years (2007-2021), as well as FRA and field validated HRGCs inventory data, were incorporated into the models. Only crashes that had previously been reported on the 560 studied HRGCs were considered for this analysis. **Figure 5.1** illustrates the approach followed for crash frequency and severity prediction model estimation based on the two datasets. A base model was first estimated by utilizing the original FRA inventory data and then another comparison model was estimated using the field-validated inventory data.

5.1 Crash Frequency Modelling

Two zero-inflated negative binomial (ZINB) models were estimated to study how data completeness and accuracy affect the parameters of a crash frequency models based on field-validated and unaltered inventory datasets. The main rationale for using the ZINB model instead of other count models is that the response variable (crashes at HRGCs) exhibited significant over-dispersion (variance > mean) with many zeros. The zero-inflated model is designed to manage response variables that have more zeros than one would anticipate in a typical count data scenario, or a significant percentage of zero values. **Figure 5.2** presents a few attributes, limitation and assumptions of ZINB model.

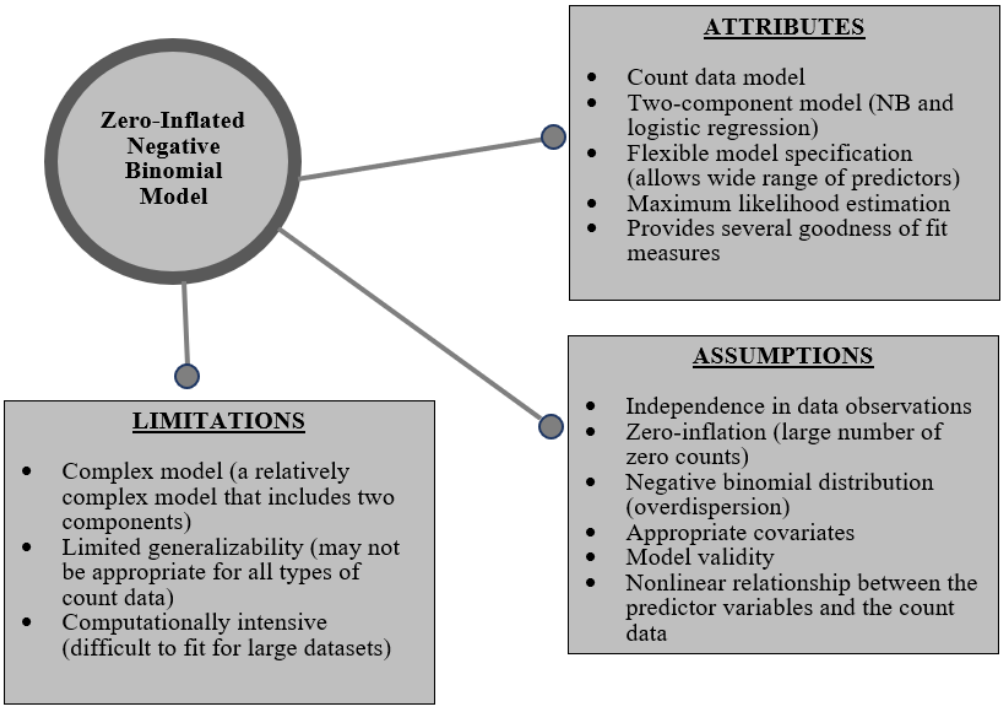

Figure 5.2 Key attributes, assumptions and limitations of ZINB model

Transportation safety analysts often prefer using zero-inflated (ZI) models for analyzing crash data, as these models have been shown to provide better statistical fit compared to traditional Poisson and Negative Binomial (NB) models. By explicitly accounting for the excess zeros commonly observed in crash data, ZI models can provide more accurate and reliable estimates of the underlying crash frequency and help identify the factors that contribute to crashes on the surface transportation network (Sharma and Landge, 2013).

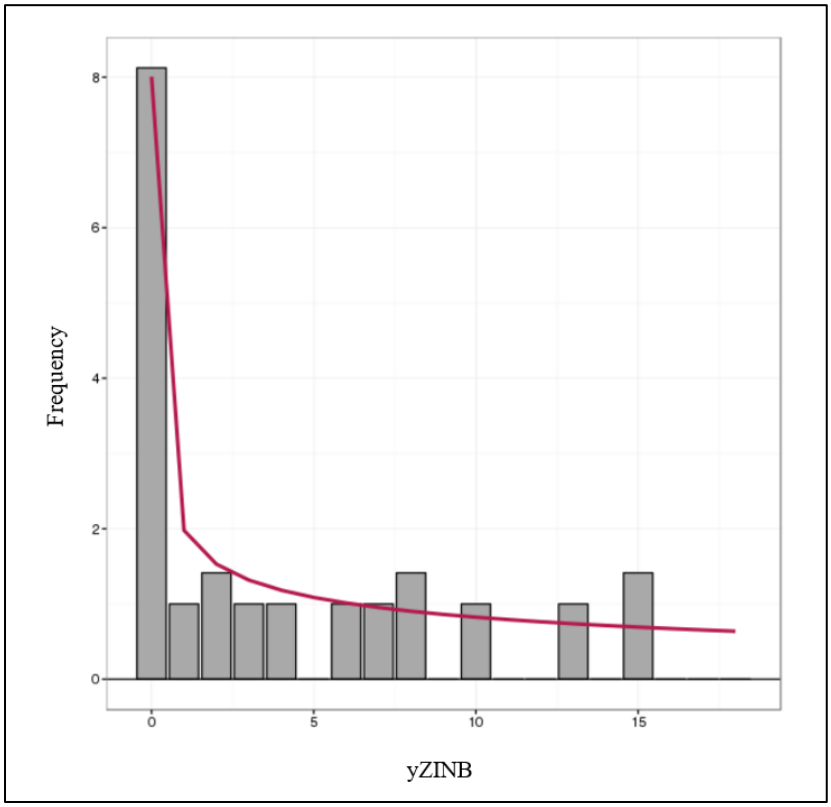

Figure 5.3 Simulated Zero-inflated Negative Binomial Distribution

A ZINB model assumes that zero outcomes are the result of two distinct processes. For example, in the case of crashes at HRGCs, the two processes are: (1) the occurrence of a crash at HRGCs, and (2) no crash at HRGCs. If there was no crash at HRGCs, the only possible outcome would be zero. If there was a crash at HRGCs, it was treated as a count process. The zero-inflated model consists of two parts: a binary model, usually a logit model, which models which of the two processes the zero outcome is associated with, and a count model, which in this case is a negative binomial model, used to model the count process for crashes at HRGCs. The expected count is expressed as a combination of the two processes. It is worth noting that the ZI-Poisson model is similar to the ZINB model, but the former assumes that the non-zero counts follow a Poisson

distribution, while the latter assumes that they follow a Negative Binomial distribution (Brod et al., 2020).

The general formula for ZINB model according to Miaou (1994) is presented as:

$$y_i = 0,1,2 \ldots \ldots \text{ with probability } \frac{\Gamma\left(\frac{1}{\alpha}+y_i\right)}{\Gamma\left(\frac{1}{\alpha}\right)\Gamma(y_i+1)}\left(\frac{1}{1+\alpha*\lambda_i}\right)^{\frac{1}{\alpha}}\left(\frac{\alpha*\lambda_i}{1+\alpha*\lambda_i}\right)^{y_i} \quad \text{Eq (15)}$$

$$\lambda_i = e^{\beta_i x_i} \quad \text{Eq (16)}$$

where x_i is i^{th} independent variable, and β_i is the coefficient of regression

For Zero Inflated Negative Binomial regression

$$y_i = 0, \text{ with probability } p_0 + \left(\frac{1}{1+\alpha*\lambda_i}\right)^{\frac{1}{\alpha}} \quad \text{Eq (17)}$$

where p_0 illustrates the probability model that includes the effects of independent variables, such as logit model.

$$p_0 = \frac{e^{r'w_i}}{1+e^{r'w_i}} \quad \text{Eq (18)}$$

r is the matrix's coefficient and w_i is the i^{th} independent variable. Furthermore, Γ(.) is Gamma function; and α represents the rate of over dispersion.

Maximum likelihood estimation (MLE) is a widely used method for estimating parameters in Poisson, Negative Binomial, and Zero-Inflated regression models (Dong et al., 2014; Raihan et al., 2019). This method involves finding the parameter values that maximize the likelihood function, which measures the probability of the observed data given the model. The MLE method is favored because it has been demonstrated to be

effective for a variety of statistical models and offers precise and effective estimates of the model parameters (Myung, 2003).

To evaluate the performance of the ZINB models, the Akaike Information Criterion (AIC) and BIC (Bayesian Information Criterion) are commonly used. The AIC is a measure of the quality of a model, considering both the goodness of fit and the complexity of the model. A lower AIC value indicates a better model fit, as it penalizes models with a larger number of parameters. However, the idea behind BIC is that the best model is the one that maximizes the likelihood of the data while penalizing for the number of parameters in the model. Therefore, AIC and BIC are often used to compare varied models and select the one that best fits the data (Bozdogan,1987).

5.1.1 Candidate Variables

Table 5.1 displays the variables present in both the FRA and field-validated inventory datasets that were deemed suitable for explaining the factors contributing to HRGCs crashes. These variables have been classified based on their nature as either discrete or continuous.

Table 5.1 Candidate Variables for Inclusion in Crash Prediction Model

Discrete Variables	**Continuous Variables**
Flashing lights	Number of bells
Gates	Number of crossbucks
Rural/Urban classification	Number of total tracks
Pavement markings	AADT
Crossing surface type	Total daily trains
Crossing Illuminated	Maximum timetable speed
Crossing angle	Number of traffic lanes

It is not necessary to compute Tolerance (TOL) or Variance Inflation Factor (VIF) when fitting a ZINB model. In linear regression models, these metrics detect multicollinearity. To ensure the stability and dependability of the model results, it may be helpful to assess the presence of multicollinearity in the data given that ZINB models frequently include several continuous predictor variables (Park et al., 2018; Chatterjee and Hadi, 2006). Collinearity between the predictor variables can result in unstable parameter estimates and inflated standard errors, even though the ZINB model can handle excess zeros and overdispersion. Therefore, it is crucial to look for multicollinearity and lessen its effects by removing or changing variables. Multicollinearity can be addressed to produce more accurate and reliable estimates of the model parameters, which enhances the ZINB model's interpretability and generalizability.

To counter multicollinearity among independent variables, those continuous variables were excluded that had VIF > 10 and TOL < 0.1 (Farooq and Ahmad, 2017). **Table 5.2** highlights the values of VIF and Tolerance estimated for the candidate variables. All candidate continuous variables that are not found to be collinear are

included in model estimation. As the base-crash prediction model is for FRA data, these multicollinearity diagnosis indices are estimated for variables in FRA data.

Table 5.2 Multicollinearity Diagnosis Indices for Candidate Variables (FRA Data)

Candidate Variables	**Coding**	**Tolerance**	**VIF**
Flashing Light Indicator	1 if there are flashing lights at crossing, 0 otherwise	0.625	1.60
Gates Indicator	1 if there are gates at crossing, 0 otherwise	0.633	1.579
Urban Functional Classification Indicator	1 if crossing is in urban area, 0 if crossing is in rural area	0.7153	1.398
AADT	Ln transformed	0.7692	1.300
Number of Bells	Count	1.01317	0.987
Number of Crossbucks	Count	0.82372	1.214
Number of Total Tracks	Count	0.75930	1.317
Number of Traffic Lanes	Count	0.98619	1.014
Total Daily Trains	Ln transformed	0.92764	1.078
Maximum Timetable Speed	Ln transformed	0.91157	1.097
Crossing Surface Type “Concrete” Indicator	1 if crossing surface is concrete, 0 others	0.62101	1.611
Crossing Surface Type “Asphalt” Indicator	1 if crossing surface is asphalting 0 others	0.75930	1.318
Crossing Illuminated Indicator	1 if crossing surface is illuminated, 0 others	0.71801	1.393
X-angle Type 1	1 if crossing angle is type I, 0 others	0.64511	1.551
X-angle Type II	1 if crossing angle is type II, 0 others	0.65700	1.523

Furthermore, to assess which discrete variables should be included in the model, only those grade crossings were considered with a 5-year crash history greater than 0. Next, the discrete variables were grouped according to their different levels and analysis was done through a boxplot chart. Due to the limited number of crossings with non-zero crash counts (only 28 crossings), the box plots did not yield any meaningful insights. In light of this, the study was kept limited to the Variance Inflation Factor (VIF) and Tolerance (TOL) indices.

5.1.2 Interpreting Base Crash Prediction Model Output

Table 5.3 presents the estimated base ZINB model that utilized the original FRA (unaltered) HRGCs inventory data. The coding of the estimated covariates is presented in **Table 5.2**. Multiple models were estimated using different sets of candidate variables. However, the final model was selected based on the lowest AIC (Akaike Information Criterion) and BIC (Bayesian Information Criterion) values. It is important to note that the model estimation was based on the requirement that the same parameters had to be chosen for both the base and comparison models, to facilitate comparative studies.

Table 5.3 Estimated Base ZINB Model Based on Original FRA Inventory Data

Variable (code name)	Estimate	Std. Error	Z-Value	P-value
Count model coefficients (negative binomial with log link)				
Constant (_cons)	-8.7615	2.5845	-3.39	0.000
Ln-transformed AADT (IAADT)	0.47160	0.1674	2.81	0.004
Warning flashing lights (WDTLIT)	-2.1517	0.6984	-3.08	0.002
Ln-transformed max train speed (IMAXTSPD)	1.40840	0.5615	2.508	0.012
Ln (theta)	10.0436	107.89	0.093	0.925
Zero-inflation model coefficients (binomial with logit link)				
Constant (_cons)	2.2206	1.0151	2.188	0.028
Ln-transformed total daily trains (ITDTRAINS)	-0.9539	0.3596	-2.652	0.007
Pearson Residuals				
Min	1Q	Median	3Q	Max
-0.72413	-0.21093	-0.11057	-0.05989	26.92213
Summary Statistics				
Number of observations = 555				
Theta = 23008.3433				
Number of iterations = 85				
Log-likelihood = -98.01				
Degrees of freedom = 7				
Inflation model = logit				
AIC = 210.0122				
BIC = 240.2197				

To explain the base ZINB model in a mathematical term, equation (16) for count part of the ZINB model may be written as:

$$\lambda_i = e^{\beta_0+\beta_1*\text{IAADT}+\beta_2*\text{WDTLIT}+\beta_3*\text{IMAXTSPD}} \quad \text{Eq (19)}$$

$$\lambda_i = e^{-8.7615+0.47160*\text{IAADT}-2.1517*\text{WDTLIT}+1.40840*\text{IMAXTSPD}} \quad \text{Eq (20)}$$

In addition, equation (18) presenting the zero-inflation part of the ZINB model can be written as:

$$p_0 = \frac{e^{r0+r'w_i}}{1+e^{r0+r'w_i}} = \frac{e^{2.2206-0.9539*ITDTRAINS}}{1+e^{2.2206-0.9539*ITDTRAINS}} \quad \text{Eq (21)}$$

The ZINB model was estimated in R (open-source programming language), using the **zeroinfl ()** function from the **pscl** package. The output in **Table 5.3** provides information about the model fitting, coefficient estimates and their standard errors, z-values, and p-values. The first part of the output, “Count model coefficients (negative binomial with log link)”, shows the coefficients and their standard errors, z-values, and p-values for the count part of the model. The coefficients are the estimated parameters of the model and are used to predict the response variable, however, the standard errors are a measure of the precision of the coefficient estimates. The z-values are the standardized coefficients and are used to test the null hypothesis that the coefficient is equal to zero. The p-values are the probability of obtaining a z-value as extreme or more extreme than the observed one under the null hypothesis.

The Ln (theta) is the natural logarithm of the dispersion parameter. Furthermore, "Zero-inflation model coefficients (binomial with logit link)", shows the coefficients and

their standard errors, z-values, and p-values for the zero-inflation part of the estimated ZINB model based on FRA dataset. The zero-inflation part of the model was used to predict the probability of a zero count. The third part of the output, “Pearson residuals”, shows the distribution of the residuals of the model. The residuals are a measure of how well the model fits the data. Ideally, the residuals should be normally distributed with a mean of zero and a constant variance. The Pearson residuals were calculated as the ratio of the observed minus the fitted value and the square root of the fitted value. In the ZINB model output based on FRA data, the minimum, first quartile, median, third quartile, and maximum of the residuals can also be seen.

Furthermore, the dispersion parameter is denoted by Theta and its estimated value is 23008.3433. In addition, the optimization algorithm used to estimate the coefficients required 85 iterations. The log-likelihood is a measure of the goodness of fit of the model, with a higher log-likelihood indicated a better fit. The coefficient of natural log-transformed Average Annual Daily Traffic (AADT) showed a positive significance with the likelihood of crashes at HRGCs, which was significant at a 5% level. On the other hand, flashing lights as a warning device (WDTLIT) showed a negative association with predicted crashes. Additionally, the coefficient of Ln-transformed maximum timetable speed (IMTTSPD) showed a positive association with crash prediction. In conclusion, the estimates revealed that crash prediction increases with an increase in maximum timetable speed, AADT, and total daily trains. However, the presence of flashing lights as a warning device was associated with a lower crash prediction at HRGCs.

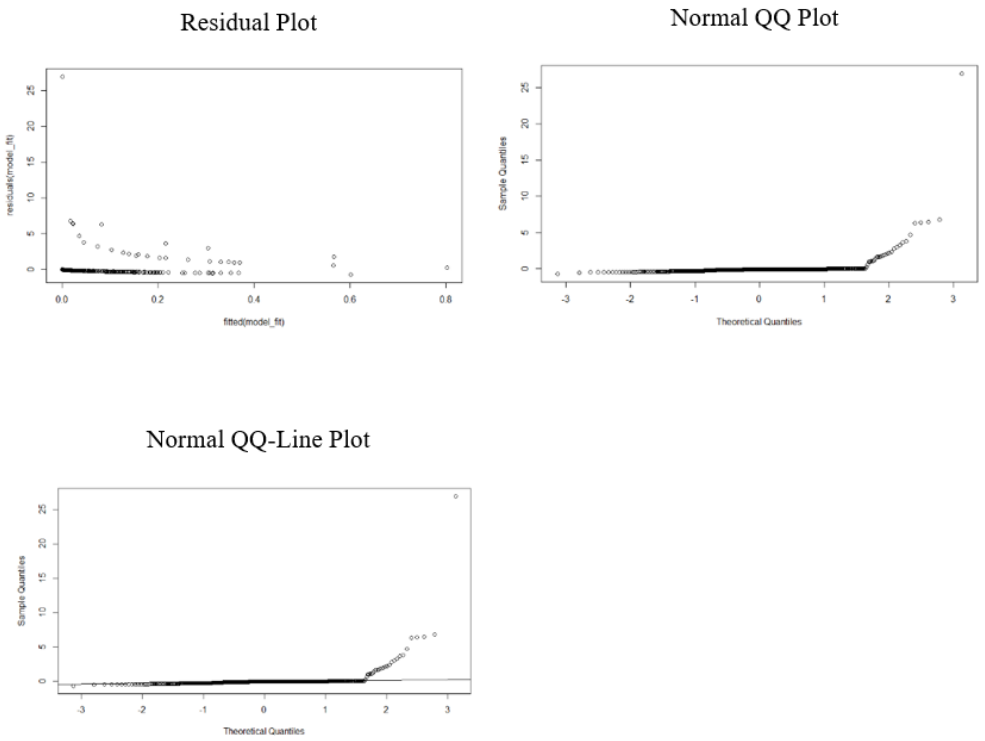

Figure 5.4 Crash Frequency Model Interpretation by Residual, Normal QQ, and QQline Plots

The residual plots and Q-Q plots were used to check the assumptions of the estimated ZINB model. The residual plot shows the residuals (the difference between the observed and predicted values) on the y-axis and the fitted values (predicted values) on the x-axis. A good model should have residuals that are randomly distributed around zero. If the residuals are not randomly distributed, it indicates that the model is not fitting the data well. Furthermore, the Q-Q plot, or quantile-quantile plots, were used to check if the residuals were normally distributed. In a normal Q-Q plot, the residuals are plotted

against a theoretical normal distribution. If the residuals follow a straight line, it indicates that the residuals are approximately normally distributed. If the residuals deviate from a straight line, it indicates that the residuals are not normally distributed. Also, The Q-Q line plot was also developed which is an extension of the normal Q-Q plot and is used to check if the residuals are normally distributed. In a normal Q-Q plot, the residuals are plotted against a theoretical normal distribution. The Q-Q line plot adds a line of best fit to the normal Q-Q plot, which allows to see if the residuals deviate systematically from the theoretical normal distribution.

The residuals plot of the estimated ZINB model prediction showed a random pattern, which suggested that the model was fitting the data well. The Q-Q plot showed that the residuals were approximately normally distributed, which was consistent with the assumption of the ZINB model (**Figure 5.4**). Furthermore, the Q-Q line plot showed that the residuals were approximately normally distributed. The residuals followed a straight line which was in agreement with the Q-Q plot. The residuals were closely aligned with the line of best fit, which meant that the residuals were approximately normally distributed. This was a good indication that the assumptions of the estimated ZINB model were met, and the model was fitting the data well.

To understand the effect of change of covariates on crash frequency, average marginal effects values were estimated (**Table 5.4**). The average marginal effects represent the change in the predicted probability of crashes at HRGCs for a one-unit change in the predictor variable, holding all other predictor variables constant. For the predictor variable IAADT, the average marginal effect was estimated to be 0.02681 which means that for a one-unit increase in the IAADT, the predicted probability of

crashes at HRGCs increased by 0.02681 on average. Furthermore, for the predictor variable WDTLIT, the AME was estimated to be -0.12232 which means that for a one-unit increase in the WDTLIT, the predicted probability of crashes decreased by 0.12232 on average.

Table 5.4 Estimated Base ZINB model Average Marginal Effects Based on Original FRA Inventory Data

Variable	**Effect**	**Std. Error**	**Z Value**	**P value**	**2.5 %**	**97.5 %**
IAADT	0.02681	0.01099	2.440	0.0146774	0.005	0.048
WDTLIT	-0.12232	0.04714	-2.595	0.0094708	-0.214	-0.029
IMTTSPD	0.08008	0.03567	2.245	0.0247619	0.010	0.149
ITDTRAINS	0.02431	0.01124	2.391	0.0166414	0.004	0.003

For the predictor variable IMTTSPD, the estimated AME was 0.08008 which means that for a one-unit increase in the IMTTSPD, the predicted probability of HRGC's crashes increased by 0.08008 on average. **Figure 5.5** provides a visual representation of the average marginal effects, showing how a unit change in each covariate affects crash prediction for the estimated Zero-Inflated Negative Binomial (ZINB) model based on the unaltered FRA inventory data for HRGCs.

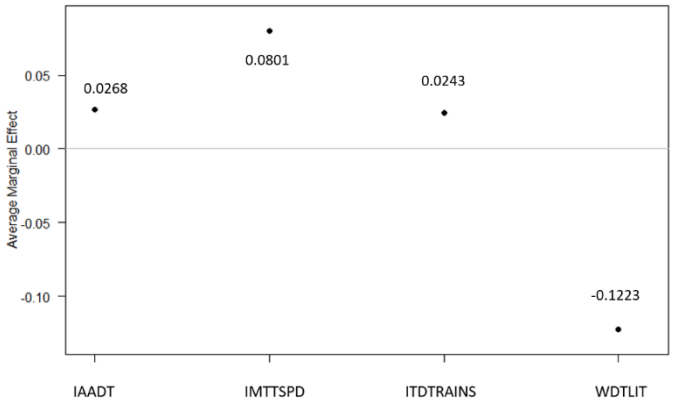

Figure 5.5 Average Marginal Effects for Crash Prediction Model Based on FRA Data

5.1.3 Interpreting Crash Prediction Output (Comparison Model)

The same parameters used in the previous model estimation based on FRA data were utilized to estimate a comparison ZINB model based on field-validated data, to understand how differences in inventory data affect crash prediction modeling. The output of the comparison ZINB model is presented in **Table 5.5**, and the coding of the variables used in the model can be found in **Table 5.2**.

Table 5.5 Estimated ZINB Comparison Model Based on Field-Validated Inventory Data

Variable (code name)	Estimate	Std. Error	Z-Value	P-value
Count model coefficients (negative binomial with log link)				
Constant (_cons)	-8.9899	2.5874	-3.47	0.000
Ln-transformed AADT (IAADT)	0.4422	0.1648	2.68	0.007
Warning flashing lights (WDTLIT)	-2.0443	0.6982	-2.92	0.003
Ln-transformed max train speed (IMAXTSPD)	1.5135	0.5667	2.67	0.007
Ln (theta)	10.4522	129.44	0.81	0.936
Zero-inflation model coefficients (binomial with logit link)				
Constant (_cons)	2.1266	1.0165	2.092	0.036
Ln-transformed total daily trains (ITDTRAINS)	-0.8793	0.3484	-2.524	0.011

Pearson Residuals				
Min	1Q	Median	3Q	Max
-0.55149	-0.22451	-0.11469	-0.05974	26.7747

Summary Statistics
Number of observations = 555
Theta = 34620.8247
Number of iterations = 84
Log-likelihood = -97.36
Degrees of freedom = 7
Inflation model = logit
AIC = 208.7161
BIC = 239.0117

To explain the comparison ZINB model (based on field validated data) in a mathematical term, equation (16) for count part of the ZINB model is presented below:

$$\lambda_i = e^{-8.9899+0.4422*\text{IAADT}-2.0443*\text{WDTLIT}+1.5135*\text{IMAXTSPD}} \quad \text{Eq (22)}$$

In addition, equation (18) presenting the zero-inflation part of the ZINB model can be written as:

$$p_0 = \frac{e^{2.1266-0.8793*ITDTRAINS}}{1+e^{2.1266-0.8793*ITDTRAINS}} \quad \text{Eq (23)}$$

The coefficients of the variables of the comparison ZINB model suggested that an increase in Ln-transformed Average Annual Daily Traffic (IAADT) and maximum timetable speed (IMTTSPD) was associated with an increase in the number of HRGCs crashes, while the presence of warning flashing lights (WDTLIT) was associated with a decrease in the number of crashes. Furthermore, the coefficient of Ln-transformed total daily trains (ITDTRAINS) suggested that an increase in the number of daily trains passing through the crossings was associated with a decrease in the probability of zero crashes. The value of theta in the model was 34620.8247, which represented the overdispersion parameter which captures the extra variation in the outcome not accounted for by the mean. The log-likelihood of the model was -97.36 on 7 degrees of freedom, and the model was optimized using the Broyden–Fletcher–Goldfarb–Shanno (BFGS) algorithm with 84 iterations. Overall, this model suggested that the number of crashes at HRGCS was influenced by Average Annual Daily Traffic (AADT), flashing

lights as warning device types at HRGCs, maximum timetable speed of trains (IMTTSPD), and total daily trains passing through the crossings (ITDTRAINS). Similar to the base ZINB model, the comparison model actually showed a better fitness of the ZINB model (**Figure 5.6**).

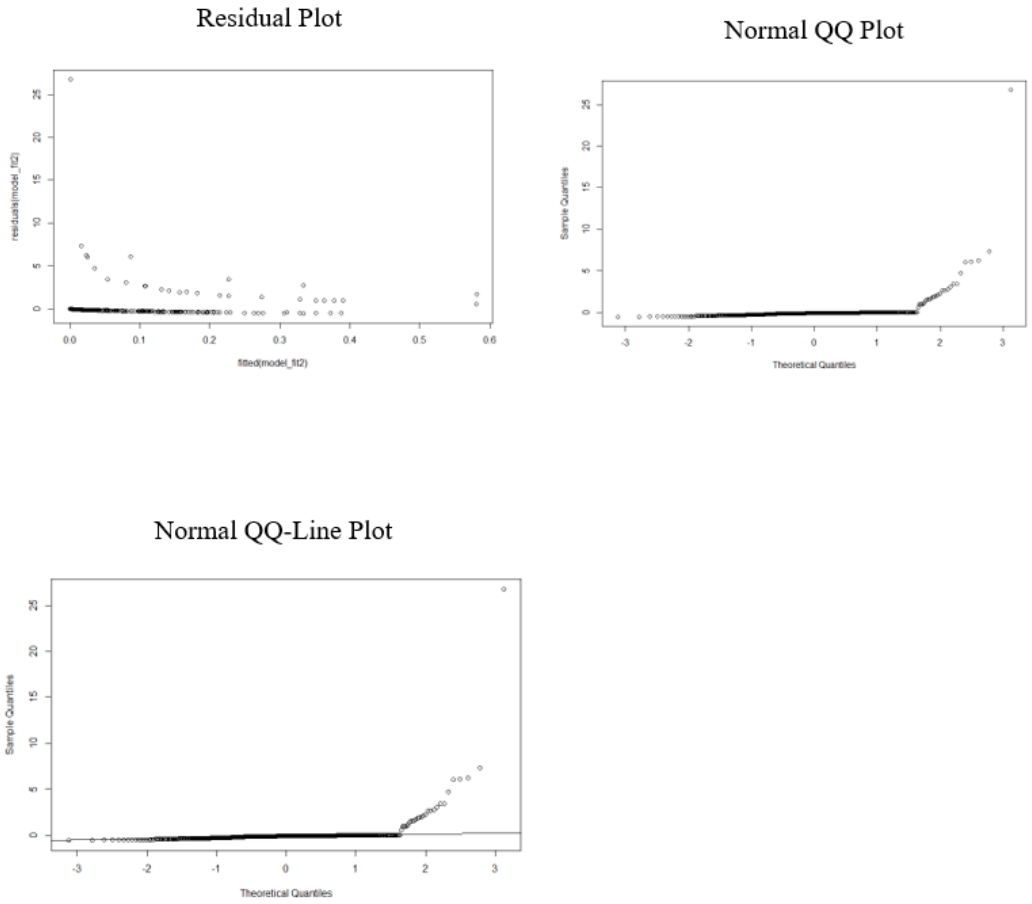

Figure 5.6 Comparison Crash Frequency Model Interpretation by Residual, Normal QQ, and QQ-line plots

Furthermore, in light of the comparison model output, it can be seen that the field validated data exhibited superior congruence with the estimated ZINB crash prediction model, as compared to the unaltered data from the FRA. This assertion is corroborated by

the Akaike Information Criterion (AIC) and Bayesian Information Criterion (BIC) values, which indicated a higher level of model fitness for the field validated data. Hence, it can also be implied that the utilization of field validated data in the estimation of the ZINB crash prediction model could yield more reliable and accurate results (**Table 5.3-5.7**).

To understand how change in independent variables affect the crash prediction estimated by the ZINB model, average marginal effects were estimated (**Table 5.6**). The results indicated that a one-unit increase in Average Annual Daily Traffic (IAADT) was associated with an average increase of 0.02546 crashes, holding all other variables constant. This effect was statistically significant at the 0.05 level, with a p-value of 0.018636. The presence of warning flashing lights (WDTLIT) was associated with an average decrease of 0.11768 crashes, holding all other variables constant. This effect was also statistically significant at the 0.05 level, with a p-value of 0.012539.

Table 5.6 Estimated Comparison ZINB model Average Marginal Effects Based on Field validated FRA Inventory Data

Variable	**Effect**	**Std. Error**	**Z Value**	**P Value**	**2.5 %**	**97.5 %**
IAADT	0.02546	0.01082	2.353	0.018636	0.00425	0.04667
WDTLIT	-0.11768	0.04713	-2.497	0.012539	-0.21006	- 0.0252
IMTTSPD	0.08714	0.03704	2.352	0.018663	0.01453	0.15974
ITDTRAINS	0.02383	0.01124	2.2912	0.021725	0.00375	0.03781

Furthermore, A one-unit increase in IMTTSPD was associated with an average increase of 0.08714 predicted crashes, holding all other variables constant. This effect was also statistically significant at the 0.05 level, with a p-value of 0.018663. Overall, these results suggested that increases in the Average Annual Daily Traffic (IAADT) and the maximum timetable speed of trains (IMTTSPD) were associated with an increase in the number of predicted crashes at HRGCs, while an increase in mean of flashing lights as warning device type (WDTLIT) at HRGCs was associated with a decrease in the number of crashes.

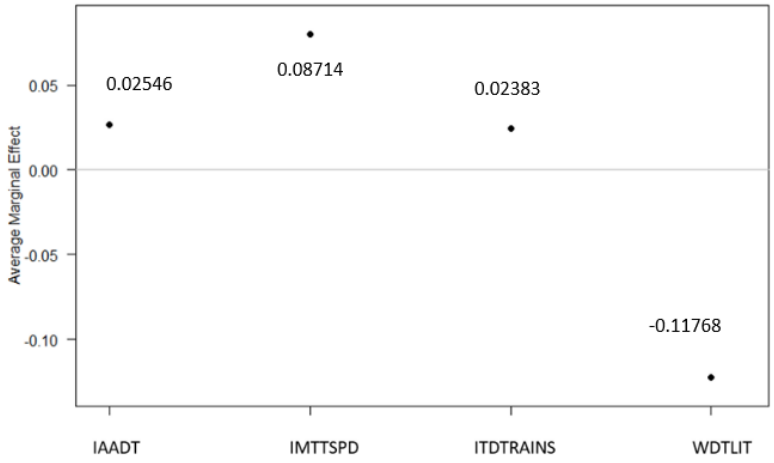

Figure 5.7 Average Marginal Effects for Crash Prediction Model Based on FV Data

Furthermore, to address Hypothesis 4 of this research, a hypothesis test based on a study by Clogg et al. (1995) was performed to determine whether the estimated coefficients of parameters from the base and comparison ZINB models were statistically significantly different from each other. The results of the test revealed that the differences between the estimated parameters' coefficients from the two models were not statistically significant (**Table 5.7**). However, there were differences in the observed crash-risk rankings of HRGCs based on the two datasets (**Table 5.8**).

Table 5.7 Comparison of Coefficients of the Base and Comparison ZINB models

Comparing Regression Coefficients of the Base and Comparison ZINB models H_0: There is no statistically significant difference between coefficients of the two models		
Compared Parameters	**Z statistic**	**P Value**
IAADT	0.00802	0.9935
WDTLIT	-0.10875	0.9134
IMAXTSPD	0.13174	0.8951
ITDTRAINS	-0.14899	0.8815

Table 5.8 Comparison of HRGCs Crash-Risk Ranking based on the FRA and Field Validated Data

RANK	Crossing ID	Predicted Value of Crash	RANK	Crossing ID	Predicted Value of Crash
1	074952M	0.801611	1	064128X	0.581035
2	074945C	0.60131	2	064129E	0.580275
3	064128X	0.566004	3	074929T	0.389984
4	064129E	0.565867	4	073326S	0.387601
5	074929T	0.370379	5	074938S	0.3775
6	073326S	0.36809	6	073456N	0.372149
7	816859H	0.359193	7	816859H	0.362821
8	073456N	0.351613	8	073342B	0.351083
9	074938S	0.347665	9	073345W	0.351083
10	073342B	0.330426	10	073455G	0.351083
11	073345W	0.330426	11	074940T	0.33412
12	073455G	0.330426	12	074956P	0.333596
13	073044B	0.314477	13	074954B	0.333596
14	074940T	0.314077	14	073316L	0.328288
15	073316L	0.307595	15	073320B	0.328288
16	073320B	0.307595	16	073340M	0.328288
17	073340M	0.307595	17	073341U	0.328288
18	073341U	0.307595	18	073459J	0.328288
19	073459J	0.307595	19	073044B	0.310565
20	074956P	0.304715	20	074860A	0.306302
21	074954B	0.304715	21	073449D	0.275392
22	074860A	0.286267	22	813669U	0.273927
23	074406N	0.277264	23	074406N	0.268641
24	813669U	0.262674	24	813592J	0.262829
25	073449D	0.255035	25	813596L	0.262829
26	813592J	0.251339	26	083524P	0.249749
27	813596L	0.251339	27	074942G	0.229268
28	083276T	0.221569	28	073292A	0.226686
29	073292A	0.215593	29	073314X	0.226686
30	073314X	0.215593	30	083276T	0.223615

5.1.4 Comparative Analysis of Estimated ZINB Models

Key factors from the regression output of base and comparison crash prediction models are as follows:

- The coefficients for Ln-transformed Maximum Timetable Speed (IMAXTSPD) and Average Annual Daily Traffic (IAADT) have positive signs in both models. However, expected magnitude differ in both models where the base model gives a higher coefficient expected-magnitude for IAADT as compared to the comparison model. Moreover, for IMAXTSPD, comparison model gives a higher coefficient expected-magnitude. The average marginal effect estimates presented in **Table 5.3** and **Table 5.5** show that a unit change in IAADT increases predicted crashes by 0.02681 in base model and 0.02545 in the comparison model.
- The coefficients for warning device type as flashing lights (WDTLIT) are negative for both models (i.e., compared to passive devices, warning flashing lights reduce predicted crashes). However, coefficient expected-magnitudes and average marginal effects of WDTLIT differ in both models where the base model estimates a higher negative value compared to the comparison model.
- The coefficients of Ln-transformed total daily trains (ITDTRAINS) in the zero-inflated part in both base and comparison models are negative indicating that the probability of excess zeros decreases with the increase in number of trains, as expected. However, coefficient expected-magnitudes of ITDTRAINS differ in both models.

- All the coefficients retained in both models indicate strong statistical significance, however based on AIC and BIC values, ZINB model based on field validated HRGCs inventory data shows a better fit.
- The hypothesis tests performed in **Table 5.7** indicated that the parameter's coefficients estimated using two different datasets on the comparative ZINB models were not statistically significantly different (Hypothesis 4 of this research). However, different crash-risk rankings were observed based on the predicted values of crashes obtained using the field-validated and FRA datasets on the estimated base and comparison ZINB models (**Table 5.8**).

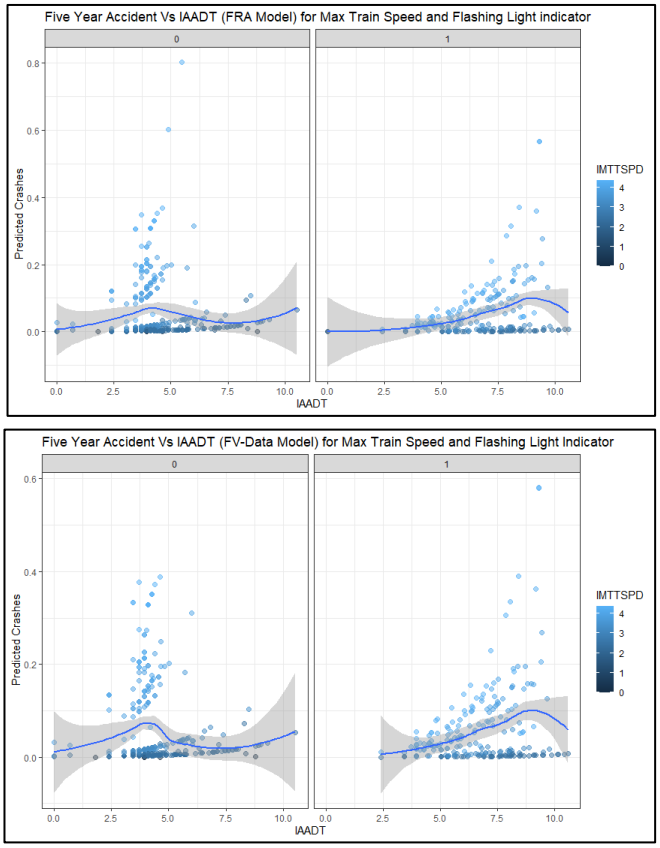

Figure 5.8 Multivariate Ggplots for Predicted Crashes of Base (FRA) and Comparison Model (FV) for Maximum Train Speed, AADT, and Flashing Light Indicator

Figure 5.8 illustrates a "Predicted crashes (expected crashes) verses AADT (natural logarithm)" charts for flashing light indicator. The charts indicated that for cases of HRGCS with no flashing lights in the original FRA dataset, the model gave higher crash predictions for maximum timetable speed ranging from 25 to 45 mph. However, the presence of flashing lights decreased predicted crashes but only for low-speed trains. The trend of predicted crashes in the field-validation based model was different. It also

illustrates estimation of higher crash predictions for no flashing lights indicator with higher speeds.

Furthermore, for convenient comparison of average marginal effects of the ZINB models, based on FRA and field validated HRGCs inventory data, **Figure 5.9** is presented.

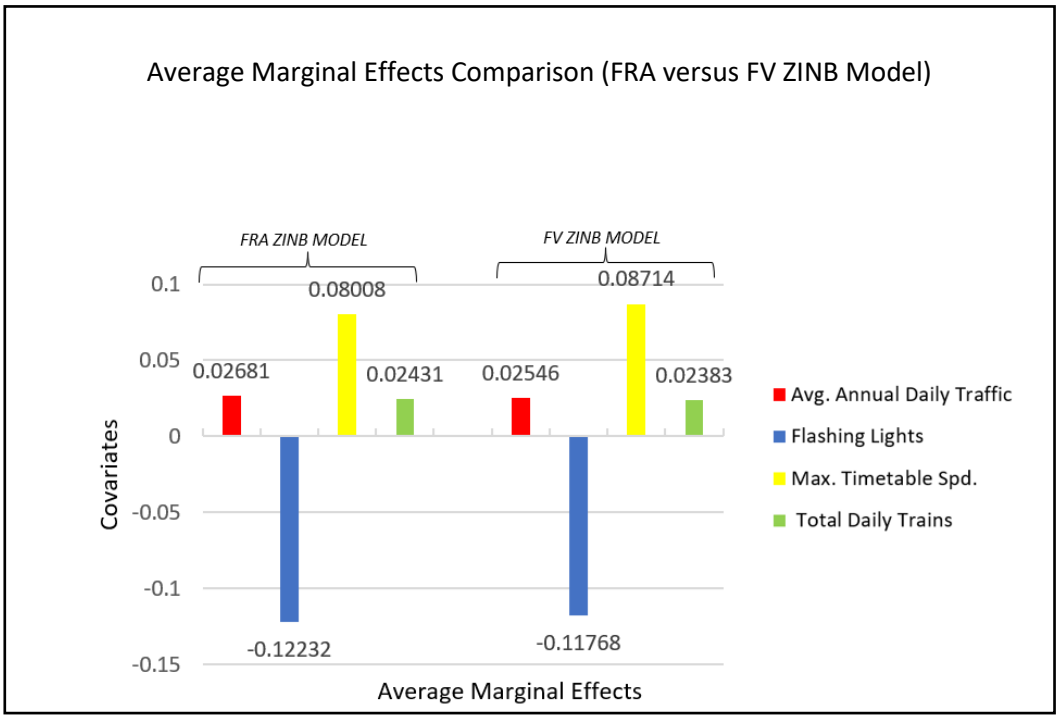

Figure 5.9 Comparison of Average Marginal Effects of Estimated Covariates of ZINB Model Based on FRA and Field Validated Data

Figure 5.9 indicated that estimated crash prediction model parameters and their respective average marginal values were different for when the models were based on unaltered FRA HRGCs inventory database and the corrected and complete (field validated) HRGCs inventory data.

5.2 Crash Severity Modelling

According to the FRA, recorded crashes on HRGCs are organized into three severity categories: fatal, injury, and property damage only (PDO). A grade crossing crash is classified as fatal if it results in at least one fatality, as injury if it leads to at least one injury, and as PDO if there are no injuries or fatalities. These severity categories are arranged in a specific order, with fatal crashes being the most severe, followed by injury crashes, and then PDO crashes, which are the least severe. By categorizing crashes based on their severity level, authorities can prioritize and allocate resources to improve safety at grade crossings, with the ultimate goal of preventing or reducing the frequency and impact of accidents (Brod et al., 2020).

To investigate the effects of data inaccuracies in HRGCs inventory data on crash severity model estimation, further analysis was conducted by estimating a crash severity model based on the FRA and field-validated dataset. As mentioned earlier, only 34 crashes were recorded on the selected 560 HRGCs in Nebraska in the 5 year-period (2016-2020), which is a small sample size for suitable crash severity model estimation. Thus, the crash severity model was estimated by integrating past 15 years (2007-2021) crash dataset with the inventory dataset of the corresponding HRGCs. Comparative studies were then performed to determine if field-validated data showed significant differences in the estimated parameter of crash prediction model when compared with the FRA dataset-based crash severity model.

A combination of different ordered logit and probit models with varying parameters was used for selection of the final crash severity model. It is noteworthy that

the study abided by the requirement of estimating two comparative models with the same parameters so that a meaningful comparison could be made. Consequently, an ordered probit model was estimated with a suitable model fitness and the same parameters. **Figure 5.10** presents some key attributes, assumptions and limitations of ordered probit model.

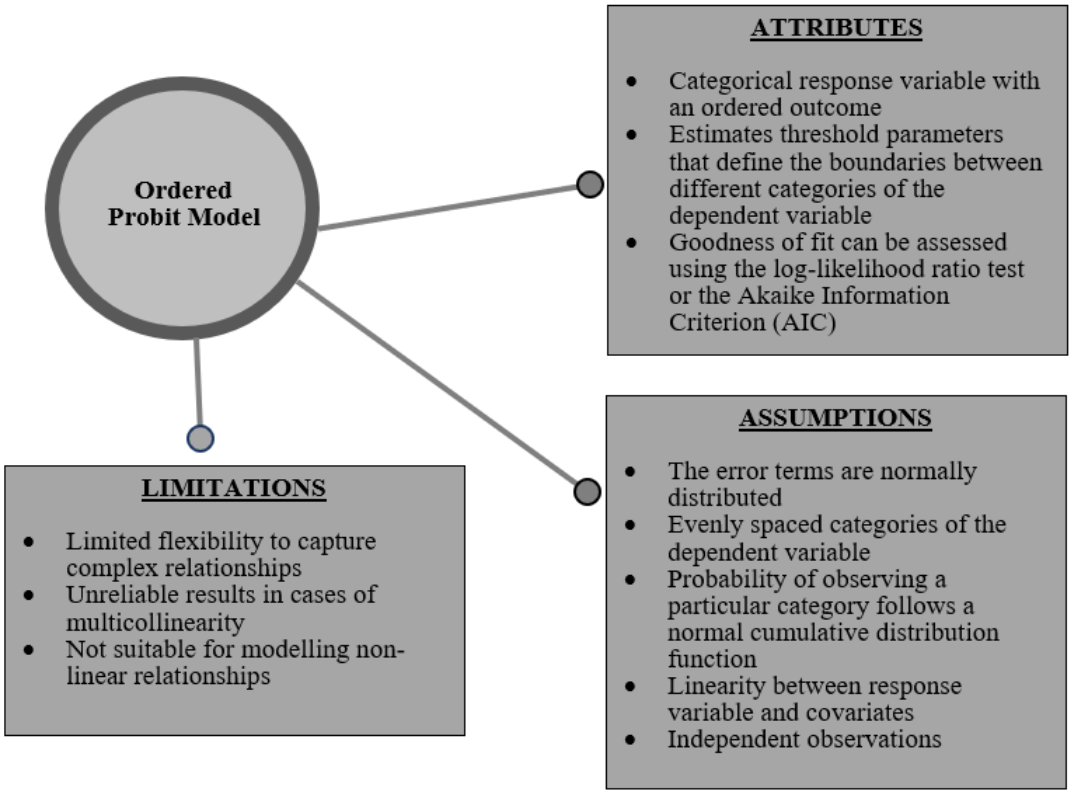

Figure 5.10 Key Attributes, Assumptions and Limitations of Ordered Probit Model

An ordered probit model is a statistical model used to analyze ordinal categorical data, which involves categories with a natural ordering (Daykin and Moffatt 2002; Farooq and Khattak, 2023). It is predicated that the observed ordinal responses are based on the relative position of the latent variable within those intervals, and that the

dependent variable is generated by an underlying continuous latent variable that is partitioned into a set of ordered intervals. The model estimates the relationship between the latent variable and one or more predictor variables, typically using maximum likelihood estimation (Kockelman and Kweon, 2002).

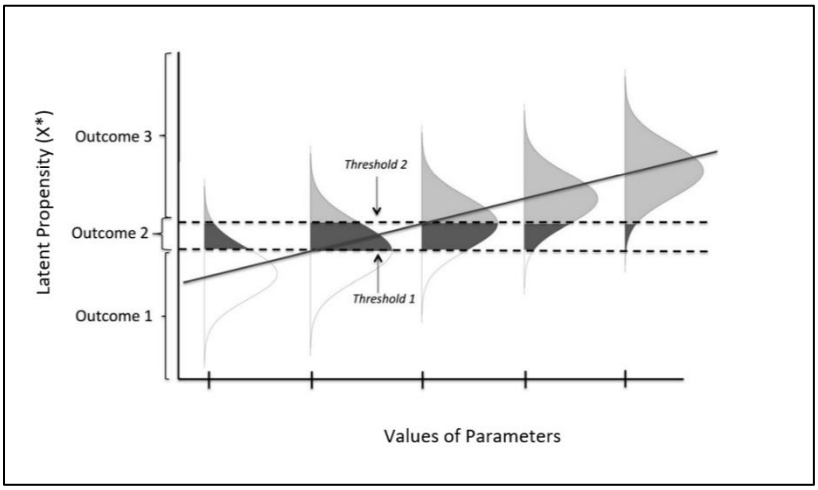

Figure 5.11 Simulated Ordered Probit Model (Shi, 2019)

For the ordered probit model with three ordered categories of HRGCs crash severity, the probabilities for each crash category are presented below. In the ordered probit model, it is assumed that the cumulative distribution function (CDF) follows a normal distribution, and the model estimates the parameters of this distribution to estimate the probabilities of the observed outcomes. The probabilities of estimating a fatal, injury and PDO crash are given as follows.

Probabilities to estimate – Fatal HRGC crash

$$P(\text{crash type} = \text{fatal HRGC crash} \mid A) \quad \text{eq (24)}$$

Probabilities to estimate – HRGC crash with Injury

$$\mathrm{P}(\text{ acctype } = \text{ injury causing HRGC crash} \mid A) \quad \text{eq (25)}$$

Probabilities to estimate-PDO HRGC crash

$$\mathrm{P}(\text{ acctype } = P\text{DO HRGC crash} \mid A) \quad \text{eq (26)}$$

Crash severity probabilities Sum to 1

$$\mathrm{P}(\text{ fatal crash } \mid A) + \mathrm{P}(\text{ injury crash} \mid A) + \mathrm{P}(\text{PDO crash} \mid A) = 1 \quad \text{eq (27)}$$

The dependent variable of the model is an observed ordinal variable X (in this study, HRGC crash severity type). The model assumes that there is a continuous, unmeasured latent variable, X^*, whose values determine the value of the observed ordinal variable X. The variable X^* has two threshold points represented by κ (the lowercase Greek letter kappa).

The value of the observed variable X depends on whether X* has crossed a threshold, as shown below:

The relationship between X and X^* is presented in equation 28.

$$X_i = \begin{cases} \text{PDO HRGC Crash,} & \text{if } X_i^* \leq \kappa_1 \\ \text{Injury HRGC Crash,} & \text{if } \kappa_1 \leq X_i^* \leq \kappa_2 \\ \text{Fatal HRGC Crash,} & \text{if } X_i^* \geq \kappa_2 \end{cases} \quad \text{eq (28)}$$

The HRGCs inventory characteristics are a function of the latent variable X. As a result, the following gives the ordered probit model to estimate for a given specification (i.e., for a chosen set of explanatory variables from the HRGCs inventory dataset):

$$P(X_i = PDO\ HRGC\ Crash) = \frac{1}{1+\exp(Z_i - \kappa_1)} \quad \text{eq (29)}$$

$$P(X_i = \text{ Injury HRGC Crash}) = \frac{1}{1+\exp(Z_i - \kappa_2)} - \frac{1}{1+\exp(Z_i - \kappa_1)} \quad \text{eq (30)}$$

$$P(X_i = \text{ Fatal HRGC Crash}) = 1 - \frac{1}{1+\exp\,(Z_i - \kappa_2)} \quad \text{eq (31)}$$

Where, the subscript i indicates an index of a recorded HRGC crash, X_i is the variable indicating HRGC crash type (fatal, injury or PDO). And κ_1 is a coefficient of the threshold separating PDO from injury crash, and κ_2 is a coefficient of the threshold separating injury from fatal crash (Kockelman and Kweon, 2002). Furthermore, the distribution of different ordinal severity levels of crashes at 560 selected HRGCs over a 15-year period (2007-2021) is presented in **Figure 5.12**. Of the total 83 crashes, 62 percent resulted in property damage only, 30 percent resulted in at least one reported injury, and unfortunately, 8 percent of crashes reported at least one fatality.

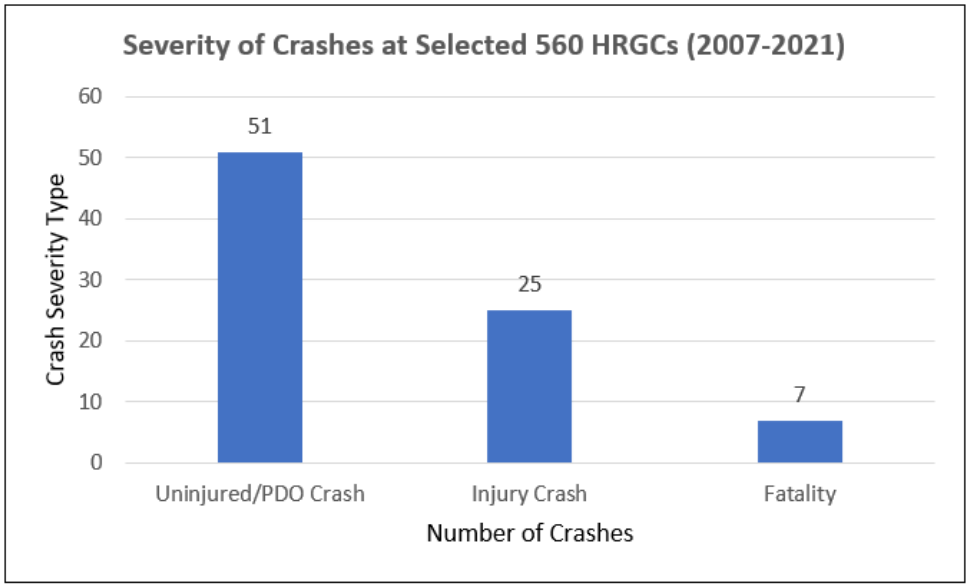

Figure 5.12 Severity of Crashes at Selected 560 HRGCs (2007-2021)

5.2.1 Interpreting Base Crash Severity Model Output

Table 5.9 presents the estimated base crash severity model that utilized the original FRA (unaltered) HRGCs inventory data. Multiple models were estimated using different sets of candidate variables. However, the final model was selected based on the lowest AIC (Akaike Information Criterion) and BIC (Bayesian Information Criterion) values. The ordered probit was developed in R by using "**ordinal**" package and "**clm**" function.

Table 5.9 Estimated Base Crash Severity Model Based on Original FRA Inventory Data

Variable (code name)	**Estimate**	**Std. Error**	**Z-Value**	**P-value**
Coefficients				
Flashing light indictor (1 if the HRGC has flashing lights as a warning device, 0 otherwise)	-0.8661	0.3842	- 2.254	0.0242
Rural_Xng indicator (1 if the HRGC is in rural area, 0 otherwise)	0.5330	0.2939	1.996	0.0459
Rdwithin_500ft indicator (1 if the HRGC is within 500 ft of the road, 0 otherwise)	0.9982	0.3971	2.513	0.0120
Threshold Coefficients				
PDO HRGC Crash | Injury HRGC Crash	0.5191	0.2511	2.067	-
Injury HRGC Crash |Fatality HRGC Crash	1.7004	0.2989	5.689	-
Summary Statistics				
Number of observations = 83				
Log-likelihood = -66.33				
Ordered model = Probit				
AIC = 142.66				
BIC = 164.21				

The results presented in **Table 5.9** can be interpreted as follows:

- Flashing Light: The coefficient of -0.8661 indicated that HRGCs with flashing lights as a warning device exhibited a lower log-odds of moving up one severity level compared to crossings without flashing lights, holding other variables constant. This coefficient is statistically significant at the 5% level (p-value=0.0242).
- Rural_Xng: The coefficient of 0.5330 indicated that HRGCs in rural areas exhibited a higher log-odds of moving up one severity level compared to HRGCs in non-rural areas, holding other variables constant. In addition, this coefficient is statistically significant at the 5% level (p-value=0.0459).
- Rdwithin_500ft: The coefficient of indicator variable showing the crossing near 500 ft of a road was positively estimated to be 0.9982 indicating that HRGCs within 500 ft of the roadway showed a higher log-odds of moving up one severity level compared to crossings further away from the roadway, holding other variables constant. This coefficient is statistically significant at the 5% level (p-value=0.0120).

The threshold coefficients estimate the cut points between the severity categories. The estimated threshold for the cut point between PDO and Injury was 0.5191. This means that if the model predicts a value greater than 0.5191, the observation is more likely to fall into severity level 2 (Injury) than level 1 (PDO). Furthermore, the estimated threshold for the cut point between Injury and Fatality was 1.700. meaning that if the model predicts a value greater than 1.700, the observation is more likely to fall into severity level 3 (Fatality) than level 2 (Injury). Overall, the results suggested that HRGCs with flashing lights as a warning device and crossings further away from the roadway indicated lower probabilities of more severe injuries in the event of a rail crossings' crash.

The HRGCs crash severity equation for base model based on unaltered FRA data takes the form:

$$Z_i = \sum_{k=1}^{3} \beta_k X_{ki} = \beta_1 \cdot \text{Flashing\_Lit} + \beta_2 \cdot \text{Rural\_Xng} + \beta_3 \cdot \text{Rdwithin\_500ft} \quad \text{eq (32)}$$

$$Z_i = -0.8661 \cdot \text{Flashing\_Lit} + 0.5330 \cdot \text{Rural\_Xng} + 0.9982 \cdot \text{Rdwithin\_500ft} \quad \text{eq (33)}$$

$$P(X_i = PDO\ HRGC\ Crash) = \frac{1}{1+\exp(Z_i - \kappa_1)} \quad \text{eq (34)}$$

$$P(X_i = \text{Injury HRGC Crash}) = \frac{1}{1+\exp(Z_i - \kappa_2)} - \frac{1}{1+\exp(Z_i - \kappa_1)} \quad \text{eq (35)}$$

$$P(X_i = \text{Fatal HRGC Crash}) = 1 - \frac{1}{1+\exp(Z_i - \kappa_2)} \quad \text{eq (36)}$$

Using **marginaleffects** functions in R, average marginal effects values were used to analyze how changes in covariate affect each ordinal category of crash severity at HRGCs. The marginal effects indicate the shift in the probability of the outcome variable when the independent variable changes from 0 to 1, while holding all other independent variables constant, when the independent variables are indicator variables (also known as dummy variables), which take on values of 0 or 1. The function computes the marginal effect using the partial derivative of the regression equation with respect to the independent variable (Farooq and Khattak, 2023; Farooq et al., 2021).

Table 5.10 Average Marginal Effects for Base Crash Severity Model Based on Original FRA Inventory Data

Group	Variable (code name)	Estimate	Std. Error	Z-Value	P-value
1 (PDO)	Flashing_Lit	0.295753	0.120519	2.454004	0.014128
	Rural_Xng	-0.182	0.094577	-1.9244	0.054305
	Rdwithin_500ft	-0.34087	0.121841	-2.79763	0.005148
2 (Injury)	Flashing_Lit	-0.17277	0.072793	-2.3734	0.017625
	Rural_Xng	0.106314	0.055894	1.902079	0.057161
	Rdwithin_500ft	0.199108	0.075319	2.643519	0.008205
3 (Fatality)	Flashing_Lit	-0.12299	0.062203	-1.97718	0.048021
	Rural_Xng	0.075689	0.04585	1.650822	0.098775
	Rdwithin_500ft	0.141759	0.065285	2.17137	0.029903

For Group 1, HRGCs without flashing lights had a 29.6 percentage point reduction in the probability of a more severe crash for each grade crossing with flashing lights added, while crossings located within 500 feet of a road intersection had a 34.1 percentage point reduction. Furthermore, rural HRGCs had an 18.2 percentage point reduction (**Table 5.10**).

For Group 2, HRGCs with flashing lights had a 17.3 percentage point reduction in the probability of a more severe crash for each crossing without flashing lights, while rural crossings had a 10.6 percentage point increase, and crossings located within 500 feet of a road intersection had a 19.9 percentage point increase. For Group 3, HRGCs with flashing lights had a 12.3 percentage point reduction in the probability of a more severe crash for each crossing without flashing lights, while rural HRGCs had a 7.6 percentage point increase, and crossings located within 500 feet of a road intersection had a 14.2 percentage point increase. In conclusion, the study found that the presence of flashing lights and the location of the crossing more than 500 feet from a road intersection were associated with a reduced probability of a more severe crash, while rural crossings were associated with an increased probability of a more severe crash.

5.2.2 Interpreting Comparison Crash Severity Model Output

Table 5.11 Estimated Crash Severity Model Based on Field-Validated Inventory Data

Variable (code name)	Estimate	Std. Error	Z-Value	P-value
Coefficients				
Flashing light indictor (1 if the HRGC has flashing lights as a warning device, 0 otherwise)	-0.8148	0.3817	- 2.135	0.03277
Rural_Xng indicator (1 if the HRGC is in rural area, 0 otherwise)	0.8902	0.3431	2.594	0.00947
Rdwithin_500ft indicator (1 if the HRGC is within 500 ft of the road, 0 otherwise)	1.0708	0.3953	2.709	0.00675
Threshold Coefficients				
PDO HRGC Crash | Injury HRGC Crash	0.5036	0.2143	2.350	-
Injury HRGC Crash |Fatality HRGC Crash	1.7283	0.2745	6.295	-
Summary Statistics				
Number of observations = 83				
Log-likelihood = -64.61				
Ordered model = Probit				
AIC = 139.21				
BIC = 159.17				

The results presented in **Table 5.11** were interpreted as follows:

- Flashing Light: The estimated coefficients show that the indicator variable for warning device type flashing light showed a negative coefficient, indicating that the presence of flashing lights as a warning device at the crossing reduces the severity of crashes. The coefficient is statistically significant at the 5% level (p-value = 0.03277).
- Rural_Xng: The indicator variable showing that HRGCs situated in a rural area "Rural_Xng" showed a positive coefficient, indicating that crashes in rural areas are more severe than those in urban areas. The coefficient is statistically significant at the 1% level (p-value = 0.00947).
- Rdwithin_500ft: The indicator variable showing that the crossing is situated withing 500 ft of the road (Rdwithin_500ft) also showed a positive coefficient, indicating that crashes that occur within 500 ft of the roadway are more severe than those that occur farther away. The coefficient is statistically significant at the 1% level (p-value = 0.00675).

The threshold coefficients estimated the cut points between the crash severity categories. The estimated threshold for the cut point between severity levels 1= PDO and 2= Injury is 0.5036, and the estimated threshold for the cut point between severity levels 2=Injury and 3=Fatality was 1.7283. Overall, the results suggested that the presence of flashing lights as a warning device can reduce the severity of crashes, and crashes in rural areas and those close to the roadway tend to be more severe.

The HRGCs crash severity equation for comparison model based on field-validated data takes the form:

$$Z_i = \sum_{k=1}^{3} \beta_k X_{ki} = \beta_1 \cdot \text{Flashing\_Lit} + \beta_2 \cdot \text{Rural\_Xng} + \beta_3 \cdot \text{Rdwithin\_500ft} \quad \text{eq (37)}$$

$$Z_i = -0.8148 \cdot \text{Flashing\_Lit} + 0.8902 \cdot \text{Rural\_Xng} + 1.0708 \cdot \text{Rdwithin\_500ft} \quad \text{eq (38)}$$

$$P(X_i = PDO\ HRGC\ Crash) = \frac{1}{1+\exp(Z_i - \kappa_1)} \quad \text{eq (39)}$$

$$P(X_i = \text{Injury HRGC Crash}) = \frac{1}{1+\exp(Z_i - \kappa_2)} - \frac{1}{1+\exp(Z_i - \kappa_1)} \quad \text{eq (40)}$$

$$P(X_i = \text{Fatal HRGC Crash}) = 1 - \frac{1}{1+\exp(Z_i - \kappa_2)} \quad \text{eq (41)}$$

Table 5.12 presents the average marginal effects of parameters of ordered probit model based on the field-validated data. The output was divided into three groups based on the values of the independent variables. For example, group 1 represents cases where Flashing_Lit, Rural_Xng, and Rdwithin_500ft are all at their minimum values. Similarly, group 2 represents cases where all the independent variables are at their mean values, and group 3 represents cases where they are all at their maximum values. Flashing light showed a positive effect on HRGCs crash severity in group 1 and a negative effect in groups 2 and 3.

Table 5.12 Average Marginal Effects for Comparison Crash Severity Model Based on Field Validated Inventory Data

Group	Variable (code name)	Estimate	Std. Error	Z-Value	P-value
1 (PDO)	Flashing_Lit	0.270566	0.118078	2.291415	0.021939
	Rural_Xng	-0.29561	0.101398	-2.91531	0.003553
	Rdwithin_500ft	-0.35557	0.115823	-3.06991	0.002141
2 (Injury)	Flashing_Lit	-0.15879	0.07127	-2.22802	0.025879
	Rural_Xng	0.173479	0.062812	2.76188	0.005747
	Rdwithin_500ft	0.208665	0.071175	2.931731	0.003371
3 (Fatality)	Flashing_Lit	-0.11177	0.058808	-1.90064	0.057349
	Rural_Xng	0.122128	0.054693	2.232958	0.025552
	Rdwithin_500ft	0.146901	0.064628	2.273036	0.023024

Rural_Xng and Rdwithin_500ft both showed negative effects on predicted HRGCs crash severity in category 1, indicating that the severity of PDO crashes decrease when they occur in rural areas or farther away from city limits. However, the severity of injury and fatality showed an increase in cases of crashes occurred in rural areas and for crossings within 500 feet of the road. According to group 3, with HRGCs being within 500 ft of the roads increased the probability of fatality to 14 percent, however, crossings in rural areas showed that the probability of an injury crash would increase to 7.5 percent. Overall, the average marginal effects give intuitive estimates like the effects calculated for the base crash severity prediction mode. However, the values of these marginal effects for all the three categories of severity are different from each other (**Table 5.12**).

Table 5.13 Comparison of Coefficients of the Base and Comparison OPM models

Comparing Regression Coefficients of the Base and Comparison OPM models H_0: There is no statistically significant difference between coefficients of the two models		
Compared Parameters	**Z statistic**	**P Value**
Flashing Light Indicator	-0.09472	0.9245
Rural Crossing Indicator	0.79067	0.4291
Road within 500 ft	0.12957	0.8969

Furthermore, to address Hypothesis 5 of this research, a hypothesis test based on a study by Clogg et al. (1995) was performed to determine whether the estimated coefficients of parameters from the base and comparison Ordered Probit models were statistically significantly different from each other. The results of the test revealed that the differences between the estimated parameters' coefficients from the two models were not statistically significant (**Table 5.13**).

5.2.3 Comparative Analysis of Estimated Ordered-Probit Models

Key factors from the regression output of ordered-probit base and comparison models are as follows:

- The coefficients for the indicator variable for flashing lights (Flashing_Lit) in both the base and comparison models for predicting crash severity showed negative signs. However, the expected magnitudes differed in both models, where the base model gave a higher coefficient expected-magnitude for Flashing_Lit as compared to the comparison model. Furthermore, the coefficients of average marginal effects for the indicator variable for flashing lights (Flashing_Lit) in both models showed different estimates for all categories (**Table 5.7-5.10**).
- The coefficients for the indicator variable for rural HRGCs (Rural_Xng) in both the base and comparison models to predict crash severity showed positive signs, indicating that crash severity increases if the crash occurs at a rural grade crossing. However, the expected magnitudes of these coefficients differed in the two models, with the base model giving a lower expected magnitude for Rural_Xng compared to the comparison model. Additionally, the coefficients for the average marginal effects of the indicator variable for rural grade crossing (Rural_Xng) in both models had different estimates for all categories. (**Table 5.7-5.10**).
- The coefficients for the indicator variable for HRGCs within 500 feet of the road (Rdwithin_500ft) in both the base and comparison models for predicting crash severity had positive signs. However, the expected magnitudes differed in both models, where the base model gave a lower coefficient expected-magnitude for Rdwithin_500ft as compared to the comparison model.

Furthermore, the coefficients of average marginal effects for the indicator variable for HRGCs within 500 feet of the road (Rdwithin_500ft) in both models had different estimates for all categories (**Table 5.7-5.10**).

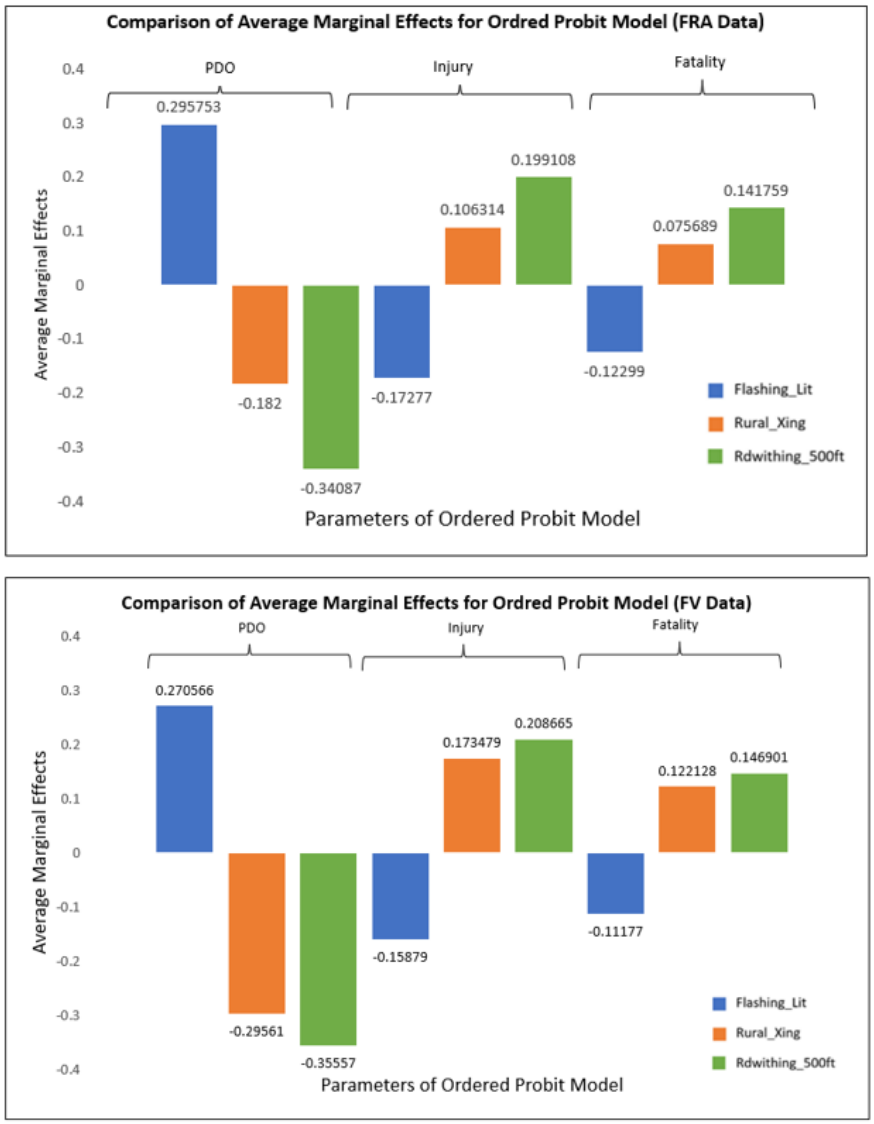

Figure 5.13 Comparison Plots of Average Marginal Effects Based on FRA and FV Ordered-Probit Model

- **Figure 5.13** presented a comparative analysis of the average marginal effects values for all categories of crash severity estimated for the base and comparison ordered probit models. From the figure, it is evident that the effect on the response variable by a change in the indicator variable for the two different models was significantly different for each case.
- All the coefficients retained in both models indicated a strong statistical significance, however based on AIC and BIC values, ordered probit model based on field validated HRGCs inventory data showed a better fit.
- There was a substantial difference between the observed and predicted value plots estimated for the base and comparison ordered probit models for crash severity prediction for selected HRGCs. **Figure 5.14** showed that the field-validated model predicted a few PDO crashes as injury crashes, while the FRA-based ordered probit model predicted many PDO crashes to be injury crashes. Similarly, the field-validated model estimated more fatality crashes for observed crash severity of PDO crashes compared to the ordered probit crash severity model based on unaltered FRA data.
- The hypothesis tests performed in **Table 5.13** indicated that the parameter coefficients estimated by the comparison OPM models, using two different datasets were not statistically significantly different.

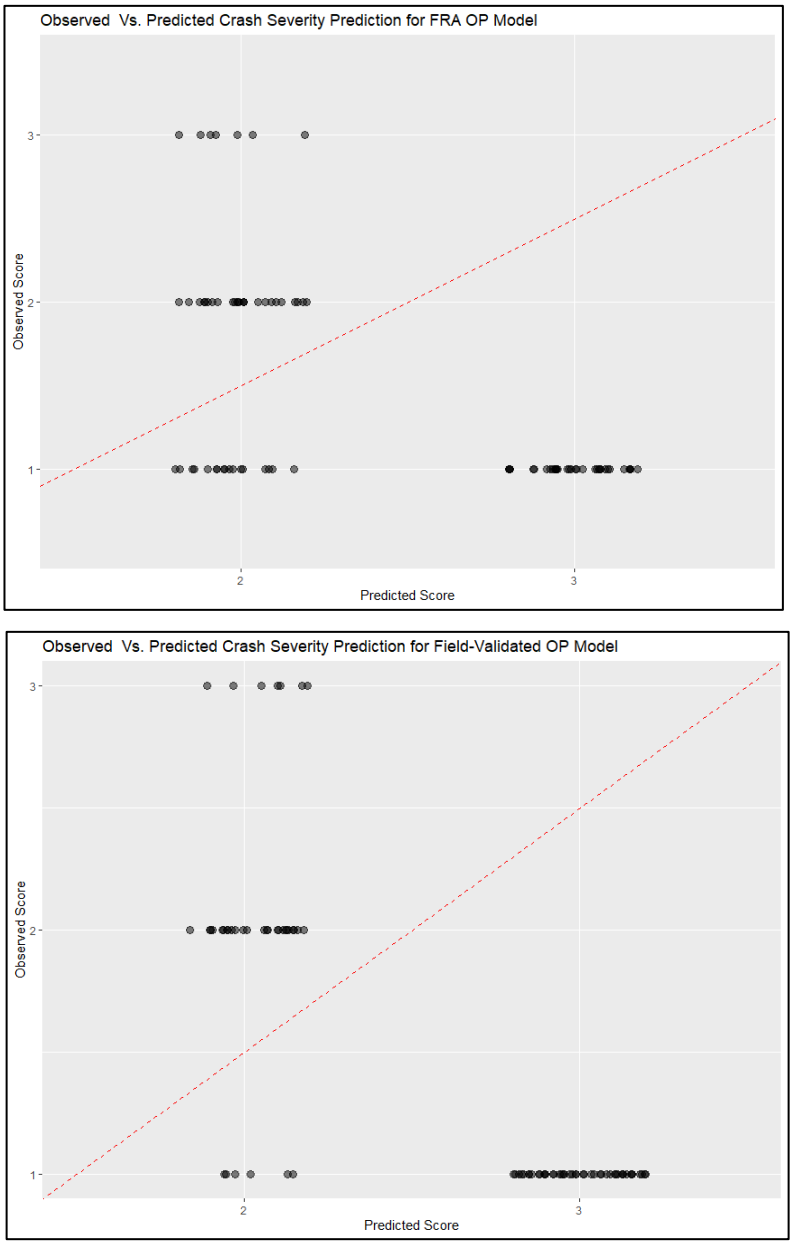

Figure 5.14 Observed Vs. Predicted Crash Severity Based on FRA and FV Ordered-Probit Model

5.3 Chapter Summary

This important chapter examined the impact of inaccuracies in HRGCs inventory data on crash prediction and severity models. To conduct a thorough analysis of how a specific parameter's coefficient differs from others and to what extent, identical parameters were used in the estimation of comparative crash frequency and severity models.

To predict crash frequency at HRGCs, two Zero-Inflated Negative Binomial models (ZINB) were estimated, using both unaltered (FRA-provided) and field-validated data for 560 selected HRGCs in Nebraska. The ZINB model was selected because the response variable, crash frequencies at HRGCs, exhibited significant over-dispersion with many zeros. To evaluate the models' performance, the Akaike Information Criterion (AIC) and Bayesian Information Criterion (BIC) were used. Candidate variables for model estimation were tested for multicollinearity to ensure the stability and dependability of the results. As volume-related factors tend to have a skewed distribution, they were Ln-transformed to meet the assumption of normality required for many statistical models. The estimated ZINB models revealed that Ln-transformed Average Annual Daily Traffic (AADT) had a positive association with the likelihood of crashes at HRGCs, while flashing lights as a warning device (WDTLIT) had a negative association with predicted crashes. Additionally, the coefficient of Ln-transformed maximum timetable speed (IMTTSPD) had a positive association with crash prediction.

Furthermore, the zero-inflated part of both crash prediction models revealed a negative association of Ln-transformed total daily trains (ITDTRAINS) with the occurrence of zero counts for crash prediction. All the coefficients retained in both models indicated strong statistical significance, but based on AIC and BIC values, the ZINB model based on field validated HRGCs inventory data showed a better fit. Finally, by estimating the average marginal effects for parameters of both ZINB models, it was found that the estimated parameters and their respective average marginal values were different for when the models were based on unaltered FRA HRGCs inventory database and the corrected and complete (field validated) HRGCs inventory data. Furthermore, two Ordered Probit models were utilized to predict crash severity, incorporating the past 15 years' crash data (2007-2021) with the corresponding inventory dataset of the HRGCs. The research performed comparative analysis to determine if the field-validated data demonstrated any significant disparities in the estimated parameter of the crash prediction model when contrasted with the FRA dataset-based crash severity model. For both Ordered Probit models, a statistically significant correlation was observed between the severity of the crash at the HRGCs and the indicator variables for rural crossings (Rural_Xng), crossings within 500 ft of the road (Rdwithin_500ft), and flashing lights (Flashing_Lit). However, the estimated average marginal effects for the parameters of the two Ordered Probit models showed statistically significant differences. Furthermore, the severity model based on the field validation and complete inventory dataset exhibited a better model fitness.

CHAPTER 6 SUMMARY, CONCLUSIONS AND FUTURE RESEARCH

This dissertation aimed to achieve five research objectives: (1) to investigate accuracy and missing values in HRGCs inventory data and to examine if the missing values in the inventory data follow any specific pattern; (2) to investigate if there are any statistical differences in expected value of HRGCs crashes estimated by employing the 2020 FRA APS model by utilizing both FRA and field-validated data; (3) to investigate if there are any statistical differences in HRGCs crash severity predictions obtained by employing the 2020 FRA APS model based on the FRA and field-validated data; (4) to assess impact of data inaccuracy on HRGCs crash frequency modelling, and lastly, (5) to assess impact of data inaccuracy on HRGCs crash severity modelling.

This chapter provides a summary of the research findings, conclusions, and recommendations aimed at improving the quality of data and reporting of HRGCs inventory data. It also highlights the limitations and contributions of this research and identifies future research avenues for enhancing the data quality of HRGCs inventory records. The ultimate objective of this research is to promote safety at HRGCs.

To achieve this goal, it is crucial to ensure that the HRGCs inventory data are accurate and up-to-date. Therefore, the recommendations proposed in this chapter address the identified limitations and provide recommendations for improving the quality of data and reporting. Moreover, this chapter underscores the significance of this research and its contributions towards enhancing the safety of HRGCs. It highlights the need for further research to build upon these findings and ensure that safety remains a top priority in the rail transportation sector. Briefly, this chapter serves as a critical resource for

stakeholders in the rail transportation industry, including policymakers, researchers, regulatory bodies, and practitioners, who are committed to improving the quality of HRGCs inventory and enhancing rail crossing safety.

6.1 Summary

The study began by examining the details of FRA Form 6180.57 and Form 6180.57, which contained inventory and crash information for HRGCs. Subsequently, three datasets were employed to identify data errors: the FRA crash database on HRGCs, the FRA inventory database on HRGCs, and the field-validated inventory database on HRGCs. Records of inventory and crash history for 560 public, at-grade, and operational HRGCs across nine counties in Nebraska were obtained from the FRA website. Upon data assessment, it was discovered that the FRA data contained several errors, such as illogical data entries and missing information. Further investigation through field validation of 560 HRGCs revealed that 5 (approx. 1%) HRGCs were either abandoned, non-operational, or non-existent in reality, despite being reported as operational in the FRA data. In addition, a total of 2,241 values were corrected, and 1,732 missing values were added, resulting in an average of 7.4% of the database values being altered at each HRGC. It was found that important inventory aspects, such as AADT, maximum timetable speed, warning device type light, count of flashing lights, gate arms, bells, surface type of main track, crossbucks, and storage distance, all had missing values. Additionally, it was observed that some inventory-related variables, such as AADT or highway speed, had impractical values that were incorrect.

To achieve the first objective, further data investigation was performed to ascertain whether the missing values in the inventory data occurred in a pattern or randomly. The MCAR test indicated that the data were not missing at random and exhibited a pattern. In addition, the descriptive statistics for validated inventory revealed that 74.11% of crossings in the study sample had a single track, 67.93% were on local streets (road-functional classification), 50% HRGCs had posted speed of 50mph, 60.71% HRGCs were in public spaces, 71.07% had no gates, and 49.55% of the HRGCs had gravel as an approach surface type.

To accomplish the third and fourth objective, the 2020 Accident Prediction and Severity Model (APS) was utilized with two different datasets: the unaltered HRGCs inventory data from FRA and the field-validated HRGCs inventory data. The crash predictions (expected crashes) and severity were compared between the two datasets through statistical analysis. A visual inspection of the crash and severity predictions for all 560 HRGCs was conducted by plotting the results of both datasets in the 2020 FRA crash prediction model. The analysis revealed differences between the two sets of predictions, with percentage differences ranging from 0 to 120%.

To determine if the differences in expected crashes were statistically significant, normality tests were conducted using Q-Q plots, histograms, and the Shapiro-Wilk test. Additionally, a non-parametric statistical test, the Wilcoxon rank-sum test, was used to compare the differences between the two crash predictions. The test results suggested that the medians of the two distributions of crash predictions differed, indicating differences in predictions. Moreover, the findings indicated that inaccurate data on HRGCs' inventory had a substantial impact on crash severity prediction for the study.

Additionally, hypothesis tests showed that the difference between FRA and field-validated crash severity predictions were statistically significant.

The fourth and fifth objective involved examining statistical differences between estimated parameters' coefficients of comparative crash prediction and severity models based on the FRA and field validated datasets. To thoroughly analyze the extent of a particular parameter's coefficient compared to others, identical parameters were included in the specifications of comparative crash frequency and severity models. Two Zero-Inflated Negative Binomial models (ZINB) were used to predict crash frequency at 560 selected HRGCs in Nebraska, using both unaltered (FRA-provided) and field-validated data. The ZINB model was chosen due to the significant over-dispersion with many zeros in the response variable "past-5 years crashes at HRGCs." To ensure the stability and reliability of the results, candidate variables for model estimation were tested for multicollinearity. The Akaike Information Criterion (AIC) and Bayesian Information Criterion (BIC) were used to evaluate the models' performance.

The ZINB models estimated that Ln-transformed Average Annual Daily Traffic (AADT) had a positive association with the likelihood of crashes at HRGCs, while flashing lights as a warning device (WDTLIT) had a negative association with predicted crashes. Moreover, Ln-transformed maximum timetable speed (IMTTSPD) had a positive association with crash prediction. In addition, the zero-inflated part of both crash prediction models revealed that Ln-transformed total daily trains (ITDTRAINS) had a negative association with the occurrence of zero counts for crash prediction. All coefficients in both models showed strong statistical significance, but the ZINB model based on field validated HRGCs inventory data had a better fit according to AIC and BIC

values. Finally, the estimated parameters' coefficients and their respective average marginal values were apparently different for when the models were based on unaltered FRA HRGCs inventory database and the corrected and complete (field validated) HRGCs inventory data. However, based on the hypothesis tests, the differences in the estimated parameters' coefficients of the ZINB model were not statistically significant.

Furthermore, two Ordered Probit models were utilized to make predictions about crash severity by incorporating the past 15 years' crash data (2007-2021) along with the corresponding inventory dataset of the HRGCs. The study then compared the field-validated data to the FRA dataset-based crash severity model to determine if there were any significant disparities in the estimated parameters' coefficients of the comparative crash prediction models. Both Ordered Probit models showed a statistically significant correlation between the severity of the crash at the HRGCs and the indicator variables for rural crossings (Rural_Xng), crossings within 500 ft of the road (Rdwithin_500ft), and flashing lights (Flashing_Lit). However, there were no statistically significant differences in the estimated parameters coefficients of the two Ordered Probit models. Furthermore, the severity model based on the field validation and complete inventory dataset demonstrated better model fit with smaller estimated standard errors for coefficient estimates of the models based on the field validated data.

6.2 Conclusions and Recommendations

The following conclusions are drawn from the research's findings.

1. The examined FRA HRGCs inventory data suffer from inaccuracies and incompleteness. While the incidence of errors and missing data is not high (roughly 7% based on the sample studied), it is important to note that such discrepancies can impact the allocation of resources, assessment of safety-risks, and decision making.

2. The crash frequency and severity predictions from the 2020 FRA APS models may be questionable as crash and severity predictions differed depending on the sample datasets used.

3. While the coefficients of estimated parameters of the crash frequency and severity prediction models were not statistically significantly different, the crash and severity models estimated using field-validated data exhibited a better fit. This implies more precise estimated model coefficients (smaller standard errors of estimated coefficients).

The following recommendations are made considering the findings to assist in raising the quality of the data on HRGCs inventory.

1. There is a need for an in-depth review of the entire FRA HRGCs inventory database. A comprehensive study is needed to identify the sources of existing inconsistencies, inaccuracies, and missing data in the FRA inventory as a step toward completeness.

2. The use of a corrected HRGCs inventory database in the 2020 FRA APS models will produce more reliable crash frequency and severity predictions.

3. Utilizing a corrected HRGCs inventory database would result in better-fitted crash and severity models with more precise estimated coefficients.

Managing inventory data and record-keeping is an arduous process that requires time, funding, and human resources. However, if more attention is paid to the accuracy of the data from the outset, the resulting predictive models and risk assessment will be more reliable and more useful for decisions regarding safety at HRGCs.

6.3 Research Limitations and Contribution

There are several limitations of this research as discussed below.

1. This research utilized inventory data specific to Nebraska, which may constrain the generalization of the research findings to the broader population of HRGCs across the United States due to differences in general topography as well as in motor vehicle drivers' behaviors.

2. The research relied on FRA crash data that met certain reporting thresholds, which could have led to underreporting.

3. The field validated data elements were limited to physical HRGCs characteristics with dynamic characteristics (e.g., AADT, daily train traffic) unverified.

4. The relatively small sample used for estimating the crash severity model consisted of only 83 crashes reported during 2007-2021.

5. The crash frequency and severity models were estimated based on a criterion of having similar significant covariates, which may have resulted in the exclusion of important variables in the models and resulted in unknown degree of misspecification in the estimated models.

6.4 Future Research

Some recommendations for future research are presented below.

1. As 2020 FRA model was based on (2016-2020) data, validation of the crash and severity predictions can be performed once five years of crash data (2021-2025) become available in the future.

2. Future study is needed to account for the effects of dynamic factors such as AADT, and train volume on crash and severity predictions.

3. A cost-benefit analysis of validation of the HRGCs inventory data will ensure that the safety benefits of proposed research recommendations outweigh the costs.

6.5 Research Contributions

The contributions of this research are listed below.

1. Using field-validated HRGCs data, inaccuracies, errors, and missing values and patterns were identified in the FRA inventory data.

2. Different crash frequencies and severity predictions were reported using the 2020 APS model and the two inventory databases.

3. Based on field-validated HRGCs inventory data, new crash and severity prediction models were estimated. This will provide guidelines to agencies and researchers for modeling crash and severity predictions based on complete datasets.

REFERENCES

Abay, K. A. Investigating the nature and impact of reporting bias in road crash data. *Transportation Research Part A: Policy and Practice*, Vol. 71, No. 1, 2015, pp. 31-45.

Abdel-Aty, M., and Keller, J. Exploring the overall and specific crash severity levels at signalized intersections. *Accident Analysis & Prevention*, Vol. 37, No. 3, 2005, pp. 417-425.

Alsop, J., and Langley, J. Under-reporting of motor vehicle traffic crash victims in New Zealand. *Accident Analysis & Prevention*, Vol. 33, No. 3, 2001, pp. 353-359.

Amoros, E., Martin, J. L., and Laumon, B. Under-reporting of road crash casualties in France. *Accident Analysis & Prevention*, Vol. 38, No. 4, 2006, pp. 627-635.

Anastasopoulos, P. C., and Mannering, F. L. A note on modeling vehicle accident frequencies with random-parameters count models. *Accident Analysis & Prevention*, Vol. 41, No. 1, 2009, pp. 153-159.

Anastasopoulos, P. C., Mannering, F. L., Shankar, V. N., and Haddock, J. E. A study of factors affecting highway accident rates using the random-parameters tobit model. *Accident Analysis & Prevention*, Vol. 45, 2012, pp. 628-633.

Anderson, M., Khattak, A. J., Farooq, M. U., Cecava, J., and Walker, C. Research on Weather Conditions and Their Relationship to Crashes (No. SPR-21 (20) M097). Nebraska Department of Transportation, 2020.

Austin, K. The identification of mistakes in road accident records: Part 1, locational variables. *Accident Analysis & Prevention*, Vol. 27, No. 2, 1995, pp. 261-276.

Barchard, K. A., and Pace, L. A. Preventing human error: The impact of data entry methods on data accuracy and statistical results. *Computers in Human Behavior*, Vol. 27, No. 5, 2011, pp. 1834-1839.

Beanland, V., Fitzharris, M., Young, K. L., and Lenné, M. G. Driver inattention and driver distraction in serious casualty crashes: Data from the Australian National Crash In-depth Study. *Accident Analysis & Prevention*, Vol. 54, 2013, pp. 99-107.

Bolukbasi, M., Mohammadi, J., and Arditi, D. Estimating the future condition of highway bridge components using national bridge inventory data. *Practice Periodical on Structural Design and Construction*, Vol. 9, No. 1, 2004, pp. 16-25.

Bozdogan, H. Model selection and Akaike's information criterion (AIC): The general theory and its analytical extensions. *Psychometrika*, Vol. 52, No. 3, 1987, pp. 345-370.

Breck, E., Polyzotis, N., Roy, S., Whang, S., and Zinkevich, M. Data Validation for Machine Learning. Presented at Conference on Systems and Machine Learning, Stanford, California, 2019.

Brod, D., Gillen, D., and Decisiontek, L. L. C. New Model for Highway-Rail Grade Crossing Accident Prediction and Severity (No. DOT/FRA/ORD-20/40). United States Department of Transportation. Federal Railroad Administration. Office of Research, Development, and Technology, 2020.

Brown, K., Sarasua, W. A., Ogle, J. H., Mammadrahimli, A., Chowdhury, M., Davis W. J., and Huynh, N. Assessment of Crash Location Improvements in Map-Based Geocoding Systems and Subsequent Benefits to Geospatial Crash Analysis (No. 15-5364). South Carolina Department of Transportation, 2015.

Burns, S., Miranda-Moreno, L., Stipancic, J., Saunier, N., and Ismail, K. Accessible and practical geocoding method for traffic collision record mapping: Quebec, Canada, case study. *Transportation Research Record: Journal of Transportation Research Board*. Vol. 2460, 2014 pp. 39-46.

Caddell, R., Hammond P., and Reinmuth S.. Roadside Features Inventory Program. http://legacy.ahmct.ucdavis.edu/wp-content/uploads/pdf/ahmct_roadside_inventory_05-2007.pdf. Accessed June 18, 2007.

Chang, L. Y., and Wang, H. W. Analysis of traffic injury severity: An application of non-parametric classification tree techniques. *Accident Analysis & Prevention*, Vol. 38, No. 5, 2006, pp. 1019-1027.

Chapman, A. D. Principles of data quality. https://www.gbif.org/document/80534/principles-of-data-quality. Global Core Biodata Resource, Accessed June 3, 2005.

Chatterjee, S., and Hadi, A. S. Regression analysis by example. John Wiley & Sons, 2006.

Chen, W. H., and Jovanis, P. P. Method for identifying factors contributing to driver-injury severity in traffic crashes. *Transportation Research Record: Journal of Transportation Research Board*, Vol. 1717, 2000, pp. 1-9.

Clogg, C. C., Petkova, E., and Haritou, A. Statistical methods for comparing regression coefficients between models. *American Journal of Sociology*, Vol. 100, No. 5, 1995, pp. 1261-1293.

Couto, A., Amorim, M., and Ferreira, S. Reporting Road victims: assessing and correcting data issues through distinct injury scales. *Journal of Safety Research*, Vol. 57, 2016., pp. 39-45.

Cummings, P. Association of seat belt use with death: a comparison of estimates based on data from police and estimates based on data from trained crash investigators. *Injury Prevention*, Vol. 8, No. 4, 2002, pp. 338-341.

Das, S., Kong, X., Lavrenz, S. M., Wu, L., and Jalayer, M. Fatal crashes at highway rail grade crossings: A US based study. *International Journal of Transportation Science and Technology*, Vol. 62, 2021, pp. 34-54.

Das, S., Kong, X., Lavrenz, S. M., Wu, L., and Jalayer, M. Fatal crashes at highway rail grade crossings: A US based study. *International Journal of Transportation Science and Technology*, Vol. 11, No. 1, 2022, pp. 107-117.

Daykin, A. R., and Moffatt, P. G. Analyzing ordered responses: A review of the ordered probit model. *Understanding Statistics: Statistical Issues in Psychology, Education, and the Social Sciences*, Vol. 1, No. 3, 2002, pp. 157-166.

Delacre, M., Lakens, D., and Leys, C. Why Psychologists Should by Default Use Welch's t-test Instead of Student's t-test. *International Review of Social Psychology*, Vol. 35, No.1, 2022, pp. 24-36.

Dong, C., Clarke, D. B., Yan, X., Khattak, A., and Huang, B. Multivariate random-parameters zero-inflated negative binomial regression model: An application to estimate crash frequencies at intersections. *Accident Analysis & Prevention*, Vol. 70, 2014, pp. 320-329.

Eluru, N., Bagheri, M., Miranda-Moreno, L. F., and Fu, L. A latent class modeling approach for identifying vehicle driver injury severity factors at highway-railway crossings. *Accident Analysis & Prevention*, Vol. 47, 2012, pp. 119-127.

Eluru, N., Bhat, C. R., and Hensher, D. A. A mixed generalized ordered response model for examining pedestrian and bicyclist injury severity level in traffic crashes. *Accident Analysis & Prevention*, Vol. 40, No. 3, 2008, pp. 1033-1054.

Erkut, E., Tjandra, S. A., and Verter, V. Hazardous materials transportation. *Handbooks in Operations Research and Management Science*, Vol. 14, 2007, pp. 539-621.

Fan, W., Kane, M. R., and Haile, E. Analyzing severity of vehicle crashes at highway-rail grade crossings: multinomial logit modeling. *Journal of the Transportation Research Forum*, Vol. 54, No. 1424-2016-118071, 2015, pp. 39-56.

Farooq, M. U., and A. Khattak. A Heterogeneity-Based Temporal Stability Assessment of Pedestrian Crash Injury Severity Using an Aggregated Crash and Hospital Dataset. Presented at Transportation Research Board (TRB) 102nd Annual Meeting, Washington DC, 2023.

Farooq, M. U., Ahmed, A., and Saeed, T. U. A statistical analysis of the correlates of compliance and defiance of seatbelt use. *Transportation Research Part F: Traffic Psychology and Behaviour*, Vol. 77, 2021, pp. 117-128.

Farooq, M. U., and Ahmad, A. An Analysis of Legislation and Level of Compliance of Key Crash Risk Factors-A Case Study of Islamabad. National University of Science & Technology Islamabad, Pakistan, 2017.

Farooq, M. U., Ghani, M. U., and Ahsan, Z. Survey on Driving Behavior and Motivational Factors Causing Aggressive Driving: A Case Study of Peshawar, Pakistan. Presented at 2nd International Conference on Emerging trends in Engineering, Management & Sciences (ICETEMS-2016), ISBN No: 978-969-23044-2-9, Peshawar, 2016.

Fatality Analysis Reporting System (FARS). https://www-fars.nhtsa.dot.gov/Main/index.aspx. Accessed: July 31, 2020.

Federal Railroad Administration (FRA). Crash and Inventory Stats on Grade Crossings. https://railroads.dot.gov/safety-data. Accessed January 1, 2022.

Federal Railroad Administration (FRA). FRA Risk Reduction Program. https://railroads.dot.gov/newsroom/press-releases/fra-launches-risk-reduction-safety-program-0 (202). Accessed Feb 5, 2020.

Federal Railroad Administration (FRA). https://railroads.dot.gov/rail-network-development/passenger-rail/system-safety-program (2019). Accessed October 17, 2019.

Federal Railroad Administration (FRA). FRA Instructions for Electronic Submission of U.S. DOT Crossing Inventory Data Grade Crossing Inventory System (GCIS), v2.9.0.0, https://railroads.dot.gov/forms-guides-publications/guides/fra-instructions-electronic-submission-us-dot-crossing-inventory. Accessed February 7, 2019.

Federal Railroad Administration (FRA). FRA Guide for Preparing Accidents/incidents Reports. https://railroads.dot.gov/sites/fra.dot.gov/files/2019-09/FRAGuideforPreparingAccIncReportspubMay2011.pdf, Accessed May 32, 2011.

Federal Railroad Administration (FRA). Federal railroad administration guide for preparing U.S. DOT crossing inventory forms. https://railroads.dot.gov/sites/fra.dot.gov/files/fra_net/18855/Crossing_Inventory_Guide_01916.pdf. Accessed November 12, 2016.

Shi, F. Learn About Ordered Probit Regression in R With Data from the General Social Survey. https://methods.sagepub.com/dataset/oprobit-in-gss-2016. Accessed January 14, 2019.

Fischhaber, P. M. Development of light rail crossing specific crash prediction models. Doctoral Dissertation, University of Colorado at Denver, 2014.

Federal Railroad Administration (FRA). Crash and Inventory data on Grade Crossings. https://railroads.dot.gov/safety-data. Accessed June 6, 2021.

Gabree, S., Chase, S., and DaSilva, M. Effect of dynamic envelope pavement markings on vehicle driver behavior at a highway-rail grade crossing. Presented at ASME/IEEE Joint Rail Conference, Washington DC, 2014.

Haleem, K., and Gan, A. Contributing factors of crash injury severity at public highway-railroad grade crossings in the US. *Journal of Safety Research*, Vol. 53, 2015, pp. 23-29.

Haleem, K., and Abdel-Aty, M. Examining traffic crash injury severity at unsignalized intersections. *Journal of Safety Research*, Vol. 41, No. 4, 2010, pp. 347-357.

Hao, W., and Daniel, J. R. Severity of injuries to motor vehicle drivers at highway–rail grade crossings in the United States. *Transportation Research Record: Journal of Transportation Research Board*, Vol. 2384, No. 1, 2013, pp. 102-108.

Hao, W., and Kamga, C. Difference in rural and urban driver-injury severities in highway–rail grade crossing accidents. *International Journal of Injury Control and Safety Promotion*, Vol. 24, No. 2, 2017, pp. 174-182.

Haque, M. M., and Sangster, J. D. Best Practices for Modeling Light Rail at Intersections (No. 26-1121-0018-009). University Transportation Center for Railway Safety, 2018.

Haque, M. S., Zhao, L., Rilett, L. R., and Tufuor, E. O. Calibration and validation of a microsimulation model of lane closures on a two-lane highway work zone. *Transportation Research Record: Journal of Transportation Research Board*, Vol. 2677, No. 3, 2023a, pp. 974-990.

Haque, M. S. Improved traffic analysis methodology for lane closure on two-lane highway work zone. Doctoral dissertation, The University of Nebraska-Lincoln, 2022.

Haque, M. S., Rilett, L. R., and Zhao, L. Impact of Platooning Connected and Automated Heavy Vehicles on Interstate Freeway Work Zone Operations. *Journal of Transportation Engineering, Part A: Systems*, Vol. 149, No. 3, 2023b, pp. (04022160) 1-13.

Helmreich, R. L. On error management: lessons from aviation. *Journal of the BMJ*, Vol. 320, No. 7237, 2000, pp. 781-785.

Hezaveh, A. M., and Cherry, C. R. A New Approach to Aggregate Crash Prediction Model: Home-Based Crash Frequency Modeling. Presented at Southern District ITE Annual Meeting, Mobile, AL, 2018.

Hu, S. R., Li, C. S., and Lee, C. K. Investigation of key factors for accident severity at railroad grade crossings by using a logit model. *Safety science*, Vol. 48, No. 2, 2010, pp. 186-194.

Huh, Y. U., Keller, F. R., Redman, T. C., and Watkins, A. R. Data quality. *Information and Software Technology*, Vol. 32, No. 8, 1990, pp. 559-565.

Imprialou, M. I. M., Quddus, M., and Pitfield, D. E. Multilevel logistic regression modeling for crash mapping in metropolitan areas. *Transportation Research Record: Journal of Transportation Research Board*, Vol. 2514, No. 1, 2015, pp. 39-47.

Imprialou, M., and Quddus, M. Crash data quality for road safety research: current state and future directions. *Accident Analysis & Prevention*, Vol. 130, 2019, pp. 84-90.

Islam, S., and Mannering, F. Driver aging and its effect on male and female single-vehicle accident injuries: Some additional evidence. *Journal of Safety Research*, Vol. 37, No. 3, 2006, pp. 267-276.

Jäckle, A. Measurement error and data collection methods: Effects on estimates from event history data (No. 2008-13). Institute for Social and Economic Research (ISER) Working Paper Series, 2008.

Jalayer, M., Zhou, H., Gong, J., Hu, S., and Grinter, M. A comprehensive assessment of highway inventory data collection methods. *Journal of the Transportation Research Forum*, Vol. 53, No. 1424-2016-117955, 2014, pp. 73-92.

Kaiser, J. Dealing with Missing Values in Data. *Journal of Systems Integration*, Vol. 5, No. 1, 2014, 12-34.

Keramati, A., Lu, P., Tolliver, D., and Wang, X. Geometric effect analysis of highway-rail grade crossing safety performance. *Accident Analysis & Prevention*, Vol. 138, No. 105470, 2020, pp. 1-14.

Khan, I. U., Lee, E., and Khan, M. A. Developing a highway rail grade crossing accident probability prediction model: a North Dakota case study. *Safety*, Vol. 4, No. 2, 2018, pp. 22-32.

Khattak, A., and M. U. Farooq. The Effects of Inaccurate and Missing Highway-Rail Grade Crossing Inventory Data on Crash Model Estimation and Crash Prediction. Presented at Transportation Research Board (TRB) 102nd Annual Meeting, Washington DC, 2023.

Khattak, A., M. U. Farooq, and A. Farhan. Motor Vehicle Drivers' Knowledge of Safely Traversing Highway-Rail Grade Crossings. Presented at Transportation Research Board (TRB) 102nd Annual Meeting, Washington DC, 2023.

Khattak, A., Kang, Y., and Liu, H. Nebraska Rail Crossing Safety Research (No. SPR-P1 (19) M091). Nebraska Department of Transportation, 2020.

Khattak, A., and Thompson, E. Development of a Methodology for Assessment of Crash Costs at Highway-Rail Grade Crossings in Nebraska (No. 25-1121-0001-422). Mid-America Transportation Center, 2012.

Khattak, A., A. Sharma, and Z. Luo. Implications of using annual average daily traffic in highway-rail grade crossing safety models. Presented at Transportation Research Board (TRB) 91st Annual Meeting, Washington DC, 2012.

Kim, K., and Yamashita, E. Y. Using ak-means clustering algorithm to examine patterns of pedestrian involved crashes in Honolulu, Hawaii. *Journal of Advanced Transportation*, Vol. 41, No. 1, 2007, pp. 69-89.

Klein, B. D., Goodhue, D. L., and Davis, G. B. Can humans detect errors in data? Impact of base rates, incentives, and goals. Management Information Systems Research Center, University of Minnesota, 1997.

Kockelman, K. M., and Kweon, Y. J. Driver injury severity: an application of ordered probit models. *Accident Analysis & Prevention*, Vol. 34, No. 3, 2002, pp. 313-321.

Landry, S., Jeon, M., Lautala, P., and Nelson, D. Getting active with passive crossings: Investigating the use of in-vehicle auditory alerts for highway-rail grade crossings. Presented at ASME/IEEE Joint Rail Conference, Columbia, South Carolina, 2016.

Lao, Y., Wu, Y. J., Corey, J., and Wang, Y. Modeling animal-vehicle collisions using diagonal inflated bivariate Poisson regression. *Accident Analysis & Prevention*, Vol. 43, No. 1, 2011, pp. 220-227.

Little, R. J. A test of missing completely at random for multivariate data with missing values. *Journal of the American Statistical Association*, Vol. 83, No. 404, 1988, pp. 1198-1202.

Liu, C., Zhao, M., Li, W., and Sharma, A. Multivariate random parameters zero-inflated negative binomial regression for analyzing urban midblock crashes. *Analytic Methods in Accident Research*, Vol. 17, 2018, pp. 32-46.

Loo, B. P. Validating crash locations for quantitative spatial analysis: a GIS-based approach. *Accident Analysis & Prevention*, Vol. 38, No. 5, 2006, 879-886.

Lord, D. Modeling motor vehicle crashes using Poisson-gamma models: Examining the effects of low sample mean values and small sample size on the estimation of the fixed dispersion parameter. *Accident Analysis & Prevention*, Vol. 38, 2006, pp. 751–766.

Lord, D., and Mannering, F. The statistical analysis of crash-frequency data: A review and assessment of methodological alternatives. *Transportation Research Part A: Policy and Practice*, Vol. 44, No. 5, 2010, pp. 291-305.

Lord, D., Washington, S., and Ivan, J. N. Further notes on the application of zero-inflated models in highway safety. *Accident Analysis & Prevention*, Vol. 39, No. 1, 2007, pp. 53-57.

Lu, P., and Tolliver, D. Accident prediction model for public highway-rail grade crossings. *Accident Analysis & Prevention*, Vol. 90, 2016, pp. 73-81.

Lu, P., Zheng, Z., Ren, Y., Zhou, X., Keramati, A., Tolliver, D., and Huang, Y. A gradient boosting crash prediction approach for highway-rail grade crossing crash analysis. *Journal of Advanced Transportation*, Vol. 34, 2020, pp. 1-10.

Malyshkina, N. V., and Mannering, F. L. Empirical assessment of the impact of highway design exceptions on the frequency and severity of vehicle accidents. *Accident Analysis & Prevention*, Vol. 42, No. 1, 2010, pp. 131-139.

Mathew, J., and Benekohal, R. F. A new accident prediction model for highway-rail grade crossings using the USDOT formula variables. *Journal of Traffic and Transportation Engineering*, Vol. 8, 2020, pp. 1-13.

Mathew, J., and Benekohal, R. F. Highway-rail grade crossings accident prediction using Zero Inflated Negative Binomial and Empirical Bayes method. *Journal of Safety Research*, Vol. 79, 2021, pp. 211-236.

Miaou, S. P. The relationship between truck accidents and geometric design of road sections: Poisson versus negative binomial regressions. *Accident Analysis & Prevention*, Vol. 26, No. 4, 1994, pp. 471-482.

Miler, M., Todić, F., and Ševrović, M. Extracting accurate location information from a highly inaccurate traffic accident dataset: A methodology based on a string-matching technique. *Transportation Research Part C: Emerging Technologies*, Vol. 68, 2016, pp. 185-193.

Moving Ahead for Progress in the 21st Century (MAP-21) http://www.fhwa.dot.gov/map21/ (2014). Accessed March 2, 2014.

Myung, I. J. Tutorial on maximum likelihood estimation. *Journal of Mathematical Psychology*, Vol. 47, No. 1, 2003, pp. 90-100.

Nam, D., Lee, J. Accident frequency model using zero probability process. Transportation Research Record: Journal of Transportation Research Board, Vol. 1973, No. 1, 2006, pp. 142-148.

NHTSA. Report to Congress NHTSA's Crash Data Collection Programs 2010. https://crashstats.nhtsa.dot.gov/Api/Public/ViewPublication/811337. Accessed April 24. 2010.

Oh, J., Washington, S. P., and Nam, D. Accident prediction model for railway-highway interfaces. *Accident Analysis & Prevention*, Vol. 38, No. 2, 2006, pp. 346-356.

Oliveira, P., Rodrigues, F., and Henriques, P. R. A formal definition of data quality problems. Presented at International Conference of Qualitative Inquiry, 2005.

Park, B. J., and Lord, D. Application of finite mixture models for vehicle crash data analysis. *Accident Analysis & Prevention*, Vol. 41, No. 4, 2009, pp. 683-691.

Park, K., Kim, J. M., and Jung, D. GLM-based statistical control r-charts for dispersed count data with multicollinearity between input variables. *Quality and Reliability Engineering International*, Vol. 34, No. 6, 2018, pp. 1103-1109.

Pasha, J., Dulebenets, M. A., Abioye, O. F., Kavoosi, M., Moses, R., Sobanjo, J., and Ozguven, E. E. A comprehensive assessment of the existing accident and hazard prediction models for the highway-rail grade crossings in the state of Florida. *Sustainability*, Vol. 12, No. 10, 2020, pp. 42-56.

Raihan, M. A., Alluri, P., Wu, W., and Gan, A. Estimation of bicycle crash modification factors (CMFs) on urban facilities using zero inflated negative binomial models. *Accident Analysis & Prevention*, Vol. 123, 2019, pp. 303-313.

Raub, R. A. Examination of highway–rail grade crossing collisions nationally from 1998 to 2007. *Transportation Research Record: Journal of Transportation Research Board*, Vol. 2122, No. 1, 2009, pp. 63-71.

Ries, R. Highway-rail grade crossing safety. *Mass Transit*, Vol. 33, No. 5, 2007, pp. 45-61.

Saccomanno, F. F., Ren, C., and Fu, L. Collision Prediction models for Highway-Rail Grade Crossings in Canada. Presented at Transportation Research Board (TRB) 82nd Annual Meeting, Washington DC, 2003.

Salmon, P. M., Lenné, M. G., Young, K. L., and Walker, G. H. An on-road network analysis-based approach to studying driver situation awareness at rail level crossings. *Accident Analysis & Prevention*, Vol. 58, 2013, pp. 195-205.

Savolainen, P. T., Mannering, F. L., Lord, D., and Quddus, M. A. The statistical analysis of highway crash-injury severities: A review and assessment of methodological alternatives. *Accident Analysis & Prevention*, Vol. 43, No. 5, 2011, pp. 1666-1676.

Senders, J. W., and Moray, N. P. Human Error: Cause, Prediction, and Reduction. ISBN: 9781000149043,1000149048. *CRC Press*. 2020.

Sharma, A. K., and Landge, V. S. Zero inflated negative binomial for modeling heavy vehicle crash rate on Indian rural highway. *International Journal of Advances in Engineering & Technology*, Vol. 5, No. 2, 2013, pp. 292-311.

Sharma, S., and Pulugurtha, S. S. Modeling crash risk at rail-highway grade crossings by track class. *Journal of Transportation Technologies*, Vol. 9, No. 03, 2019, pp. 261-274.

Souleyrette, R., Stout, T., and Williams, T. Crash Data Validation: An Iowa Case Study (No. CTRE Project 06-256), Iowa Department of Transportation, 2007.

Srinivasan, K. K. Injury severity analysis with variable and correlated thresholds: ordered mixed logit formulation. *Transportation Research Record: Journal of Transportation Research Board*, Vol. 1784, No. 1, 2002, pp. 132-141.

Stanton, N. A., and Salmon, P. M. Human error taxonomies applied to driving: A generic driver error taxonomy and its implications for intelligent transport systems. *Safety Science*, Vol. 47, No. 2, 2009, pp. 227-237.

Suer, M. What is Data Quality and Why is it important? https://www.alation.com/blog/what-is-data-quality-why-is-it-important/#:~:text=Why%20Is%20it%20Important%20to,quality%20versus%20low%2Dquality). Accessed January 2, 2021.

Tayi, G. K., and Ballou, D. P. Examining data quality. *Communications of the ACM*, Vol. 41, No. 2, 1998, pp. 54-57.

U.S. Department of Transportation. Federal-Aid Policy Guide: Part 924 - Highway Safety Improvement Program. 1991. https://www.fhwa.dot.gov/legsregs/directives/fapg/cfr0924.htm. Accessed October 24, 2018.

Veregin, H. Data quality parameters. *Geographical Information Systems*, Vol. 1, 1998, pp. 177-189.

Wang, R. Y., Ziad, M., Lee, and Y. W. Data quality. Springer Science & Business Media, 2006.

Wang, X., and Abdel-Aty, M. Analysis of left-turn crash injury severity by conflicting pattern using partial proportional odds models. *Accident Analysis & Prevention*, Vol. 40, No. 5, 2008a, pp. 1674-1682.

Wang, X., and Abdel-Aty, M. Modeling left-turn crash occurrence at signalized intersections by conflicting patterns. *Accident Analysis & Prevention*, Vol. 40, No. 1, 2008b, pp.76-88.

Wang, Z., Chen, H., and Lu, J. J. Exploring impacts of factors contributing to injury severity at freeway diverge areas. *Transportation Research Record: Journal of Transportation Research Board*, Vol. 2102, No. 1, 2008b, pp. 43-52.

Watson, A., Watson, B., and Vallmuur, K. How accurate is the identification of serious traffic injuries by Police? The concordance between Police and hospital reported traffic injuries. Presented at Australasian Road Safety Research, Policing and Education Conference, Sydney, 2013.

Watson, A., Watson, B., and Vallmuur, K. Estimating under-reporting of road crash injuries to police using multiple linked data collections. *Accident Analysis & Prevention*, Vol. 83, 2015, pp. 18-25.

Welch, B. L. The significance of the difference between two means when the population variances are unequal. *Biometrika*, Vol. 29, No. 4, 1938, pp. 350-362.

Westerlund, J. Testing for error correction in panel data. *Oxford Bulletin of Economics and Statistics*, Vol. 69, No. 6, 2007, pp. 709-748.

Xie, Y., and Zhang, Y. Crash frequency analysis with generalized additive models. *Transportation Research Record: Journal of the Transportation Research Board*, 2008, pp. 39–45.

Yan, X., Richards, S., and Su, X. Using hierarchical tree-based regression model to predict train–vehicle crashes at passive highway-rail grade crossings. *Accident Analysis & Prevention*, Vol. 42, No. 1, 2010, 64-74.

Yap, B. W., and Sim, C. H. Comparisons of various types of normality tests. *Journal of Statistical Computation and Simulation*, Vol. 81, No. 12, 2011, pp. 2141-2155.

Yau, K. K., Wang, K., and Lee, A. H. Zero-inflated negative binomial mixed regression modeling of over-dispersed count data with extra zeros. *Biometrical Journal: Journal of Mathematical Methods in Biosciences*, Vol. 45, No. 4, 2003, pp.437-452.

Young, R. K., and Liesman, J. Estimating the relationship between measured wind speed and overturning truck crashes using a binary logit model. *Accident Analysis & Prevention*, Vol. 39, No.3, 2007, pp. 574-580.

Zhao, S., Iranitalab, A., and Khattak, A. J. A clustering approach to injury severity in pedestrian-train crashes at highway-rail grade crossings. *Journal of Transportation Safety & Security*, Vol. 11, No.3, 2019, pp. 305-322.

Zhao, S., and Khattak, A. J. Injury severity in crashes reported in proximity of rail crossings: The role of driver inattention. *Journal of Transportation Safety and Security*, Vol. 10, No. 6, 2018, pp. 507-524.

Zheng, Z., Lu, P., and Pan, D. Predicting highway–rail grade crossing collision risk by neural network systems. *Journal of Transportation Engineering, Part A: Systems*, Vol. 145, No. 8, 2019, pp. 80-96.

APPENDIX A U.S. DOT CROSSING INVENTORY FORM

U. S. DOT CROSSING INVENTORY FORM

DEPARTMENT OF TRANSPORTATION
FEDERAL RAILROAD ADMINISTRATION

OMB No. 2130-0017

Instructions for the initial reporting of the following types of new or previously unreported crossings: For public highway-rail grade crossings, complete the entire Inventory Form. For private highway-rail grade crossings, complete the Header, Parts I and II, and the Submission Information section. For public pathway grade crossings (including pedestrian station grade crossings), complete the Header, Parts I and II, and the Submission Information section. For Private pathway grade crossings, complete the Header, Parts I and II, and the Submission Information section. For grade-separated highway-rail or pathway crossings (including pedestrian station crossings), complete the Header, Part I, and the Submission Information section. For changes to existing data, complete the Header, Part I Items 1-3, and the Submission Information section, in addition to the updated data fields. Note: For private crossings only, Part I Item 20 and Part III Item 2.K. are required unless otherwise noted. An asterisk * denotes an optional field.

A. Revision Date *(MM/DD/YYYY)* _____/_____/_____	B. Reporting Agency ☐ Railroad ☐ Transit ☐ State ☐ Other	C. Reason for Update *(Select only one)* ☐ Change in Data ☐ Re-Open ☐ New Crossing ☐ Date Change Only ☐ Closed ☐ Change in Primary Operating RR ☐ No Train Traffic ☐ Admin. Correction ☐ Quiet Zone Update	D. DOT Crossing Inventory Number

Part I: Location and Classification Information

1. Primary Operating Railroad	2. State	3. County

4. City / Municipality ☐ In ☐ Near	5. Street/Road Name & Block Number *(Street/Road Name)* | ** (Block Number)*	6. Highway Type & No.

7. Do Other Railroads Operate a Separate Track at Crossing? ☐ Yes ☐ No If Yes, Specify RR	8. Do Other Railroads Operate Over Your Track at Crossing? ☐ Yes ☐ No If Yes, Specify RR

9. Railroad Division or Region ☐ None	10. Railroad Subdivision or District ☐ None	11. Branch or Line Name ☐ None	12. RR Milepost *(prefix) | (nnnn.nnn) | (suffix)*

13. Line Segment *	14. Nearest RR Timetable Station *	15. Parent RR *(if applicable)* ☐ N/A	16. Crossing Owner *(if applicable)* ☐ N/A

17. Crossing Type ☐ Public ☐ Private	18. Crossing Purpose ☐ Highway ☐ Pathway, Ped. ☐ Station, Ped.	19. Crossing Position ☐ At Grade ☐ RR Under ☐ RR Over	20. Public Access *(If Private Crossing)* ☐ Yes ☐ No	21. Type of Train ☐ Freight ☐ Intercity Passenger ☐ Commuter ☐ Transit ☐ Shared Use Transit ☐ Tourist/Other	22. Average Passenger Train Count Per Day ☐ Less Than One Per Day ☐ Number Per Day ____

23. Type of Land Use
☐ Open Space ☐ Farm ☐ Residential ☐ Commercial ☐ Industrial ☐ Institutional ☐ Recreational ☐ RR Yard

24. Is there an Adjacent Crossing with a Separate Number? ☐ Yes ☐ No If Yes, Provide Crossing Number	25. Quiet Zone *(FRA provided)* ☐ No ☐ 24 Hr ☐ Partial ☐ Chicago Excused Date Established

26. HSR Corridor ID ☐ N/A	27. Latitude in decimal degrees *(WGS84 std: nn.nnnnnnn)*	28. Longitude in decimal degrees *(WGS84 std: -nnn.nnnnnnn)*	29. Lat/Long Source ☐ Actual ☐ Estimated

30.A. Railroad Use *	31.A. State Use *
30.B. Railroad Use *	31.B. State Use *
30.C. Railroad Use *	31.C. State Use *
30.D. Railroad Use *	31.D. State Use *
32.A. Narrative *(Railroad Use)* *	32.B. Narrative *(State Use)* *

33. Emergency Notification Telephone No. *(posted)*	34. Railroad Contact *(Telephone No.)*	35. State Contact *(Telephone No.)*

Part II: Railroad Information

1. Estimated Number of Daily Train Movements

1.A. Total Day Thru Trains *(6 AM to 6 PM)*	1.B. Total Night Thru Trains *(6 PM to 6 AM)*	1.C. Total Switching Trains	1.D. Total Transit Trains	1.E. Check if Less Than One Movement Per Day ☐ How many trains per week? ____

2. Year of Train Count Data *(YYYY)*	3. Speed of Train at Crossing 3.A. Maximum Timetable Speed *(mph)* _____ 3.B. Typical Speed Range Over Crossing *(mph)* From _____ to _____

4. Type and Count of Tracks

Main _____ Siding _____ Yard _____ Transit _____ Industry _____

5. Train Detection *(Main Track only)*
☐ Constant Warning Time ☐ Motion Detection ☐AFO ☐ PTC ☐ DC ☐ Other ☐ None

6. Is Track Signaled? ☐ Yes ☐ No	7.A. Event Recorder ☐ Yes ☐ No	7.B. Remote Health Monitoring ☐ Yes ☐ No

U. S. DOT CROSSING INVENTORY FORM

A. Revision Date *(MM/DD/YYYY)* | PAGE 2 | **D. Crossing Inventory Number** *(7 char.)*

Part III: Highway or Pathway Traffic Control Device Information

1. Are there Signs or Signals? ☐ Yes ☐ No

2. Types of Passive Traffic Control Devices associated with the Crossing

2.A. Crossbuck Assemblies *(count)* | 2.B. STOP Signs *(R1-1)* *(count)* | 2.C. YIELD Signs *(R1-2)* *(count)*

2.D. Advance Warning Signs *(Check all that apply; include count)* ☐ None
☐ W10-1 _____ ☐ W10-3 _____ ☐ W10-11 _____
☐ W10-2 _____ ☐ W10-4 _____ ☐ W10-12 _____

2.E. Low Ground Clearance Sign *(W10-5)*
☐ Yes *(count_____)*
☐ No

2.F. Pavement Markings
☐ Stop Lines ☐ Dynamic Envelope
☐ RR Xing Symbols ☐ None

2.G. Channelization Devices/Medians
☐ All Approaches ☐ Median
☐ One Approach ☐ None

2.H. EXEMPT Sign *(R15-3)*
☐ Yes
☐ No

2.I. ENS Sign *(I-13)* Displayed
☐ Yes
☐ No

2.J. Other MUTCD Signs ☐ Yes ☐ No
Specify Type _____ Count _____
Specify Type _____ Count _____
Specify Type _____ Count _____

2.K. Private Crossing Signs *(if private)*
☐ Yes ☐ No

2.L. LED Enhanced Signs *(List types)*

3. Types of Train Activated Warning Devices at the Grade Crossing ***(specify count of each device for all that apply)***

3.A. Gate Arms *(count)*
Roadway _____
Pedestrian _____

3.B. Gate Configuration
☐ 2 Quad ☐ Full *(Barrier)* Resistance
☐ 3 Quad ☐ Median Gates
☐ 4 Quad

3.C. Cantilevered *(or Bridged)* Flashing Light Structures *(count)*
Over Traffic Lane _____ ☐ Incandescent
Not Over Traffic Lane _____ ☐ LED

3.D. Mast Mounted Flashing Lights *(count of masts)* _____
☐ Incandescent ☐ LED
☐ Back Lights Included ☐ Side Lights Included

3.E. Total Count of Flashing Light Pairs

3.F. Installation Date of Current Active Warning Devices: *(MM/YYYY)*
_____/_____ ☐ Not Required

3.G. Wayside Horn
☐ Yes Installed on *(MM/YYYY)* _____/_____
☐ No

3.H. Highway Traffic Signals Controlling Crossing
☐ Yes ☐ No

3.I. Bells *(count)*

3.J. Non-Train Active Warning
☐ Flagging/Flagman ☐ Manually Operated Signals ☐ Watchman ☐ Floodlighting ☐ None

3.K. Other Flashing Lights or Warning Devices
Count _____ Specify type _____

4.A. Does nearby Hwy Intersection have Traffic Signals?
☐ Yes ☐ No

4.B. Hwy Traffic Signal Interconnection
☐ Not Interconnected
☐ For Traffic Signals
☐ For Warning Signs

4.C. Hwy Traffic Signal Preemption
☐ Simultaneous
☐ Advance

5. Highway Traffic Pre-Signals
☐ Yes ☐ No
Storage Distance * _____
Stop Line Distance * _____

6. Highway Monitoring Devices *(Check all that apply)*
☐ Yes - Photo/Video Recording
☐ Yes – Vehicle Presence Detection
☐ None

Part IV: Physical Characteristics

1. Traffic Lanes Crossing Railroad
Number of Lanes _____
☐ One-way Traffic
☐ Two-way Traffic
☐ Divided Traffic

2. Is Roadway/Pathway Paved?
☐ Yes ☐ No

3. Does Track Run Down a Street?
☐ Yes ☐ No

4. Is Crossing Illuminated? *(Street lights within approx. 50 feet from nearest rail)* ☐ Yes ☐ No

5. Crossing Surface *(on Main Track, multiple types allowed)* Installation Date * *(MM/YYYY)* _____/_____ Width * _____ Length * _____
☐ 1 Timber ☐ 2 Asphalt ☐ 3 Asphalt and Timber ☐ 4 Concrete ☐ 5 Concrete and Rubber ☐ 6 Rubber ☐ 7 Metal
☐ 8 Unconsolidated ☐ 9 Composite ☐ 10 Other *(specify)* _____

6. Intersecting Roadway within 500 feet?
☐ Yes ☐ No If Yes, Approximate Distance *(feet)* _____

7. Smallest Crossing Angle
☐ 0° – 29° ☐ 30° – 59° ☐ 60° - 90°

8. Is Commercial Power Available? *
☐ Yes ☐ No

Part V: Public Highway Information

1. Highway System
☐ (01) Interstate Highway System
☐ (02) Other Nat Hwy System (NHS)
☐ (03) Federal AID, Not NHS
☐ (08) Non-Federal Aid

2. Functional Classification of Road at Crossing
☐ (0) Rural ☐ (1) Urban
☐ (1) Interstate ☐ (5) Major Collector
☐ (2) Other Freeways and Expressways
☐ (3) Other Principal Arterial ☐ (6) Minor Collector
☐ (4) Minor Arterial ☐ (7) Local

3. Is Crossing on State Highway System?
☐ Yes ☐ No

4. Highway Speed Limit _____ MPH
☐ Posted ☐ Statutory

5. Linear Referencing System *(LRS Route ID)* *

6. LRS Milepost *

7. Annual Average Daily Traffic *(AADT)*
Year _____ AADT _____

8. Estimated Percent Trucks
_____ %

9. Regularly Used by School Buses?
☐ Yes ☐ No Average Number per Day _____

10. Emergency Services Route
☐ Yes ☐ No

Submission Information - *This information is used for administrative purposes and is not available on the public website.*

Submitted by _____ Organization _____ Phone _____ Date _____

Public reporting burden for this information collection is estimated to average 30 minutes per response, including the time for reviewing instructions, searching existing data sources, gathering and maintaining the data needed and completing and reviewing the collection of information. According to the Paperwork Reduction Act of 1995, a federal agency may not conduct or sponsor, and a person is not required to, nor shall a person be subject to a penalty for failure to comply with, a collection of information unless it displays a currently valid OMB control number. The valid OMB control number for information collection is 2130-0017. Send comments regarding this burden estimate or any other aspect of this collection, including for reducing this burden to: Information Collection Officer, Federal Railroad Administration, 1200 New Jersey Ave. SE, MS-25 Washington, DC 20590.

APPENDIX B U.S. DOT CROSSING CRASH/INCIDENT FORM

DEPARTMENT OF TRANSPORTATION
FEDERAL RAILROAD ADMINISTRATION (FRA)

HIGHWAY-RAIL GRADE CROSSING ACCIDENT/INCIDENT REPORT

OMB No. 2130-0500

	1a. Alphabetic Code	1b. Railroad Accident/Incident No.
2. Name of Other Railroad or Other Entity Filing for Equipment Involved in Train Accident/Incident	2a. Alphabetic Code	2b. Railroad Accident/Incident No.
3. Name of Railroad or Other Entity Responsible for Track Maintenance *(single entry)*	3a. Alphabetic Code	3b. Railroad Accident/Incident No.
4. U.S. DOT Grade Crossing Identification Number	5. Date of Accident/Incident	6. Time of Accident/Incident AM ☐ PM ☐

7. Nearest Railroad Station	8. Subdivision	9. County	10. State Abbr. / Code
11. City *(if in a city)*	12. Highway Name or Number Public ☐ Private ☐		

Highway User Involved

13. Type — Code
A. Auto B. Truck C. Truck-trailer D. Pick-up truck E. Van F. Bus G. School bus H. Motorcycle J. Other motor vehicle K. Pedestrian M. Other *(specify)*

14. Vehicle Speed *(est. mph at impact)*

15. Direction *(geographical)* — Code
1. North 2. South 3. East 4. West

16. Position — Code
1. Stalled or stuck on crossing 2. Stopped on crossing 3. Moving over crossing 4. Trapped on crossing by traffic 5. Blocked on crossing by gates

20a. Was the highway user and/or rail equipment involved in the impact transporting hazardous materials? — Code
1. Highway user 2. Rail equipment 3. Both 4. Neither

Rail Equipment Involved

17. Equipment — Code
1. Train *(units pulling)* 2. Train *(units pushing)* 3. Train *(standing)* 4. Car(s) *(moving)* 5. Car(s) *(standing)* 6. Light loco(s) *(moving)* 7. Light loco(s) *(standing)* 8. Other *(specify)* A. Train pulling – RCL B. Train pushing – RCL C. Train standing – RCL D. EMU Locomotive(s) E. DMU Locomotive(s)

18. Position of Car Unit in Train

19. Circumstance — Code
1. Rail equipment struck highway user 2. Rail equipment struck by highway user

20b. Was there a hazardous materials release by — Code
1. Highway user 2. Rail equipment 3. Both 4. Neither

20c. State here the name and quantity of the hazardous material released, if any.

21. Temperature *(Specify if minus)* ° F	22. Visibility *(single entry)* — Code 1. Dawn 2. Day 3. Dusk 4. Dark	23. Weather *(single entry)* — Code 1. Clear 2. Cloudy 3. Rain 4. Fog 5. Sleet 6. Snow		
24. Type of Equipment Consist *(single entry)* — Code 1. Freight Train 2. Passenger Train-Pulling 3. Commuter Train-Pulling 4. Work train 5. Single Car 6. Cut of cars 7. Yard/switching 8. Light loco(s) 9. Maint./inspect. Car A. Spec. MoW Equip. B. Passenger Train-Pushing C. Commuter Train-Pushing D. EMU E. DMU	25. Track Type Used by Rail Equipment Involved — Code 1. Main 2. Yard 3. Siding 4. Industry	26. Track Number or Name		
27. FRA Track Class (1-9, X)	28. Number of Locomotive Units	29. Number of Cars	30. Consist Speed *(Recorded speed, if available)* — Code R - Recorded E - Estimated MPH	31. Time Table Direction — Code 1. North 3. East 2. South 4. West
32. Type of Crossing Warning — 1. Gates 2. Cantilever FLS 3. Standard FLS 4. Wig wags 5. Hwy. traffic signals 6. Audible 7. Crossbucks 8. Stop signs 9. Watchman 10. Flagged by crew 11. Other *(specify)* 12. None Code(s)	33. Signaled Crossing Warning — Code *(See reverse side for instructions and codes)*	34. Roadway Conditions — Code A. Dry B. Wet C. Snow/slush D. Ice E. Sand, Mud, Dirt, Oil, Gravel F. Water (Standing, Moving)		
35. Location of Warning — Code 1. Both sides 2. Side of vehicle approach 3. Opposite side of vehicle approach	36. Crossing Warning Interconnected with Highway Signals — Code 1. Yes 2. No 3. Unknown	37. Crossing Illuminated by Street Lights or Special Lights — Code 1. Yes 2. No 3. Unknown		
38. Highway User's Age	39. Highway User's Gender — Code 1. Male 2. Female	40. Highway User Went Behind or in Front of Train and Struck or was Struck by Second Train — Code 1. Yes 2. No 3. Unknown	41. Highway User — Code 1. Went around the gate 2. Stopped and then proceeded 3. Did not stop 4. Stopped on crossing 5. Other *(specify)* 6. Went around/thru temporary barricade *(if yes, see instructions)* 7. Went thru the gate 8. Suicide/Attempted suicide	
42. Driver Passed Standing Highway Vehicle — Code 1. Yes 2. No 3. Unknown	43. View of Track Obscured by *(primary obstruction)* — Code 1. Permanent structure 2. Standing railroad equipment 3. Passing train 4. Topography 5. Vegetation 6. Highway vehicles 7. Other *(specify)* 8. Not obstructed			

Casualties to:	Killed	Injured	44. Driver was — Code 1. Killed 2. Injured 3. Uninjured	45. Was Driver in the Vehicle? — Code 1. Yes 2. No
46. Highway-Rail Crossing Users			47. Highway Vehicle Property Damage *(est. dollar damage)*	48. Total Number of Vehicle Occupants *(including driver)*
49. Railroad Employees			50. Total Number of People on Train *(include passengers and train crew)*	51. Is a Rail Equipment Accident/ Incident Report Being Filed? — Code 1. Yes 2. No
52. Passengers on Train				

53a. Special Study Block — Video Taken? ☐ Yes ☐ No; Video Used? ☐ Yes ☐ No	53b. Special Study Block

54. Narrative Description *(Be specific, and continue on separate sheet if necessary)*

55. Typed Name & Title	56. Signature	57. Date

NOTE: This report is part of the reporting railroad's accident report pursuant to the accident reports statute and, as such shall not "be admitted as evidence or used for any purpose in any suit or action for damages growing out of any matter mentioned in said report...." 49 U.S.C. 20903. See 49 C.F.R. 225.7 (b).

INSTRUCTIONS FOR COMPLETING BLOCK 33

Only if Types 1 - 6, Item 32 are indicated, mark in Block 33 the status of the warning devices at the crossing at the time of the accident, using the following codes:

1. Provided minimum 20-second warning.
2. Alleged warning time greater than 60 seconds.
3. Alleged warning time less than 20 seconds.
4. Alleged no warning.
5. Confirmed warning time greater than 60 seconds.
6. Confirmed warning time less than 20 seconds.
7. Confirmed no warning.

If status code 5, 6, or 7 was entered, also enter a letter code

explanation from the list below: A. Insulated rail vehicle.

B. Storm/lightning damage.

C. Vandalism.

D. No power/batteries dead.

E. Devices down for repair.

F. Devices out of service.

G. Warning time greater than 60 seconds attributed to accident-involved train stopping short of the crossing, but within trackcircuit limits, while warning devices remain continuously active with no other in-motion train present.

H. Warning time greater than 60 seconds attributed to track circuit failure (e.g., insulated rail joint or rail bonding failure, trackor ballast fouled, etc.).

J. Warning time greater than 60 seconds attributed to other train/equipment within track circuit limits.

K. Warning time less than 20 seconds attributed to signals timing out before train's arrival at the crossing/island circuit.

L. Warning time less than 20 seconds attributed to train operating counter to track circuit design direction.

M. Warning time less than 20 seconds attributed to train speed in excess of track circuit's design speed.

N. Warning time less than 20 seconds attributed to signal system's failure to detect train approach.

P. Warning time less than 20 seconds attributed to violation of special train operating instructions.

R. No warning attributed to signal system's failure to detect the train.

S. Other cause(s). Explain in Narrative Description.

This collection of information is mandatory under 49 CFR 225 and is used by FRA to monitor national rail safety. Public reporting burden is estimated to average 2 hours per response, including the time for reviewing instructions, searching existing databases, gathering, and maintaining the data needed, and completing and reviewing the collection of information. The information collected is a matter of public record, and no confidentiality is promised to any respondent. Please note that an agency may not conduct or sponsor, and a person is not required to respond to a collection of information unless it displays a currently valid OMB control number. The OMB control number for this collection is 2130-0500.

FORM FRA F 6180.57 (Rev. 08/10) OMB approved 7/30/2022, Approval expires 07/31/2023