Micro and Macro Level Safety Analysis at Railroad Grade Crossings

Submitted to NURail

By:

Juan C. Medina
Shiyu Shen
Rahim F. Benekohal
University of Illinois at Urbana-Champaign

Revised March 2016

1. Introduction

Railroad grade crossings are potential conflict points between train and highway vehicles, and train and pedestrians. Grade crossings pose a risk to all the travelers and the degree of risk depends on factors such as train and vehicle volumes, presence and operation of traffic control devices (TCD), speed of trains and vehicles, geometry of the crossing, interaction with interconnected intersection, pedestrians, number of tracks, etc. The risk assessment at grade crossings is important to prioritize locations for safety improvements and for assigning economic resources to locations that could provide the greatest returns in investments.

To assess the safety and risk of grade crossings, one may utilize macroscopic models derived from regression analysis using data from an entire state or country. Macroscopic analysis is mainly used to study the correlation between crossing characteristics (i.e. conflicting volumes and their physical characteristics) and the past accident frequency, to generate predictions about accidents in the future. A common macroscopic model currently used is the US DOT Accident Prediction Formula, which predicts the annual number of crashes at a given crossing. While this model provides useful information for ranking grade crossings for resource allocation, it was developed in the late 1970s and thus it has potential for improvement.

The analysis of macroscopic models is one of the two main topics from this study, where a series of alternative models are explored and compared to the US DOT formula. Through these comparisons, it is possible to understand strengths and weaknesses of the current state-of-practice and also to identify models for future improvements. A multi-scale analysis using macro models with input from a micro analysis is also planned for the future.

In addition to macroscopic analysis, the second main topic discussed here is the analysis of accidents from a microscopic perspective, where individual characteristics of accidents at a given location are investigated to determine potential contributing factors. A microscopic analysis procedure is described in this study takes into account detailed information such as driver characteristics, surrounding development type, lighting conditions, weather and visibility, direction of travel, speed and location of vehicle and train, etc. The proposed microscopic analysis is used to discover local trends that could not be identified at a macro scale and may help in assessing the risk at a crossing. Examples for the use of the micro analysis are provided to show the benefits of such a combined methodology and its potential use as a tool for on-site inspection.

The next sections describe the macroscopic approach using the US DOT formula, the exploration of alternative models, and their performance for assessing the risk using data from Illinois as an example. Then, the micro analysis is described and exemplified as a tool for risk analysis and trend discovery. Finally, conclusions are presented along with proposed directions for future research and improved risk assessments.

2. Macroscopic Analysis of Accidents at Railroad-Highway Grade Crossings

As mentioned above, the analysis of accident data at railroad-highway grade crossings is often performed at a macroscopic scale, where models are created using datasets that include crossings at a regional or national level to identify general accident trends. These models are mainly used to analyze the correlation between crossing characteristics (i.e. conflicting volumes and their physical characteristics) and the past accident frequency, and thus, to generate predictions about accidents in the future.

Safety improvements on grade crossings have been studied since the railroads have been in place in the second half of the nineteenth century, when gates had already been conceived as a mean for crossing protection. However, recent mainstream estimation of risks at grade crossings date back to the early 1970s with the consolidation of an inventory of all crossings, developed by public agencies (FHWA, FRA, AAR) and private railroads.

A crossing inventory, along with a collection of accident records at grade crossings, were essential to put in practice a systematic method for identifying and prioritizing grade crossings for safety improvements, in accordance to the mandates in the Federal-Aid Policy Guide, which contains the HSIP (US DOT and Federal Highway Administration, 2007).

After these databases were available, they could be used for estimation of risks, accidents, and accident severity at a large scale. A number of approaches to analyze safety at grade crossings over time have taken a number of forms including hazard or risk indices (e.g. New Hampshire Hazard Index), accident prediction models (e.g. U.S. DOT Accident Prediction Model), or methodologies to find hot or black spots (e.g. Saccomanno et al., 2004), among many others. The techniques and approaches for accident analysis cover several areas ranging from a variety of traditional regression analyses (e.g. Miranda-Moreno et al., 2005), to clustering (e.g. Anderson, 2009), to Bayesian modifications (e.g. Washington and Oh, 2006), to behavioral models (e.g. Leibowitz, 1985), just to mention a few.

Conflicting volumes and crossing characteristics that are associated with the locations with higher accidents can be seen as potential contributing factors, and improvements on crossings with these characteristics can be planned based on the models' results. Also, the prediction models may be useful in identifying "hot spots" and allocating monetary resources.

In this Chapter, an approach that is commonly used by public agencies (the US DOT Accident Prediction Formula) is described and its prediction results are compared to alternative statistical models.

Comparison of the most significant contributing variables from the DOT formula and the models developed in this study could provide insights on different trends at a local or regional level (as opposed to a nationwide level), in particular for the state of Illinois, which is the focus in this study. In this regard, it is noted that the statistical models were created based on data from the state of Illinois, while the existing US DOT formula was created using crossings nationwide.

The following sections describe the US DOT formula and the statistical models, highlighting the crossing characteristics that appear as the most relevant contributing factors to explain the accident frequency. Then, the models are compared and similarities and differences are presented.

2.1. The US DOT Accident Prediction Formula

A common method to rank grade crossings for resource allocation and improvements is to predict the likelihood of an accident over a time period. This value can also be used as a measure of the risk of an accident, and therefore, it is of importance for agencies responsible for the safety of grade crossings. The Railroad-Highway Grade Crossing Handbook describes the use of the US DOT Accident Prediction

Formula, which can be used for assessing the likelihood of an accident at a given crossing based on its traffic and geometric characteristics.

The US DOT Accident Prediction Formula has three elements that result in an initial accident prediction, a second accident prediction, and a final collision prediction. The functional forms of these three elements were the result of multiple regression based on the FRA database from 1975. The current version of the US DOT Prediction Formula is presented in the Railroad-Highway Grade Crossing Handbook, Revised Second Edition from 2007 by the FHWA. The current formula has the same coefficients for the crossing characteristics presented in 1980 (Mengert, 1980), except for the third model element which consists of a normalizing constant that is factored in the model right before the final prediction is obtained. The variables included in the models or their coefficients have not been modified since the original study was published, and thus there could be potential for improvements in the identification of newer variables, different effects of the contributing factors, and the overall predictions.

2.1.1. Background

The research project that lead to the development of the current US DOT accident prediction formula was conducted in the late 1970s and presented in a report titled "Rail-Highway Crossing Hazard Prediction Research Results" by Peter Mengert (1980).

The main goal of this study was to determine absolute predicted number of crashes at grade crossings and identifying causal factors. The study states, however, that there is no connection between prediction and isolating factors that are causally related to accidents, thus the formulas are not intended to be used partially or using only some of the variables it contains. The end result was expected to serve as input for a resource allocation procedure that was being developed by the US DOT.

Different approaches were analyzed in the process of finding an appropriate functional form for the model. A subset from the original FRA database was created by selecting crossings with accidents and assigning a dummy variable of 1, as opposed to other crossings that were assigned a zero value. The database was explored using an initial set of 51 variables that included exposure, geometric, and other characteristics of all crossings. However, models with too many variables (say 10 or more, based on the study) may suffer of collinearity, limiting the usability of some of them.

The first attempts at finding a functional form for the model relied on liner regression to obtain a relative rank of predictability among the selected variables. Also, it is explained that the decision of dividing the predictions by warning devices was based on previous studies. The analysis of crossings with crossbucks was conducted first, noticing that the variables related to vehicular and train volumes contributed the most to the models.

Expressions were obtained using only volume variables, and several functional forms were explored. Other variables were added to the model once the expression for the volume variables were found, following a step-wise procedure. Among the first findings, it was said that a Log 10 functional form was found more adequate than others tested. The final linear regressions showed that volume variables accounted for about 90% of the predicted powers of the regression.

Nonetheless, linear models were deemed not good enough because results were not better than those obtained with previous models, namely the New Hampshire and Coleman-Steward. In addition, it was noted that there was a concentration of the probability of accidents at the two ends of the spectrum in the models. This is, an approximation that assumes a linear distribution between the probabilities of having 0 or 1 crashes may not be adequate.

The next steps focused on non-linear methods and at most 6 variables (at that time) in order to reduce noise in the models. The dataset was carefully divided into test and validation subsets, each with the same number of observations per device and accidents.

The model construction for the non-linear model using the test subset followed similar steps as in the linear model attempts, and include the following:

- Construct the best volume model, only with volume variables
- Refine the best volume model, shaping it to a polynomial up to the third degree
- Incorporate non-volume variables

These steps and additional refinements lead to the final models that now are part of the US DOT prediction formula used today.

Understanding how the current prediction formula was created is a stepping stone in potential improvements using other prediction models. Lessons learned and contributions from the study briefly described above were put to use in the development of the models described in this chapter. Before that, the next section presents a short introduction of the actual US DOT prediction formula and how it is used today.

2.1.2. Implementation of the Accident Prediction Formula

The process to obtain a predicted number of accidents in a year for a given crossing starts with the estimation of the initial collision prediction using the basic formula, defined as:

$$a = K \times EI \times MT \times DT \times HP \times MS \times HT \times HL$$

where:

a = initial collision prediction, collisions per year at the crossing

K = formula constant

EI = factor for exposure index based on product of highway and train traffic

MT = factor for number of main tracks

DT = factor for number of through trains per day during daylight

HP =factor for highway paved (yes or no)

MS = factor for maximum timetable speed

HT = factor for highway type

HL = factor for number of highway lanes

There are three different equations, one for each warning device: passive devices, flashing lights, and automatic gates.

It should be noted that not all variables in the model are significant for all three types of warning devices. Some of the variables have a value of 1 indicating that it doesn't affect the model outcome, as it is seen in Figure 2.1. Expressing the three equations in terms of the same variables is convenient. For example, for crossings with flashing lights the coefficients for the terms *HP*, *MS*, and *HT* are always 1, but of course these variables have an active role for a different group (passive devices).

Crossing Category	Formula Constant K	Exposure Index Factor EI	Main Tracks Factor MT	Day Thru Trains Factor DT	Highway Paved Factor HP	Maximum Speed Factor MS	Highway Type Factor HT	Highway Lanes Factor HL
Passive	0.002268	$\left(\frac{c \times t + 0.2}{0.2}\right)^{0.3334}$	$e^{0.2094mt}$	$\left(\frac{d + 0.2}{0.2}\right)^{0.1336}$	$e^{-0.6160(hp-1)}$	$e^{0.0077ms}$	$e^{-0.1000(ht-1)}$	1.0
Flashing Lights	0.003646	$\left(\frac{c \times t + 0.2}{0.2}\right)^{0.2953}$	$e^{0.1088mt}$	$\left(\frac{d + 0.2}{0.2}\right)^{0.0470}$	1.0	1.0	1.0	$e^{0.1380(hl-1)}$
Gates	0.001088	$\left(\frac{c \times t + 0.2}{0.2}\right)^{0.3116}$	$e^{0.2912mt}$	1.0	1.0	1.0	1.0	$e^{0.1036(hl-1)}$

c = annual average number of highway vehicles per day (total both directions)

t = average total train movements per day

mt = number of main tracks

d = average number of thru trains per day during daylight

hp = highway paved, yes = 1.0, no = 2.0

ms = maximum timetable speed, mph

ht = highway type factor value

hl = number of highway lanes

Highway Type	Inventory Code	ht Value
Rural		
Interstate	01	I
Other principal arterial	02	2
Minor arterial	06	3
Major collector	07	4
Minor collector	08	5
Local	09	6
Urban		
Interstate	11	1
Other freeway and expressway	12	2
Other principal arterial	14	3
Minor arterial	16	4
Collector	17	5
Local	19	6

Figure 2.1. U.S. DOT Collision Prediction Equations for Crossing Characteristics Factors (Source: Railroad-Highway Grade Crossing Handbook, FHWA, 2007)

After obtaining the initial collision prediction value (a), a second collision prediction value (B) is computed using the actual crash history of the crossing and the value a, as shown below:

$$B = \frac{T_0}{T_0 + T}(a) + \frac{T}{T_0 + T}\left(\frac{N}{T}\right)$$

Where:

B = second collision prediction, collisions per year at the crossing

N/T = collision history prediction, collisions per year, where N is the number of observed collisions in T years at the crossing

T_0 = Formula weighting factor, $T_0 = \frac{1.0}{(0.05+a)}$

Finally, the second collision prediction is adjusted using a single normalizing constant for each type of warning device. The FRA explains that the constants need to be periodically adjusted in order to keep the procedure matched with the current accident trends, the current number of open crossings, and the change in the warning devices. The most recent normalizing constants date back to 2010 and are shown in Table 2.1, obtained from the periodic document released by the FRA (FRA, 2010).

Table 2.1. Accident Prediction and Resource Allocation Procedure Normalizing Constants

WARNING DEVICE GROUPS	NEW	PRIOR YEAR CONSTANTS							
	2010	**2007**	**2005**	**2003**	**1998**	**1992**	**1990**	**1988**	**1986**
(1) Passive	**.4613**	.6768	.6407	.6500	.7159	.8239	.9417	.8778	.8644
(2) Flashing Lights	**.2918**	.4605	.5233	.5001	.5292	.6935	.8345	.8013	.8887
(3) Gates	**.4614**	.6039	.6513	.5725	.4921	.6714	.8901	.8911	.8131

2.2. Alternative Models for Accident Prediction

In addition to the US DOT Model, other statistical models are explored to obtain accident frequency predictions at grade crossings. These models have been used in the past for event counts where frequencies of such events are generally low, such as in our case. The three models explored are Poisson, negative binomial, and a zero-inflated negative binomial.

2.2.1. Poisson

A common way to model event count data, which may contain positive and integer events with low frequency such as accidents at grade crossings, is using a Poisson model. A Poisson distribution is a discrete probability distribution that can describe the frequency of events over a time period, and can be defined by its mean μ_i. The probability of observing y_i events over a given time period can be described by:

$$P(Y = y_i|x_i) = \frac{e^{-\mu_i}{\mu_i}^{y_i}}{y_i!}$$

Where *e* is the base of the natural logarithm (*e* = 2.71828) and *y*! is the factorial of *y*.

A regression analysis to fit a Poisson distribution and to model to the accident count data can be performed using a log-linear model. A log-linear function ensures that the predicted number of accidents is positive, and the model is fit to predict the logarithm of the average number of accidents using a linear combination of predictors.

In this particular application, the predictors are the explanatory variables with significant influence in the model, e.g. vehicle and train volumes, geometry characteristics, etc. Thus, a standard linear model of the form $\mu_i = a + x_i'\beta$ can be transformed using the log function as:

$$log(\mu_i) = log(a) + x_i'\beta$$

Where, *a* is an intercept, β are the parameters estimated in the model fitting, and x_i are the explanatory variables. Therefore, β_i represents the expected change in the left-hand side of the expression (the log of the mean) when the variable x_i changes by one unit. A predictor for the mean itself can be obtaining by exponentiation of the previous equation, leading to:

$$\mu_i = b + e^{x_i'\beta}$$

Where *b* is an intercept or an offset. This log-linear model will be used to create an accident prediction formula and will be compared to the U.S. DOT model.

However, it is noted that the data may not adequately fit a Poisson distribution, particularly due to the variance being greater than the mean, one of the Poisson's main properties. Therefore, additional models that may account for "excess" of variance or "over-dispersion" were also investigated and are described next.

2.2.2. Negative Binomial

Another functional form that can be used to model count data is a negative binomial model. This model is of most interest when the variance of the predicted variable increases at a higher rate than the mean (when a Poisson model does not fit properly).

A set of parameters commonly used to describe a negative binomial model are the expected mean of the predicted variable $E(y_i) = \mu_i$ and the variance, denoted as:

$$Var(y_i) = \mu_i + \omega{\mu_i}^2$$

Where, $\mu_i = \alpha\theta$ and ω= $\frac{1}{\alpha}$. $Here, \alpha$ and θ are the shape and scale parameters of a gamma distribution that describe the distribution of the mean in a Poisson distribution. Thus, the predicted variable follows Poisson, and its mean follows a gamma distribution. Since α and θ are positive values, it is easy to observe that a negative binomial model can accommodate over-dispersion (greater variance) but it cannot adequately deal with a Poisson distribution when there is under-dispersion.

In a negative binomial regression, the predicted variable can be expressed in terms of the new distribution with mean μ_i and variance $\mu_i + \omega{\mu_i}^2$ and the coefficients of the explanatory variables will be estimated using a log link function with an equation form similar to that used for the Poisson distribution:

$$y_i = NB(\mu_i, \omega)$$

$$log(y_i) = log(c) + x_i'\beta$$

Where NB is a negative binomial distribution.

2.2.3. Zero-inflated Negative Binomial

In cases where the data displays high frequency of zero counts, such as data with accident counts at railroad-highway crossings, zero-inflated models can be considered to improve the fit of the regression analysis.

The main idea behind zero-inflated models is to consider that there are two types of individuals in the population trying to be modeled. One type contains the individuals distributed according to a given event count distribution, say a Poisson or a negative binomial, and the other type contains those individuals with zero even counts. Thus, a zero-inflated model is a two-part model that accounts for the existence of excessive zeroes using the probability of being part of one category (the zero category), leaving the remaining individuals to be part of a second category that is fitted to a given distribution.

Conceptually, if y_i is the number of accidents at a crossing, a zero-inflated negative binomial distribution can be generated as:

$$y_i \sim \begin{cases} 0 & with\ probability\ \varphi_i \\ g(y_i) & with\ probability\ 1-\varphi_i \end{cases}$$

Where φ_i is the process that generates individuals with event counts and $g(y_i)$ is the process that generates the number of accidents following a negative binomial distribution. Thus, the probability of $Y_i = y_i$ is:

- Individuals with zero counts:

$$P(y_i = 0|x_i) = \varphi_i + (1-\varphi_i)g(0)$$

- Individuals with counts greater than zero:

$$P(y_i > 0|x_i) = (1-\varphi_i)g(y_i)$$

It follows that zero-inflated models can be constructed using link functions such as logit or logistic. In this study, a logit link function was used to model the excessive zeroes thereby extending a generalized linear model. More specifically, the following model definition implemented in the Statistical Analysis Software (SAS) was selected:

$$P(y_i) = \begin{cases} \varphi_i + (1-\varphi_i)(1+k\lambda) & for\ y_i = 0 \\ (1-\varphi_i)\dfrac{\Gamma\left(y+\frac{1}{k}\right)}{\Gamma(y+1)\Gamma\left(\frac{1}{k}\right)}\dfrac{(k\mu)^y}{(1+k\lambda)^{y+\frac{1}{k}}} & for\ y_i > 0 \end{cases}$$

Where k is the negative binomial dispersion parameter and λ is the underlying distribution mean such that $E(Y) = \mu = (1-\varphi_i)\lambda$ and $Var(Y) = \mu + \left(\frac{\varphi_i}{1-\varphi_i} + \frac{k}{1-\varphi_i}\right)\mu^2$

2.3. Development of Alternative Models using FRA Inventory and Accident Data

The analysis performed in this study was based on the current crossing inventory data from the FRA website (http://safetydata.fra.dot.gov/), as of fall 2013. Only crossings from the state of Illinois were analyzed in this study, and results from the database were post-processed and filtered for at-grade public crossings that remain open.

It is noted that the FRA database is kept up to date based on accident reports submitted by law by the railroad companies within 30 days after the month to which they pertain, and the inventory data reports from sight-survey data about individual crossings and is provided voluntarily by states and railroads. Therefore, accident frequencies are expected to be accurate and up to date, but the available records indicating the state of the crossings are based on the last voluntary data submission to the inventory. Records will reflect the latest information available, thus values for variables such as vehicular and train volume could be from any point in time depending on the update.

Also, the history of modifications to grade crossings, including safety improvements is not in the database, which reflects only the latest reported information. Thus, if a crossing had an improvement in the type of warning device, the date and nature of this change is not kept in the inventory database, and it may only be partially observed in the variables of the accident reports, had any occurred after this improvement.

These characteristics of the database pose some challenges for analysis of accident trends, and while it is very useful, caveats on the limitations of the data and the accuracy of the reported numbers need to be stated.

In terms of accident records used in this study, a query for individual accidents at grade crossings by year was requested through the FRA website for the last 10 years of available data, from 2003 to 2012. Similar to the grade crossing inventory, only accidents at grade public crossings that remain open were considered.

After the datasets were obtained, the two files (inventory and accidents) were merged to find the locations with accidents and without accidents, as well as the accident frequency in the analysis period. At this point, the total number of crossings with accidents was observed to be 938, and the number of crossings without accidents was 6072.

2.3.1. Potential Predictor Variables

Further observation of the data prompted the need for data cleaning. Fields for some variables of interest were found to be empty or inconsistent, and thus their distribution was analyzed to determine their usefulness in a regression analysis. These fields were the following:

- AADT
- AADTYEAR (Year for AADT)
- TOTALTRN (Total trains)
- MAXTTSPD (Maximum Timetable Speed)
- PCTTRUK (Estimate Percent Trucks)
- HWYSPEED (Posted Highway Speed)

Data exploration for these fields returned the range and distributions shown in Tables 2.2 and 2.3 for the crossings with and without accidents.

TABLE 2.2. Range and distribution of variables of interest – crossings with accidents

Percentile	Crossings with Accidents (n = 887)					
	AADTYEAR**	**AADT**	**TOTALTRN***	**MAXTTSPD**	**PCTTRUK**	**HWYSPEED**
100% Max	2012	39500	8008 (215 *)	79	75	55
99%	2011	28200	171	79	32	55
95%	2011	19700	106	79	16	50
90%	2011	14300	72	79	11	40
75% Q3	2010	6800	47	70	7	30
50% Median	2009	1500	21	60	4	0
25% Q1	2008	300	8	40	0	0
10%	2006	75	4	20	0	0
5%	2003	50	2	10	0	0
1%	1987	10	1	10	0	0
0% Min	1973	9	1	5	0	0

* Eliminated crossings with train volumes >2000

** Eliminated crossings with AADTs older than 2000

TABLE 2.3. Range and distribution of variables of interest – crossings without accidents

Percentile	Crossings without accidents (n = 6072)					
	AADTYEAR**	AADT	TOTALTRN*	MAXTTSPD	PCTTRUK	HWYSPEED
100% Max	2013	45500	8008 (255*)	79	2250	55
99%	2011	22600	156	79	30	55
95%	2011	10500	62	79	17	45
90%	2011	5950	43	79	12	35
75% Q3	2010	1750	21	60	7	20
50% Median	2009	400	8	49	2	0
25% Q1	2007	100	3	25	0	0
10%	2005	25	2	10	0	0
5%	2002	25	1	10	0	0
1%	1987	10	1	5	0	0
0% Min	1973	1	1	1	0	0

* Eliminated crossings with train volumes >2000

** Eliminated crossings with AADTs older than 2000

Based on the data, it was decided to discard crossings with AADT previous to the year 2000, avoiding crossings with information considerably outdated. A total of 23 crossings with accidents and 185 crossings without accidents were in this category. In addition, crossings with a total daily train count greater than 2000 were eliminated, as they would represent an entry code instead of actual train traffic (for example, the code 8008 was found multiple times). A total of 3 crossings with accidents and 13 without accidents were in this category. The maximum daily train volume in the remaining crossings was 255 trains.

After eliminating the abovementioned crossings, the total number of locations with accidents was 861 and without accidents it was 5874. This was the final database used in the analysis.

Exploration of other variables such as the percentage of trucks and the highway speed (also shown in Tables 2.2 and 2.3) indicate that these variables may not be adequate for the models. About half of the crossings had missing information on these fields (a value of zero), thus including them in a regression analysis could generate inconsistent results.

In addition, the distribution of values and categories for other variables of interest is shown in Table 2.4. These variables were considered worth of further investigation as predictors in the statistical models and were included in the initial regressions to determine their statistical significance.

TABLE 2.4. Categories and range of other variables considered in the analysis

Variable	Categories	Crossings with accidents	Crossings without accidents
SPSEL (Train detection)	CWT (constant warning time)	47.30%	29.90%
	DC/AFO (direct current/audio frequency overlay)	27.30%	31.00%
	None	22.20%	36.70%
	Older code not used anymore/missing value	3.20%	2.40%
XANGLE (Minmum crossing angle)	1 (0◦ - 29◦)	4.41%	2.49%
	2 (30◦ - 59◦)	16.26%	16.24%
	3 (60◦ - 90◦)	79.33%	81.27%
WDCODE (Warning device code)	1 - no signs or signals	0.12%	0.61%
	2 - other signs or signals	0.00%	0.07%
	3 - crosbucks	14.05%	29.95%
	4 - stop signs	1.16%	0.80%
	5 - special active WD	0.12%	0.46%
	6 - other active WD	0.23%	1.36%
	7 - flash lights	13.94%	23.41%
	8 - all other gates	69.92%	41.96%
	9 - four quad gates	0.46%	1.38%
TOTAL TRACKS (Sum of MAINTRK and OTHERTRK)	1	45.76%	66.50%
	2	37.17%	25.54%
	3	13.01%	6.01%
	4	2.44%	1.38%
	5	0.70%	0.36%
	6	0.58%	0.15%
	7	0.00%	0.07%
	8	0.23%	0.00%
	9	0.12%	0.00%
TRAFICLN (Number of traffic lanes crossing RR)	1	11.85%	17.79%
	2	69.45%	75.57%
	3	1.86%	0.92%
	4	15.80%	5.16%
	5	0.35%	0.32%
	6	0.23%	0.15%
	7	0.00%	0.05%
	8	0.35%	0.03%
	9	0.12%	0.00%
HWYNEAR (Nearby intersecting highway?)	<75ft	42.86%	35.17%
	75-200 ft	3.48%	2.96%
	200-500 ft	2.90%	2.74%
	N/A	50.75%	59.12%

2.3.2. Accident Frequencies

The accident frequencies in the 861 locations in Illinois, together with those without accidents are shown in Table 2.5. The highest frequency in the 10-year period between 2003 and 2012 was observed at one location with 9 accidents, followed by a location with 7 accidents. As expected, the number of crossings with fewer accidents start increasing rapidly in each frequency group. From this table it is also noted that it may be worth exploring zero-inflated models to accommodate the high number of locations without any accident.

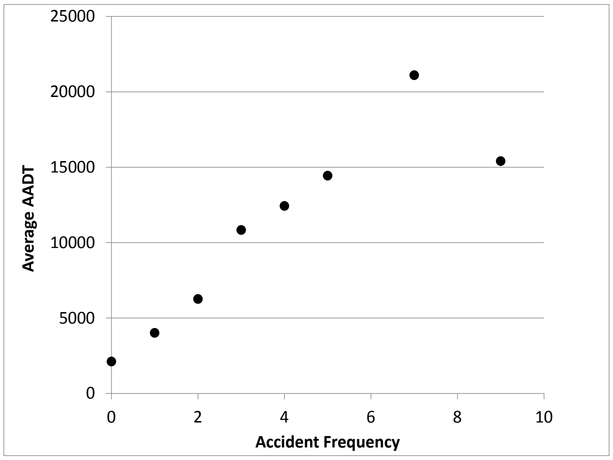
TABLE 2.5. Distribution of accident frequencies in Illinois (between 2003 and 2012)

Frequency	Locations in the analysis	%
0	5874	87.22%
1	685	10.17%
2	128	1.90%
3	31	0.46%
4	11	0.16%
5	4	0.06%
7	1	0.01%
9	1	0.01%
Total	**6735**	**100.00%**

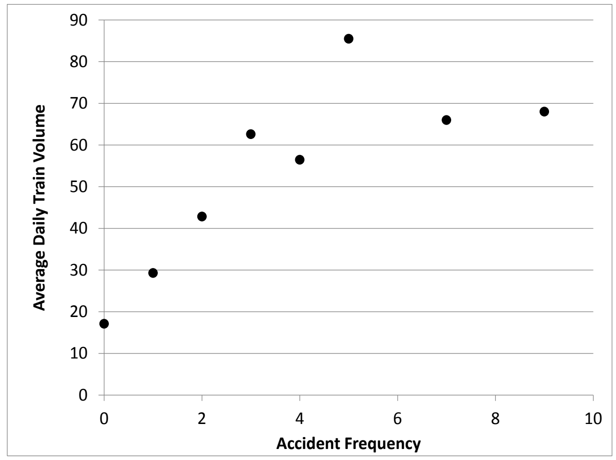
After deciding the final number of locations to be included in the analysis, additional exploration of the relationships between the values of some variables of interest and the accident frequency was conducted. This preliminary exploration helped understanding if there were clear trends that can explain an increase in accident frequency using a single variable in a model.

The distribution of the accident frequencies with different volume variables (train and vehicular traffic, number of tracks and traffic lanes), which have shown the greatest influence in previous models including the US DOT model, are shown in Figure 2.2. As expected, the volume (or exposure) variables have a direct relationship with the accident frequency, where locations with higher accident frequencies tend to have a greater volume/exposure values such as greater AADT, train volume, number of tracks and number of traffic lanes. However, the magnitude of the relation between accident frequency and the value of the variable is not constant, indicating that the potential influence of each of these variables may vary.

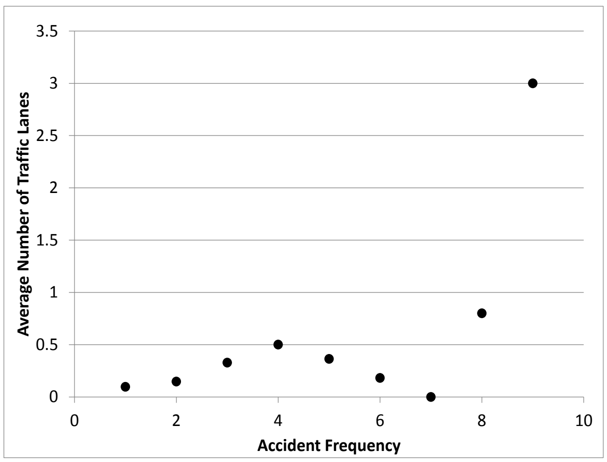
In addition, as part of observing the distribution of volume/exposure variables, a measure of exposure was created to evaluate the accident frequency for each of four types of warning control devices: passive, active, gates, and quad gates. Quad gates are analyzed separate from other type of gate configurations since they can prevent gate drive-arounds, similar to treatments using elements such as raised medians, barrier posts, raised traffic dots, etc.



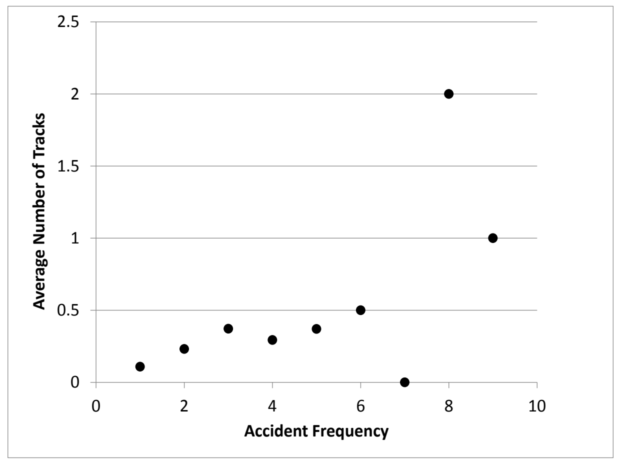
a – ADDT



b – Train Volume



c – Traffic Lanes



d – Railroad Tracks

Figure 2.2. Relation between Accident Frequency and Exposure Variables

Exposure was defined as the product of train and vehicular volume, and the accident frequency was normalized by the exposure measure in order to compare the warning devices. Thus, for each warning device type, the normalized average accident frequency shown in Figure 2.3 was calculated as follows:

$$Normalized\ Frequency = \frac{Average\ Frequency * 100}{Exposure}$$

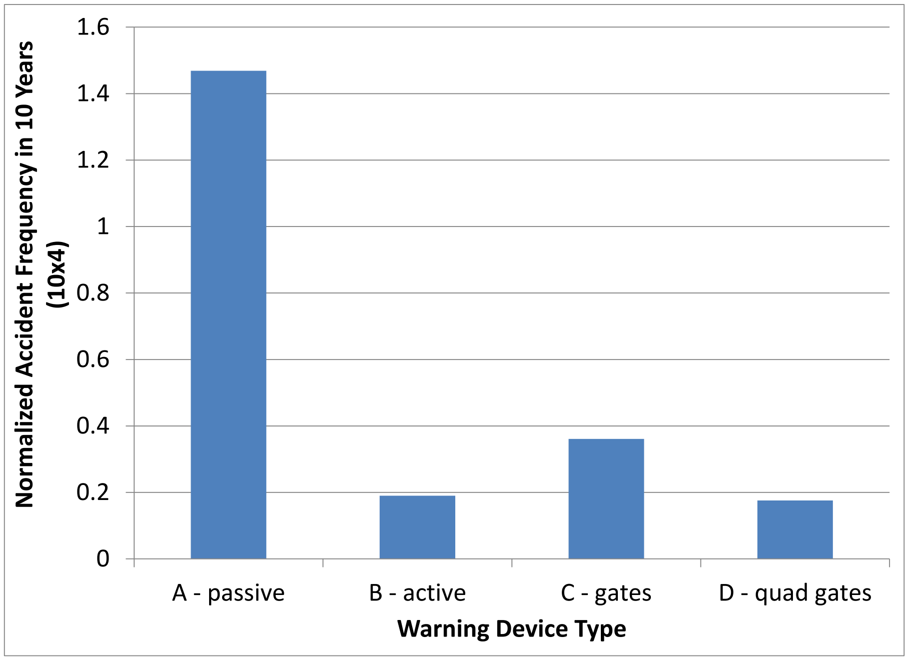


Figure 2.3. Relation between normalized accident frequency and warning device

Figure 2.3 shows that, as expected, crossings with passive warning devices have a much higher average risk per conflict compared to other types of crossings. Also, crossings with quad gates show reduced risks compared to those with other type of gates. Recall that the exposure levels are different for each warning type, being on average highest at crossings with gates, thus finding risk levels at locations with quad gates similar to those with other type of active warning device is a significant finding.

2.3.3. Variables Analyzed in the Regression Models

The set of variables that after the observational examination were considered suitable for being tested as potential contributing factors in the models are listed below. The exact labels used in the FRA database are shown in capitals, and the descriptions of the variable are shown in parenthesis.

- AADT (Annual average daily traffic)
- TOTALTRN (Total trains)
- MAXTTSPD (Maximum timetable speed)

- Control Device (Variable created from WDCODE):
  - Passive: includes no sings or signals, crossbucks, and stop signs.
  - Active: Signals, bells, wigwags, flashing lights, other active devices
  - Gates: All gates except quad gates
  - Quad Gates
- XANGLE (smallest crossing angle)
- Total Tracks (Total number of tracks: Sum of MAINTRK and OTHERTRK)
- TRAFFICLN (Number of traffic lanes)
- HWYNEAR (Nearby intersecting highway)

On the other hand, other variables that seemingly had good potential as contributing factors were not included in the models given their frequent missing or inaccurate values. Examples of those are the percentage of trucks, posted highway speed limit, and train detection.

The results for the three models: Poisson, negative binomial, and zero-inflated negative binomial are described next.

2.3.4. Poisson Model for all Crossings

A model including the variables listed above was evaluated using the GENMOD procedure in SAS, specifying a Poisson distribution and a log link function. The goodness of fit measures for such model returned by SAS are shown below in Table 2.6, where the model does not seem to fit well the data because Pearson chi-square value divided by the degrees of freedom is greater than 1, indicating that the data may be over-dispersed.

Table 2.6. Criteria for Assessing Goodness of Fit – Poisson Regression Analysis

Criterion	DF	Value	Value/DF
Deviance	6723	4096.642	0.6093
Pearson Chi-Square	6723	8188.518	1.218

However, to illustrate the results of the regression, an illustration of the model results and the accident data is shown in Figure 2.4. The model parameters, their estimates, and their significance are shown in Table 2.7. The average number of accidents predicted by the model was 0.16494 accidents per crossing for a total of 1111 accidents for all accidents together. In comparison, the FRA database had a total of 1114 accidents, thus the model was effective at predicting the overall accident frequency. However, the variance of the model was 0.03, which is significantly lower than in the database (0.25) and evidencing the over-dispersion indicated in the model fitting.

Table 2.7. Results of the Poisson Regression for all crossings

Parameter	Category	DF	Estimate	Standard Error	Pr > ChiSq
Intercept		1	-4.247	0.522	<.0001
aadt3		1	3.90E-05	5.13E-06	<.0001
total_train3		1	0.007	0.001	<.0001
max_ttspeed3		1	0.004	0.002	0.0093
control_device	A	1	0.889	0.510	0.0814
control_device	B	1	0.959	0.509	0.0598
control_device	C	1	1.471	0.503	0.0034
control_device	D	0	0	0	.
cross_angle3	1	1	0.435	0.139	0.0017
cross_angle3	2	1	0.053	0.083	0.5215
cross_angle3	3	0	0.000	0.000	.
total_tracks		1	0.165	0.035	<.0001
traf_lanes3		1	0.129	0.041	0.0016
hwy_near2	1	1	0.163	0.062	0.0086
hwy_near2	2	0	0	0	.
Scale		0	1	0	

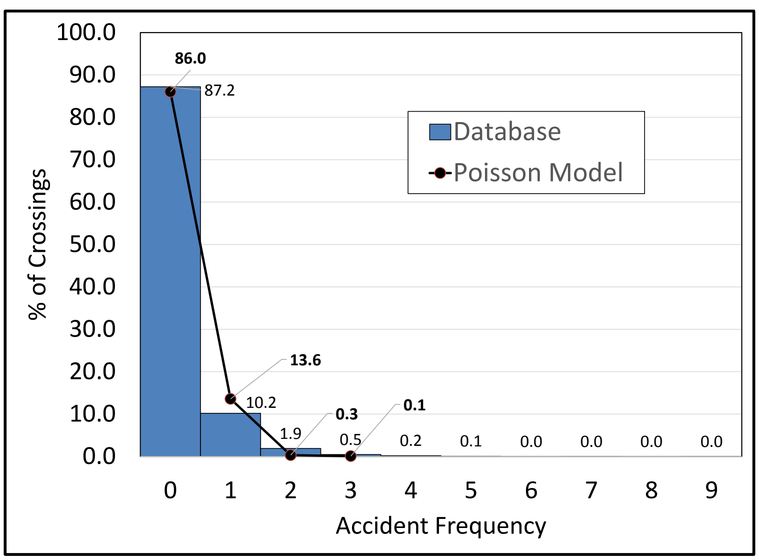


Figure 2.4. Results of the Poisson Model Compared to Field Data

A scatter plot of the residuals is shown in Figure 2.5. Note that the crossings ID were created from the original FRA database, and thus they are not in any particular order. The total number of observations in the scatter plot is 6735, as described earlier. The residuals are estimated as the difference between the observed number of accidents (y_i) and the predicted value from the model (f_i). The sum of square errors (SSE), the root mean squared error (RMSE), and the AIC (Akaike's Information Criterion), AICC (corrected Akaike's Information Criterion), and BIC (Bayesian Information Criterion) values (Table 2.8) were also found to determine the fit of the model comparatively with other models developed in this chapter:

$$SSE_{Poisson} = \sum_{1}^{6735} (y_i - f_i)^2 = 1514.5$$

$$RMSE_{Poisson} = \left(\frac{SSE}{n - v}\right)^{\frac{1}{2}} = \left(\frac{1514.5}{6735 - 12}\right)^{\frac{1}{2}} = 0.4746$$

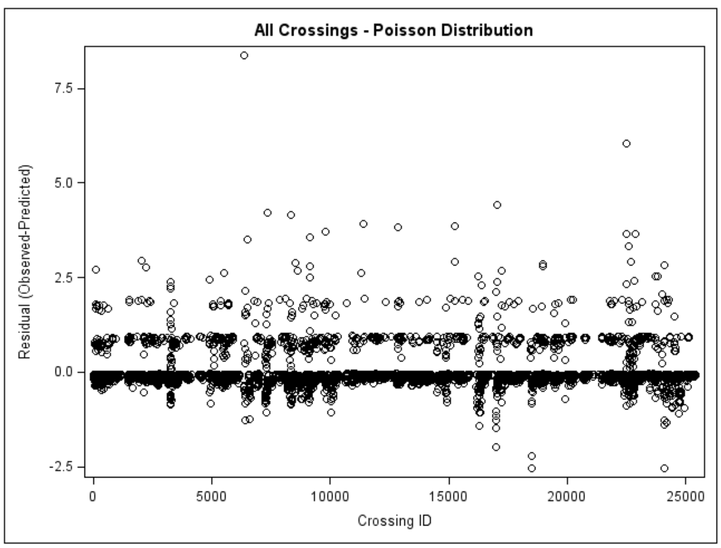


Figure 2.5. Scatter plot of the Residuals for the Poisson Model

Table 2.8. Criteria for Poisson Model

Criteria	Value
AIC (smaller is better)	5975.65
AICC (smaller is better)	5975.7
BIC (smaller is better)	6057.43

2.3.5. Negative Binomial Model for all Crossings

After identifying over-dispersion in the Poisson model and observing the predicted number of crashes with respect to the FRA database, a model using a negative binomial distribution was fitted to the data. Similar to the Poisson model, the GENMOD procedure in SAS was used, but this time specifying a negative binomial distribution with a log link function. The same variables were initially included in the model. Results of the goodness of fit are shown in Table 2.9.

Table 2.9. Criteria for Assessing Goodness of Fit – Negative Binomial Regression Analysis

Criterion	DF	Value	Value/DF
Deviance	6723	3119.932	0.4641
Pearson Chi-Square	6723	6860.395	1.0204

The negative binomial model shows a better overall fit compared to the Poisson model based on the Pearson chi-square values and its ratio when divided by the degrees of freedom. As a whole, the model generated an average of 0.1675 accidents per crossing, for a total of 1128 accidents (compared to 1114 accidents in the FRA database). The regression results and the comparison of the generated distribution with field data is shown in Table 2.10 and Figure 2.6.

Table 2.10. Results of the Negative Binomial Regression for all crossings

Parameter	Category	DF	Estimate	Standard Error	Pr > ChiSq
Intercept		1	-4.1458	0.554	<.0001
aadt3		1	4.50E-05	6.88E-06	<.0001
total_train3		1	0.0081	0.0012	<.0001
max_ttspeed3		1	0.0038	0.0019	0.0429
control_device	A	1	0.845	0.5347	0.114
control_device	B	1	0.9003	0.5343	0.092
control_device	C	1	1.4235	0.5275	0.007
control_device	D	0	0	0	.
cross_angle3	1	1	0.4251	0.1754	0.0153
cross_angle3	2	1	0.0529	0.095	0.5776
cross_angle3	3	0	0	0	.
total_tracks		1	0.1263	0.0429	0.0032
traf_lanes3		1	0.131	0.05	0.0088
hwy_near2	1	1	0.1361	0.0708	0.0547
hwy_near2	2	0	0	0	.
Scale		1	1.1563	0.1553	

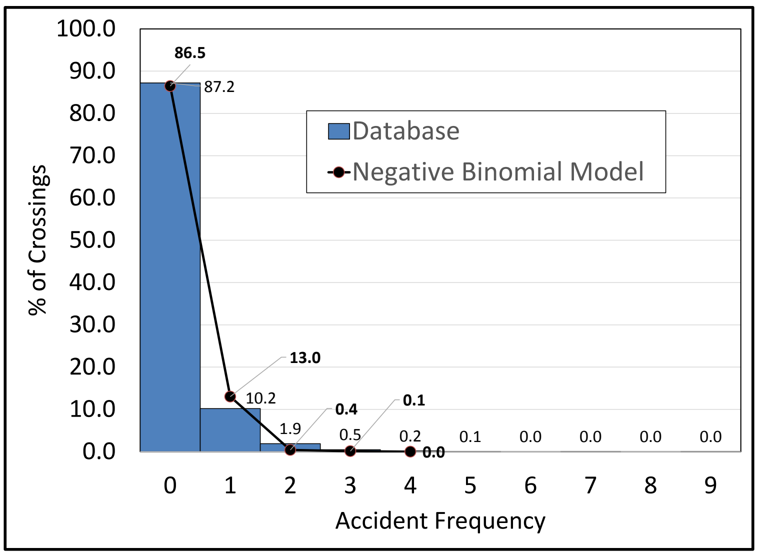


Figure 2.6. Results of the Negative Binomial Model Compared to Field Data

Similar to the analysis with a Poisson distribution, the residuals, the SSE and the RMSE are provided for the negative binomial model, as well as criteria for assessing the goodness of fit (Table 2.11), and the residuals (Figure 2.7) as follows:

$$SSE_{NB} = \sum_{1}^{6735} (y_i - f_i)^2 = 1534.7$$

$$RMSE_{NB} = \left(\frac{SSE}{n-v}\right)^{\frac{1}{2}} = \left(\frac{1534.7}{6735-12}\right)^{\frac{1}{2}} = 0.4778$$

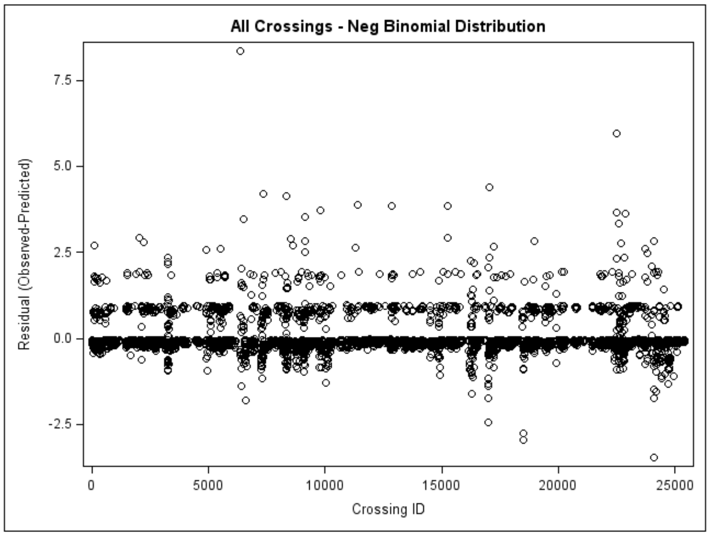


Figure 2.7. Scatter plot of the Residuals for the Negative Binomial Model

Table 2.11. Criteria for Negative Binomial Model

Criteria	Value
AIC (smaller is better)	5848.5
AICC (smaller is better)	5848.55
BIC (smaller is better)	5937.09

2.3.6. Zero Inflated Negative Binomial (ZINB) Model for all Crossings

A third model was also obtained after fitting a zero inflated negative binomial (ZINB) distribution to the data. As described in Section 2.2.3, this model can account for a greater proportion of observations with zero event counts, thus it was possible to obtain a better fit than with a standard negative binomial distribution. The deviance and Pearson chi-square goodness of fit for the ZINB model are shown in Table 2.12, indicating a better fit than with the NB model in the previous section. The results of the regression with the significance levels of each variable, as well as the distribution of the values compared to the data is shown in Table 2.13 and Figure 2.8. It is noted that the total number of crashes predicted by the ZINB model was the same as those from the NB model.

Table 2.12. Criteria for Assessing Goodness of Fit – ZINB Regression Analysis

Criterion	DF	Value	Value/DF
Deviance		5766.969	
Pearson Chi-Square	6722	6754.581	1.0048

Table 2.13. Results of the ZINB Regression for all crossings

Parameter	Category	DF	Estimate	Standard Error	Pr > ChiSq
Intercept		1	-3.3021	0.5665	<.0001
aadt3		1	4.90E-05	7.00E-06	<.0001
total_train3		1	0.0043	0.0014	0.0016
control_device	A	1	0.7077	0.5482	0.1968
control_device	B	1	0.7326	0.5476	0.1809
control_device	C	1	1.1188	0.5448	0.04
control_device	D	0	0	0	.
cross_angle3	1	1	0.4263	0.173	0.0137
cross_angle3	2	1	0.0553	0.0949	0.5597
cross_angle3	3	0	0	0	.
total_tracks		1	0.1164	0.0422	0.0058
traf_lanes3		1	0.1253	0.049	0.0106
hwy_near2	1	1	0.1404	0.0711	0.0484
hwy_near2	2	0	0	0	.
Dispersion		1	0.8088	0.1433	

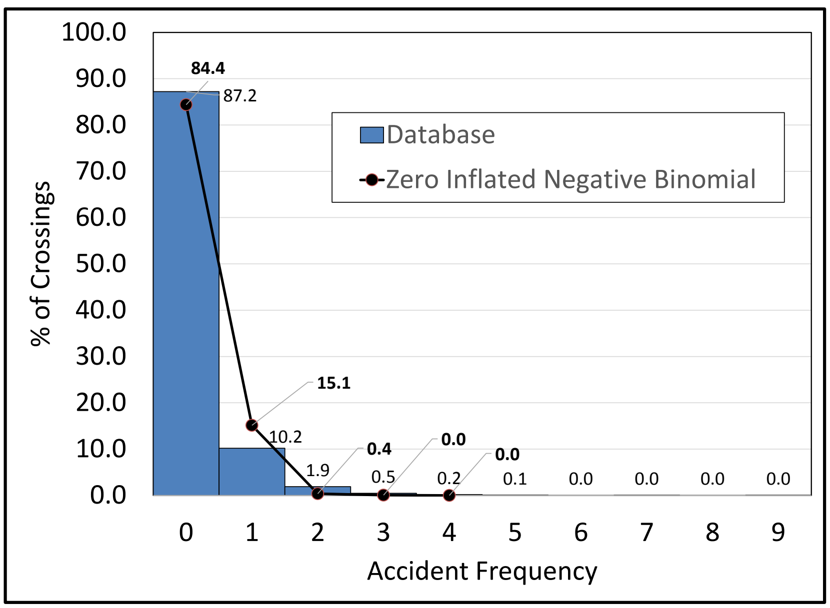


Figure 2.8. Results of the ZINB Model Compared to Field Data

The goodness of fit criteria for the ZINB model are provided in Table 2.14, Figure 2.9, and the following criteria:

$$SSE_{ZINB} = \sum_{1}^{6735} (y_i - f_i)^2 = 1539.6$$

$$RMSE_{ZINB} = \left(\frac{SSE}{n - v}\right)^{\frac{1}{2}} = \left(\frac{1539.6}{6735 - 12}\right)^{\frac{1}{2}} = 0.4785$$

It is noted that the zero inflated model has to estimate two additional parameters for the probability of an observation having a zero value: An intercept and an estimate for the total number of trains (the variable chosen for the zero model).

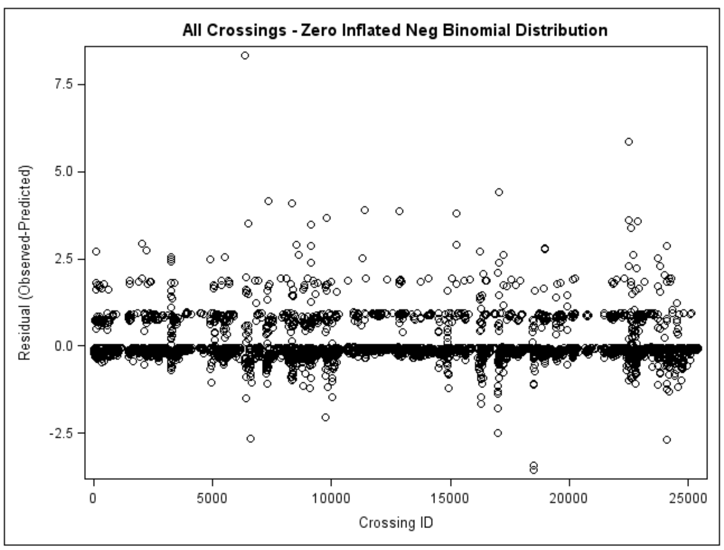


Figure 2.9. Scatter plot of the Residuals for the ZINB Model

Table 2.14. Criteria for ZINB Model

Criteria	Value
AIC (smaller is better)	5795.75
AICC (smaller is better)	5795.82
BIC (smaller is better)	5897.98

2.3.7. Summary of the Three Models by Warning Device

Results of fitting the three models to predict the accident frequency at the grade crossings in the FRA database showed that some (thoughsmall) improvement was obtained by changing from a Poisson distribution, to a negative binomial, to a zero inflated negative binomial.

Also, a closer look at the models reveals that the model performance across the different types of warning devices was not necessarily the same. Table 2.15 shows the residuals and the square standard errors (SSE) for the three models described above, by warning device: Passive (Device A), active (Device B), gates (Device C), and quad gates (Device D).

In particular, it is observed that the residuals and the SSE for the crossings with gates (Device C) had a larger deviation from the actual data, mainly due to greater accident frequencies in this group. Crossings with Device C were the only ones having 5 or more accidents (6 crossings), and also 9 of the 11 crossings with 4 accidents.

In addition, the proportion of zero values between warning device groups varies significantly for Device C. Specifically, for devices A, B, and D, about 92% to 95% of the crossings did not have accidents, compared to about 79% of the crossings with Device C.

Table 2.15. Summary of the Three Models by Warning Device

Model	Warning Device	N	Residuals		SSE	
			sum	var	sum	var
Poisson	Device A	1978	0.02	0.08	161.0	0.15
	Device B	1605	0.3	0.13	210.2	0.56
	Device C	3067	2.8	0.37	1139.2	3.33
	Device D	85	0.02	0.05	4.0	0.04
	Total	**6735**	**3.1**		**1514.5**	
Negative Binomial	Device A	1978	0.5	0.08	160.8	0.15
	Device B	1605	0.5	0.13	210.2	0.56
	Device C	3067	-14.6	0.38	1159.5	3.33
	Device D	85	-0.3	0.05	4.2	0.04
	Total	**6735**	**-13.9**		**1534.7**	
Zero-inflated Negative Binomial	Device A	1978	0.5	0.08	160.3	0.15
	Device B	1605	-0.1	0.13	209.5	0.54
	Device C	3067	-14.2	0.38	1165.4	3.32
	Device D	85	-0.6	0.05	4.3	0.04
	Total	**6735**	**-14.3**		**1539.6**	

These findings indicate that models created for each warning device, may offer advantages over the models just presented above. However, the sample size of crossings with quad gates is not enough for reliable model estimation, as it is mentioned in the following section, thus devices C and D (all types of gates) were modeled together. Also, predictions from such model could be directly compared to the current US DOT prediction formula.

The next section presents device-specific formulas created from a zero inflated negative binomial model, since this functional form dealt somewhat more favorably with the data over-dispersion and had the best goodness of fitness values.

2.4. Alternative Models by Warning Device

Models for the three categories of warning devices described in previous sections were created based on a zero-inflated negative binomial (ZINB) distribution. It is noted that even though the warning device categorization presented here is similar to the one used in the US DOT formula, data analysis indicated that separating crossing with quad gates from those with other gate configurations could be advantageous

(Figure 2.3). Quad gates provide a drive-around protection by blocking traffic at both entry and exit lanes, and this feature seems to play a significant role in accident frequency. However, the number of crossings with quad gates (85 in total) was too low for properly fitting the model, thus crossings with gates regardless of the gate configuration were analyzed altogether (Devices C and D).

In order to properly test the models, the original FRA dataset was divided into two subsets: one dataset for building the models, and one dataset for validation. The two datasets were defined such that there were the same number of crossings in each warning device category and the same number of crossings by accident frequency. The general summary of the two datasets is shown in Table 2.16.

Table 2.16. Building and validation datasets for model developing

Warning Device	Accident Frequency	Number of Crossings	
		Subset 1	**Subset 2**
Device A (Passive Warning Devices)	0	923	923
	1	60	60
	2	6	5
	3	0	1
	Total	**989**	**989**
Device B (Active Warning Devices)	0	741	741
	1	51	50
	2	8	9
	3	2	1
	4	1	1
	Total	**803**	**802**
Device C + D (Gates)	0	1273	1273
	1	232	232
	2	50	50
	3	14	13
	4	4	5
	5	2	2
	7	1	1
	Total	**1576**	**1576**
Total		**3368**	**3367**
		6735	

The models for each warning device type were generated using the same contributing factors as those described for all the crossings together and using the observations in Subset 1 from Table 2.16. After the coefficients were obtained, the models were applied to crossings in Subset 2 and the accidents were predicted. The US DOT model was also implemented and the expected accident frequency was obtained using the crossing characteristics, the past accident history, and the normalizing constant, as described in Section 2.1.2.

In light of the adjustment performed to the US DOT model's initial prediction using the accident frequency, in this section, the results from the ZINB model were also adjusted using the same methodology. In this way, the models could be compared by using the direct results from the ZINB regression, and also by adjusting them by the accident frequency.

In this particular case, there is an inherent advantage of adjusting the predictions by the accident frequency, since the objective is to predict the exact same number of accidents that have occurred in the analysis period. A perfect model is expected to predict the same accident frequency that has been observed in the past, assuming that the crossing conditions, conflicting volumes, or any other contributing factor do not change. This is precisely the situation modeled in this case, thus adjusting by the accident history will always bias the model results towards improved predictions. A different situation may be observed for true prediction, when past history and the desired prediction may differ. This case is explored in the next section.

The ZINB models and the results for each of the warning devices, together with the US DOT predictions and comparisons with the actual data, are described next.

2.4.1. Crossings with Passive Warning Devices (Device A)

The initial ZINB model for this warning device type was the following:

$$Accidents = \left(1 - \frac{1}{1 + e^{3.8023-(0.0976*trains)}}\right)$$
$$* \, e^{-3.0052+(0.000151*aadt)+(0.005593*trains)-(0.1140*tracks)-(0.05865*lanes)+f_{angle}+f_{near}}$$

$$f_{angle} = \begin{cases} 0.1843 & if\ 0 < angle < 30 \\ -0.1601 & if\ 30 \leq angle < 60 \\ 0 & if\ 60 \leq angle \leq 90 \end{cases}$$

$$f_{near} = \begin{cases} 0.1951 & if\ distance\ to\ nearest\ intersection < 200ft \\ 0 & otherwise \end{cases}$$

Where *trains* and *aadt* are the daily train and vehicular volume, *tracks* and *lanes* are the number of tracks and lanes, and f_{angle} and f_{near} are factors for the different categories of crossing angle and distance from nearest intersection, respectively. The first term in the equation is of the form (1- φ), where φ represents the probability of a crossing of having zero observations (the zero-model), followed by a factor with the negative binomial model.

However, even though the results of the original regression are shown in the equation above, including all variables explored in the previous chapter, not all variables turned out to be significant. The Pvalues of each variable and the final decision about including it in the model or not is shown in Table 2.17.

Table 2.17. Regression results from the ZINB model for passive warning devices (Device A)

Variable	Category	Initial Regression		Final Regression
		Pvalue	Selected for Final Model	Pvalue
Intercept	-	< 0.0001	✓	< 0.0001
aadt	-	0.0328	✓	0.0383
total_train	-	0.008	✓	0.0098
cross angle	1 (0°-29°)	0.8071	x	N/A
	2 (30°-59°)	0.6346		
	3 (60°-90°)	-		
total_tracks	-	0.6603	x	N/A
traf_lanes	-	0.8008	x	N/A
hwy_near	1 (0-200ft)	0.1951	x	N/A
	2 (>200ft)	-		
Intercept (zero model)	-	0.35	✓	0.35 *
total_train (zero model)	-	0.16	✓	0.18 *

* Model fit improved significantly with the zero-inflated model even though the Pvalues for the zero model estimates were high. A type 3 analysis for total_train showed a Pvalue of 0.10, showing significance in the model. Model convergence was only achieved if the ZINB distribution was fitted

Based on the variables selected in Table j2.17, and confirming an improvement in the model results by means of the goodness of fit criteria (Table 2.18), the model for crossings with passive warning devices was only defined in terms of the daily train and vehicular volume, as follows:

$$Accidents = \left(1 - \frac{1}{1 + e^{4.2446-(0.1015*trains)}}\right) * e^{-3.1167+(0.000138*aadt)+(0.05069*trains)}$$

Table 2.18. Criteria for Assessing Goodness of Fit – ZINB Passive Warning Devices

Criterion	DF	Value	Value/DF
Pearson Chi-Square	979	979.952	1.001

2.4.2. Crossings with Active Warning Devices (Device B)

In a similar way, the initial regression analysis for active warning devices (Device B) included all variables described for Device A, but not all variables were significant predictors. It was noted that the regression for the zero inflated model indicated that the total train volume was not a good predictor of the probability of zero accidents. Instead, the AADT showed to be a better predictor for this particular set of devices and it was used to determine the results of the initial regression for variables selection, shown in Table 2.19.

Table 2.19. Regression results from the ZINB model for active warning devices (Device B)

Variable	Category	Initial Regression		Final Regression
		Pvalue	Selected for Final Model	Pvalue
Intercept	-	< 0.0001	✓	< 0.0001
aadt	-	0.0107	✓	0.0089
total_train	-	0.0107	✓	0.0226
cross angle	1 (0°-29°)	0.1021	✓	0.0746
	2 (30°-59°)	0.0533		0.0667
	3 (60°-90°)	-		-
total_tracks	-	0.1378	x	N/A
traf_lanes	-	0.0025	✓	0.0044
hwy_near	1 (0-200ft)	0.6657	x	N/A
	2 (>200ft)	-		
Intercept (zero model)	-	0.0453	✓	0.0438
aadt (zero model)	-	0.0276	✓	0.0275

After observing these results, the regression analysis was performed again without the total number of tracks and the variable for the distance from the crossing to the nearest intersection. The final regression without these variables improved the model fit (Table 2.20) and resulted in the following expression, using the coefficients from the last column in Table 2.19:

$$Accidents = \left(1 - \frac{1}{1 + e^{-(4.8251-(0.0003*aadt))}}\right) * e^{-3.8738+(0.000132*aadt)+(0.02451*trains)+(0.4523*lanes)+f_{angle}}$$

$$f_{angle} = \begin{cases} 1.0324 & if\ 0 < angle < 30 \\ -0.8104 & if\ 30 \le angle < 60 \\ 0 & if\ 60 \le angle \le 90 \end{cases}$$

Table 2.20. Criteria for Assessing Goodness of Fit – ZINB Active Warning Devices

Criterion	DF	Value	Value/DF
Pearson Chi-Square	795	781.29	0.983

2.4.3. Crossings with Gates (Devices C+D)

The initial regression for crossings with gates showed low influence of the crossing angle and the number of traffic lanes. All other variables were significant. The final regression improved the model fit by having the zero-inflated term, similar to the observation for crossings with active warning devices. The significance of each variable in the regressions is shown in Table 2.21.

Table 2.21. Regression results from the ZINB model for crossings with gates (Devices C and D)

Variable	Category	Initial Regression		Final Regression
		Pvalue	Selected for Final Model	Pvalue
Intercept	-	< 0.0001	✓	< 0.0001
aadt	-	< 0.0001	✓	< 0.0001
total_train	-	< 0.0001	✓	< 0.0001
cross angle	1 (0°-29°)	0.18	x	N/A
	2 (30°-59°)	0.12		
	3 (60°-90°)	-		
total_tracks	-	0.0005	✓	0.0003
traf_lanes	-	0.18	x	N/A
hwy_near	1 (0-200ft)	0.044	✓	0.0283
	2 (>200ft)	-		-
Intercept (zero model)	-	0.2586	✓	0.2521*
total_train (zero model)	-	0.2563	✓	0.2495*

The final Pearson chi-square results (Table 2.22) and the full model for crossings with gates is also shown below:

$$Accidents = \left(1 - \frac{1}{1 + e^{-(40.52-(0.2317*trains))}}\right) * e^{-2.5701+(0.000055*aadt)+(0.01037*trains)+(0.2227*tracks)+f_{near}}$$

$$f_{near} = \begin{cases} 0.2511 & if\ distance\ to\ nearest\ intersection < 200ft \\ 0 & otherwise \end{cases}$$

Table 2.22. Criteria for Assessing Goodness of Fit – ZINB Crossings with Gates

Criterion	DF	Value	Value/DF
Pearson Chi-Square	1569	1577.93	1.006

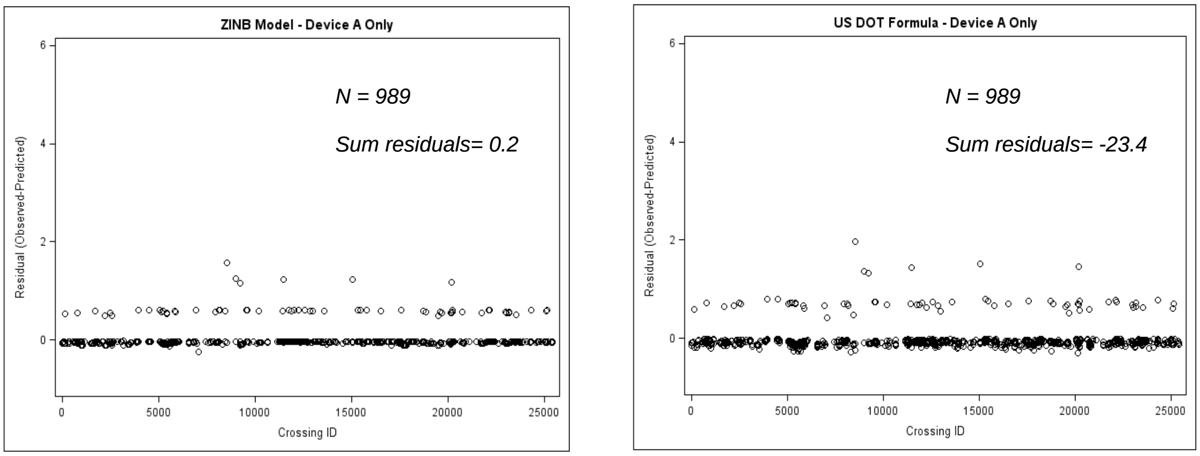
2.5. Comparison of US DOT Accident Prediction Formula and Alternative Models - by Device

After defining the models for each of the warning devices, comparisons were made between the model results and the results from the US DOT accident prediction formula in terms of general prediction accuracy, the overall distribution of the predictions, and the ability to reproduce the crossing ranking based on actual accident frequency. Recall that models were developed using a subset of the total data, and a completely separate set of observations for the validation or comparison of the models, as described in section 2.4 and shown in Table 2.16. The same validation dataset used for the ZINB models was used for the predictions with the US DOT formula.

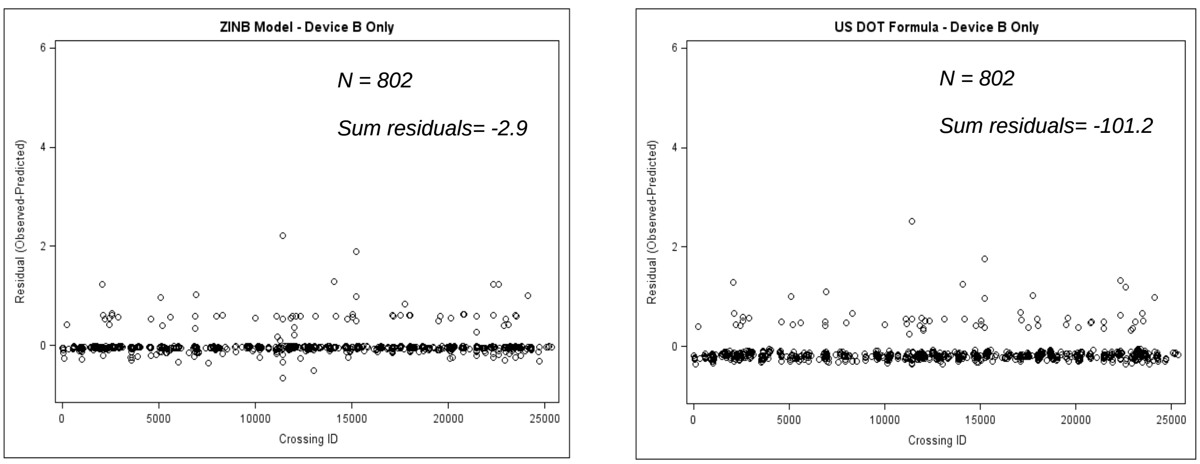
2.5.1. General Prediction Accuracy

For the general prediction accuracy, the residuals were observed to identify biases in the model predictions. The residuals were defined as the difference between the actual accident count and the predicted value, and were plotted for each warning device, as it is shown in Figure 2.10. It is noted that even though the crossings are labeled from 1 through 25000+, there were only as many points as crossings per device category, thus for device A there were 989 points, for Device B there were 802 points, and for Device C+D there were 1576 points.

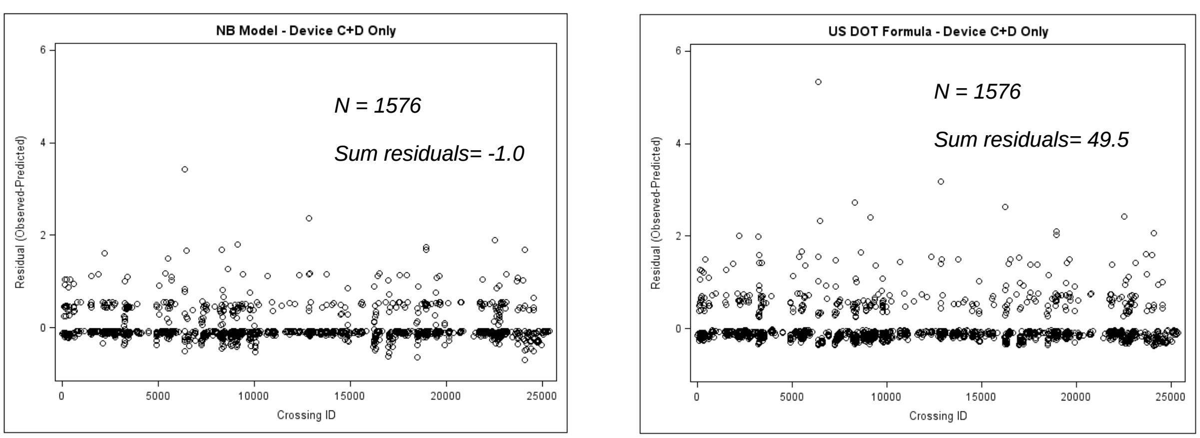
From the residual plots, it is observed that both models have a tendency to have negative residuals at crossings with zero accidents, given that the prediction value is always greater than zero. Also, the overall ZINB predicts a better overall number of accidents (sum of residuals) than the US DOT model, which is expected given that the regressions were developed using the "test data", which is similar than the "validation data". In addition to the residuals, a better description of the over and under prediction can be observed by plotting the predicted and the observed values and observing where the predictions fall with respect to a diagonal line (predicted = observed). These plots are shown in Figure 2.11.



A- Device A (ZINB model and US DOT formula)

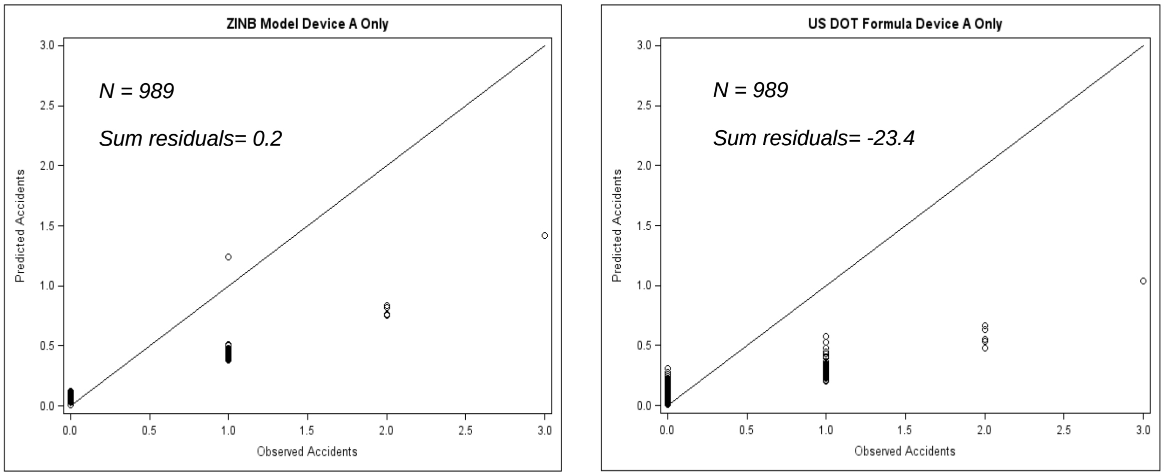


B- Device B (ZINB model and US DOT formula)

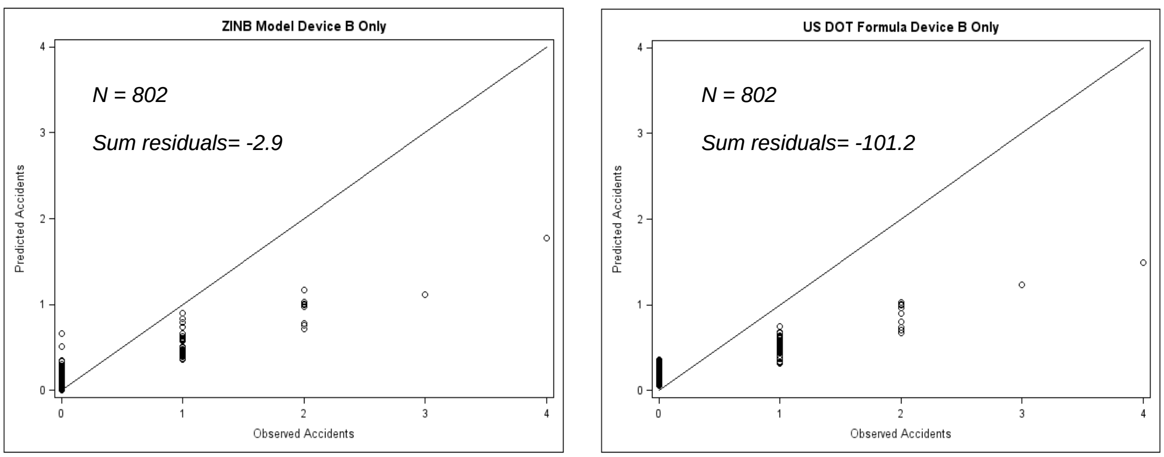


C- Devices C and D (ZINB model and US DOT formula)

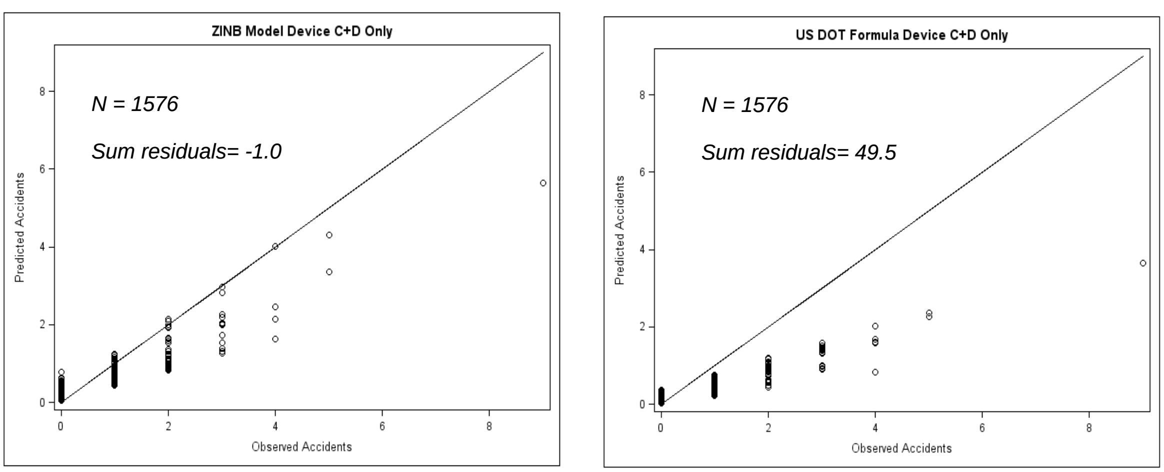
Figure 2.10. Scatter plot of the Residuals using the Validation Subset



A- Device A (ZINB model and US DOT formula)



B- Device B (ZINB model and US DOT formula)



C- Devices C and D (ZINB model and US DOT formula)

Figure 2.11. Predicted and Observed Accident Frequencies using the Validation Subset

The sum of squared errors (SEE) was also observed for the models and the US DOT formula, as a measure of the accuracy of the model predictions. A summary of the SEE is shown in Table 2.23.

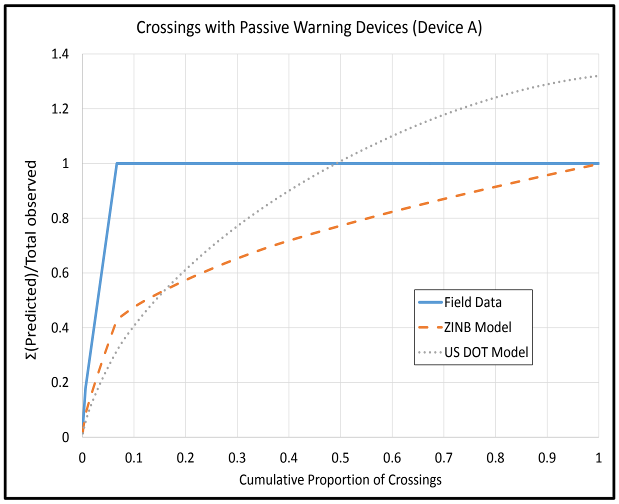
Table 2.23. Sum of Squared Errors for ZINB and US DOT Formula

Warning Device	Sum of Squared Errors (SSE)			
	ZINB Model		US DOT Formula	
	sum	var	sum	var
Device A	32.1	0.02	50.6	0.05
Device B	38.6	0.07	62.0	0.08
Device C+D	164.6	0.19	291.0	0.86
Total	235.3	-	403.5	-

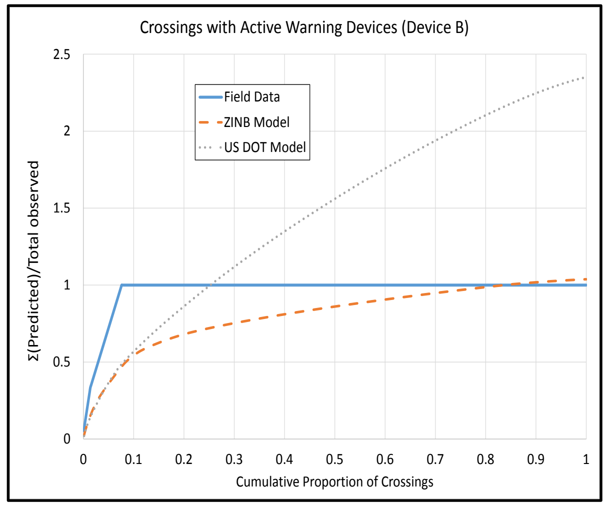
2.5.2. Overall distribution of the predictions

An aspect that was also important to determine for the models was the deviation of the cumulative number of crashes over the cumulative number of crossings. This is a measure similar to the power factors used by XX to evaluate the models in the development of the US DOT formula in 1980. Two plots were generated for each warning device category: one for the proportion of crashes with respect to the actual data , and one for the proportion of the accidents relative to each model.

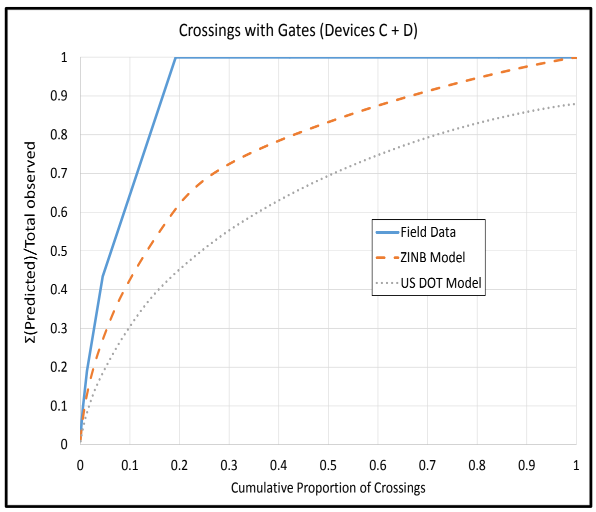
The plots for the proportion of crashes with respect to the actual data show how the cumulative predictions match the total number of crashes when the crossings are sorted from highest to lowest. Figure 2.12 shows absolute prediction accuracy and also the shape of the function in absolute terms. Absolute (or raw) predictions are important as the magnitude of the predicted number of crashes could be used in benefit/cost analysis and resource allocation rankings.



A- Passive Warning Devices



B- Active Warning Devices



C- Crossings with Gates

Figure 2.12. Absolute predictions for the ZINB Model and the US DOT Formula

2.5.3. Crossing Rankings

Lastly, the comparison of the models and the US DOT formula includes the rankings based on accident frequency. These rankings were found by sorting the crossings by the predicted accident frequency from largest to smallest. Different top lists were generated and compared between the field data, the ZINB models and the US DOT formula, as shown in Tables 2.24, 2.25, and 2.26.

Table 2.24. Ranking for crossings with passive warning devices (Device A)

Data or Model	Total Accidents		
	Top 5 Crossings	Top 10 Crossings	Top 66 (all crossings with accidents)
Field Data	12	18	74
ZINB Model	11	18	74
US DOT Model	11	18	66

* Max accident frequency at a crossing = 3

Table 2.25. Ranking for crossings with active warning devices (Device B)

Data or Model	Total Accidents			
	Top 5 Crossings	Top 10 Crossings	Top 20 Crossings	Top 61 (all crossings with accidents)
Field Data	13	23	29	75
ZINB Model	13	20	29	73
US DOT Model	13	22	29	71

* Max accident frequency at a crossing = 4

Table 2.26. Ranking for crossings with gates (Devices C and D)

Data or Model	Total Accidents			
	Top 5 Crossings	Top 10 Crossings	Top 20 Crossings	Top 303 (all crossings with accidents)
Field Data	27	45	75	410
ZINB Model	26	42	71	398
US DOT Model	27	44	70	373

* Max accident frequency at a crossing = 9

2.5.4. An Example of True Prediction

In this section, an example to illustrate the actual prediction ability of the DOT prediction formula and the alternative model is presented. Five years of data (from 2003 to 2007) was used to develop an alternative model and to predict the number of crashes that were observed from 2008 to 2012 (also a 5-year span).

The zero-inflated negative binomial (ZINB) model described in previous sections was used to obtain the predictions. Regressions were run for each of the three groups of warning device types (passive, active warning devices, and gates) to obtain the model coefficients. This first prediction was then corrected by the actual accident history of each of the crossings using not only the procedure described in the DOT prediction formula (and applied above), but also using two additional methods that could be considered for future use: an empirical Bayes approach, and the average of the prediction and the past accident history.

The empirical Bayes approach described by Persaud et al. (2001) was implemented for the prediction correction. In this approach, the initial prediction from the model is updated based on the past accident frequency at a given crossing, the length of the time period analyzed, and the dispersion parameter estimated in the model. This correction was used by Persaud et al. to obtain estimates of the accident frequency for a before-after study if the location being analyzed was not modified in the after period. Thus, in our case, it provides with an estimation of the accidents assuming that no changes took place at the crossing during the prediction period. The formulation for the expected annual number of accidents in the before period (m_b), assumed the same for the after period, is the following:

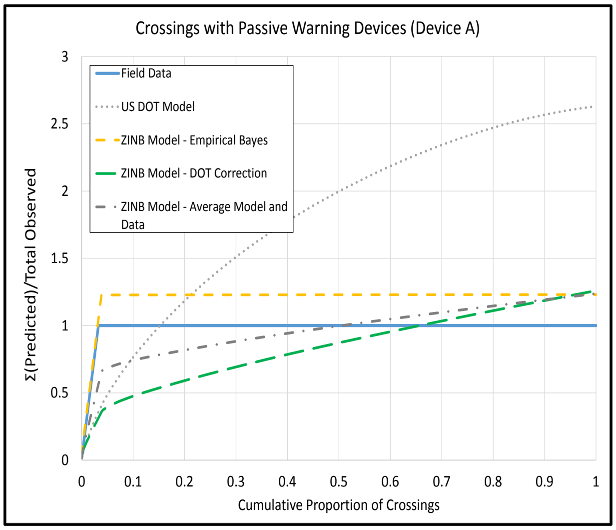
$$m_b = \frac{k + x_b}{\frac{k}{P} + y_b}$$

Where k is the dispersion parameter estimated in the model, $x_b$is the actual number of crashes observed in the before period (i.e. from 2003 to 2007), and y_b is the length of the before period.

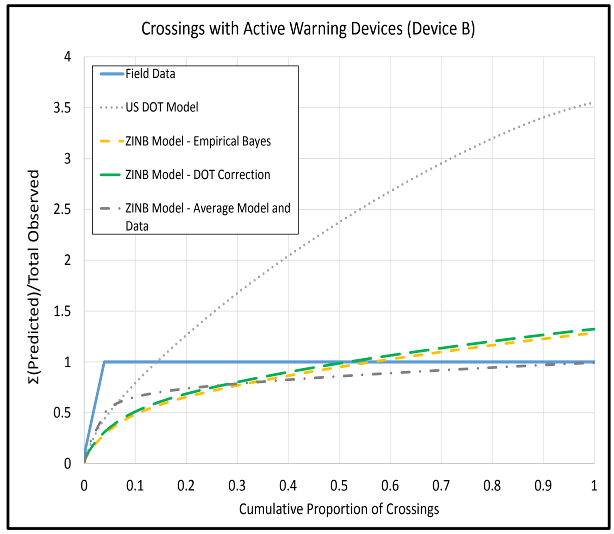
It is important to note that this exercise does not use the accident frequency to be predicted at any point, and thus it is an example of true prediction. However, even though the performance of the two models (US DOT and the ZINB) for this particular dataset is indicative of their capabilities, it is not a direct estimation of their performance in other datasets. With this caveat, the results of the prediction models are shown as follows.

Also, it is noted that for the US DOT Formula, the normalizing constant issued by FRA in 2007 was used to estimate the accidents between 2008 and 2012.

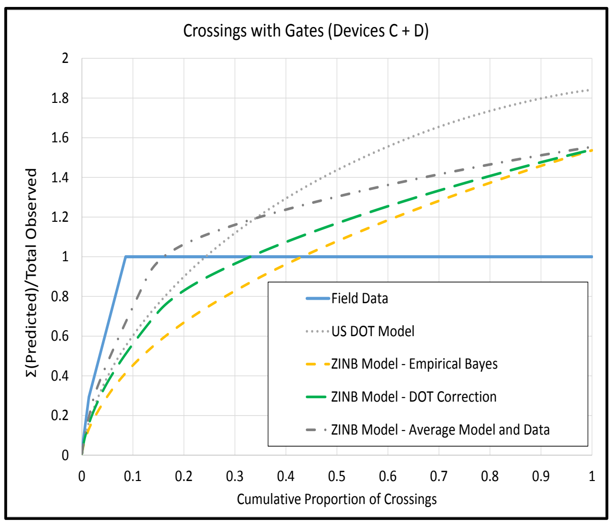
Figure 2.13 shows the absolute number of accidents predicted by the two models and using different adjustments of the initial predictions for the ZINB model.



A- Passive Warning Devices



B- Active Warning Devices



C- Crossings with Gates

Figure 2.13. Absolute predictions for 2008-2012 with the ZINB Model and the US DOT Formula

From figure 2.13, it is observed how the two models compared with the actual distribution of the total accidents observed from 2008 to 2012. The US DOT formula produced predictions that overestimated the accident frequency for all three types of devices, by a factor between 1.8 and 3.6 times the observed accidents, whereas the ZINB model had a closer fit with factors ranging between 1 and 1.6 times the observed accidents depending on the adjustment method.

The type of correction applied to the initial model prediction did have a significant effect, mostly on the shape of the curve for the accident distribution. Higher accident frequencies were observed at crossings with active warning devices and gates, where the corrections using an average of the predicted and the past accident history had the greatest effects and projected the initial portion of the curve closer to the actual data. On the other hand, for locations with passive warning devices, the empirical Bayes placed significant weight to the past history and concentrated most accidents in a few locations and significantly closer to the data.

In general, regardless of the correction used to include the past history, the ZINB model predicted a more accurate total number of accidents for all three groups compared to the results from the US DOT formula.

Nonetheless, the general distribution of the accidents is only one of many possible criteria to compare the two models. The crossing rankings obtained from the two models and their comparison with the rankings based on actual accidents observed from 2008 to 2012 were also investigated, as shown in Table 2.27.

Table 2.27. Rankings predicted by different methods for 2008-2012

Warning Device	Ranking Method	Number of crashes predicted in top locations			Cumulative accidents observed in top locations		
		Top 10	Top 20	Top 50	Top 10	Top 20	Top 50
Passive (Device A)	Data (observed)	11	21	51	11	21	51
	US DOT Formula	6	10	19	1	2	7
	ZINB - Empirical Bayes	14	24	54	2	3	7
	ZINB - DOT correction	6	9	17	1	2	7
	ZINB - Average model and data	8	14	29	2	3	7
Active (Device B)	Data (observed)	16	26	56	16	26	56
	US DOT Formula	8	13	26	4	5	8
	ZINB - Empirical Bayes	6	9	17	3	5	6
	ZINB - DOT correction	7	10	18	3	5	5
	ZINB - Average model and data	11	19	35	2	7	9
Gates (Devices C and D)	Data (observed)	28	48	101	28	48	101
	US DOT Formula	18	31	58	11	23	44
	ZINB - Empirical Bayes	15	25	46	13	24	40
	ZINB - DOT correction	18	30	56	13	22	40
	ZINB - Average model and data	20	35	69	11	19	33

Table 2.27 shows two types of results from the models and contrasted with the actual data. The first one is the number of crashes predicted in top locations, with result shown for top 10, 20, and 50

locations for each type of warning device. To obtain these values the data was sorted based on the accident prediction values, and predictions of the top locations were added to determine how many accidents were expected at the locations with higher risk values. In general, models underestimate the number of accidents at these locations, particularly those with higher frequencies, as it has been shown in previous sections. The predictions of the ZINB models fluctuated significantly depending on the correction for accident history that was applied, without a single correction being dominant in all three types of warning devices.

The second set of results in the last three columns of Table 2.27 shows the actual number of accidents observed in the top locations chosen by each model. As expected, prediction with macro models for low accident frequencies in the past, such as those crossings with passive warning devices is more difficult and models did not accurately rank most of the locations compared to the ranking based on the actual data. For example, in the top 50 locations of all models, only 7 accidents were actually observed between 2008 and 2012, whereas in the ranking by frequency it is possible to find 51 accidents in the top 50. Predictions improved with higher frequencies, being the most accurate the ones found for the top locations with gates.

It is noted that given the randomness involved in accident prediction for low frequencies, more accurate modeling of the accidents for a group of crossings may not necessarily result in a more accurate ranking. This can be seen with the ZINB models, which from Figure 2.13 are observed to have a better fit for the accident distribution, but resulted in similar rankings to those from the US DOT formula.

3. Micro Analysis of Accidents at Railroad-Highway Grade Crossings

As described in the Introduction, in addition to a macro analysis, accidents could be analyzed using a micro approach to include variables that may be significant at a local or microscopic level. Such procedure may be able to account not only for traditional variables such as conflicting volumes and geometry, but also for variables that describe the position and direction of trains and vehicles, areas surrounding the crossing, driver demographics, and the environmental conditions when the accidents occurred, among others.

The proposed microscopic analysis may be similar to manual or informal procedures already in place at some jurisdictions, but the methodology proposed here attempts to enumerate a few basic and practical steps to identify trends based on crash history using a tree diagram and a sketch of the crossing. Results of the micro analysis can then be combined with results from the macroscopic model for a more accurate identification of crossings for safety improvements.

In this chapter, the steps to perform a simple micro analysis are described, followed by examples of applications using data from Illinois. The methodology for micro analysis is in its simplest form and targeted for manual implementation at the level of a single crossing at a time. The authors plan to extend this approach to more complex tasks with automated sorting of the attributes in the tree structure, temporal analysis of the accident occurrences, and for corridor and regional analysis.

3.1. Description of the Micro Analysis Procedure

The procedure for the proposed micro analysis for one single crossing is simple and it is described as follows:

1. The location of the train and vehicles or pedestrians involved in each of the accidents is mapped using a marker (e.g. circle, star) on a sketch of the crossing that can be easily constructed based on an aerial image. Markers should also be numbered to associate accident locations with other accident characteristics.

2. In the sketch, a table with the position, direction, and speed of train and vehicles is created. This table will help in visualizing possible patterns when all information for the crossing is extracted.

3. Then, a tree diagram is created to classify the accidents using the following variables: highway user type involved in the accident, action of the motorist, highway user direction, time table direction, circumstance of accident, vehicle driver age and gender, and weather condition and visibility. An illustration of the variables and their possible values is shown in Figure 3.1. The tree diagram helps identify trends in the accidents and therefore, possible contributing factors.

4. Trends observed in the tree structure can be further explored using information extracted from other variables, and images of the crossing, among other sources.

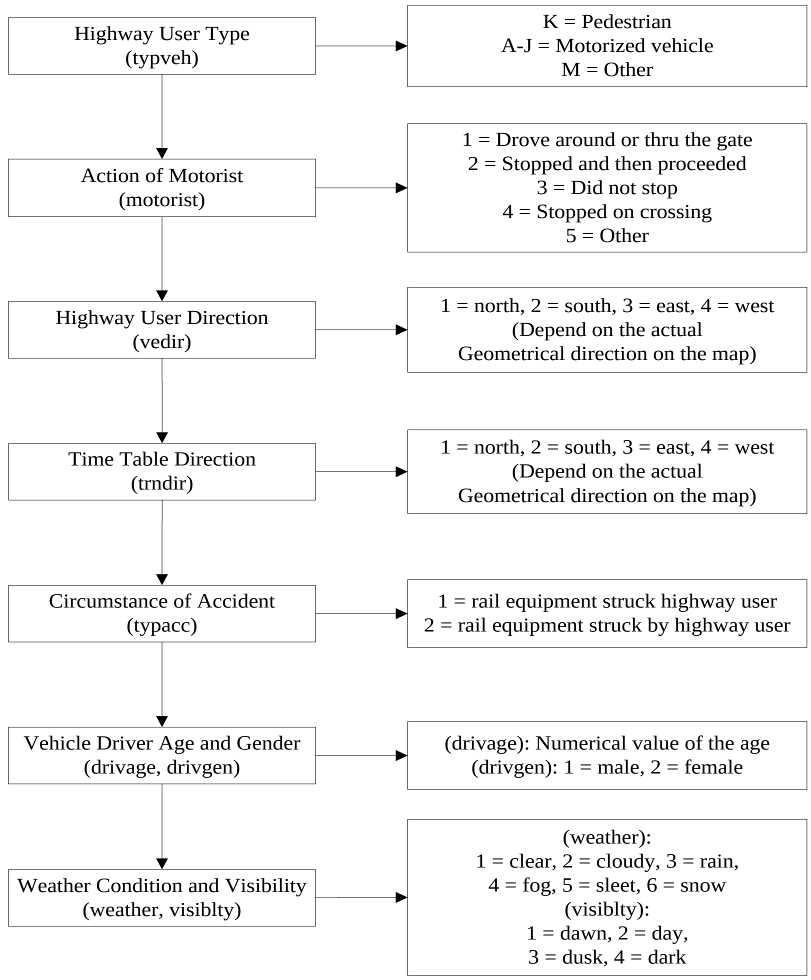
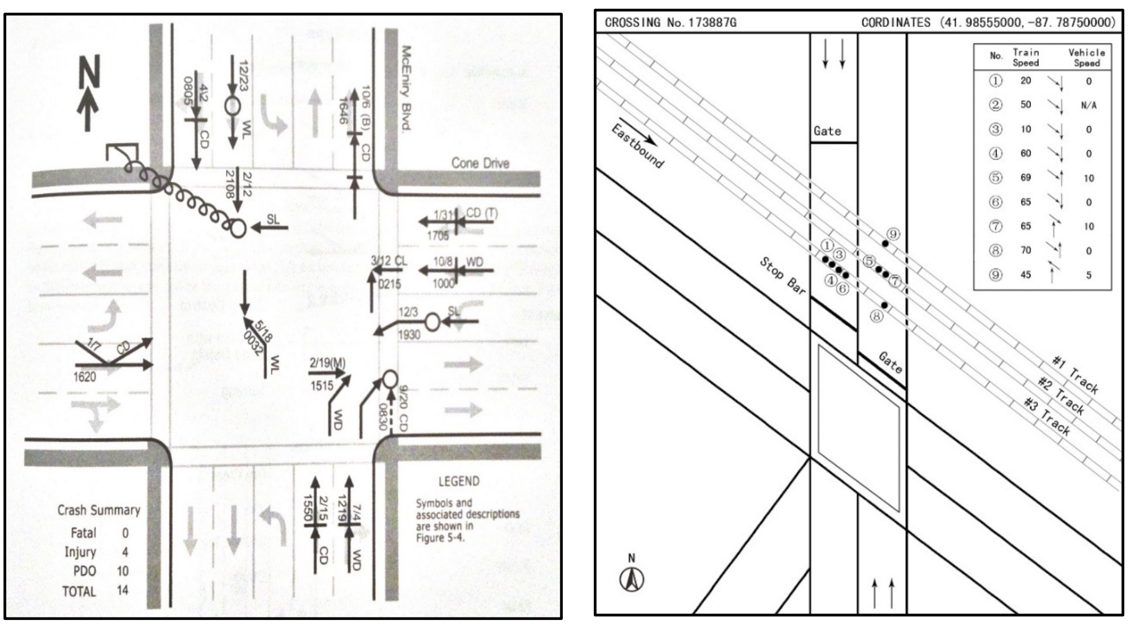


Figure 3.1. Variables for proposed micro analysis

The proposed micro analysis has similarities and differences with other proposed methods for detail analysis of accidents. Two of such methods are the diagnosis used for roadway safety management process from the Highway Safety Manual (AASHTO, 2010), and a field investigation procedure specifically for grade crossings proposed in the Railroad-Highway Grade Crossing Handbook, Revised Second Edition, 2007.

In the HSM, the diagnosis has three main steps: 1) review previous safety data from police reports, 2) access other supporting documentation about the site condition, such as construction plans, land use mapping, etc., which could be used to provide historical site context and define the roadway environment, and 3) to conduct field investigation to the site, which could serve to validate the previous data gathered from police report and other supporting documentation.

In the first step, the previous safety data is compiled from police reports to create descriptive crash statistics, which could be used to find specific type of crashes that exceed the threshold proportion. The statistics include multiple types of crash data which could be categorized as crash identifiers (e.g. date, day of week, time of day), crash type (e.g. rear-end, sideswipe, angle, etc.), crash severity, sequence of events (e.g. direction of travel and location of parties involved), and contributing circumstances (e.g. parties involved, road condition, lighting condition, etc.). In addition, crashes are summarized by location using three different diagrams or maps: collision diagrams, condition diagrams and crash mapping. The collision diagram shows the crashes represented by arrows with the direction of the vehicles and the type of crashes, vehicle type, crash type, vehicle movement, severity, road surface and lighting. This diagram is similar to the sketch produced in the micro analysis, which was developed for grade crossings. Additional information not included in a collision diagram and shown in the sketch includes the speeds of the train and vehicles. Examples of the collision diagram from HSM and the proposed sketch are shown in Figure 3.2.



A- HSM Collision Diagram B- Sketch in micro analysis

Figure 3.2. Collision Diagram from HSM and Proposed Sketch in Micro Analysis

Other diagrams described in the HSM include the condition diagram and crash mapping. The condition diagram is a sketch of the site containing several important characteristics, including roadway features (e.g. lane configuration, traffic control, pedestrian, bicycle, and transit facilities in the vicinity of the site, etc.), land uses and pavement conditions. These collision and condition diagrams can be integrated together to show a further relation between crashes and road conditions. The crash mapping is a way of using GIS to conduct spatial analysis for crashes, such as analyzing crash data with other GIS data (e.g. the presence of schools, posted speed limit signs, etc.), report crash clusters by certain query conditions and showing the crash density along a corridor.

The sketch proposed in the micro analysis is similar to the end result of the first step in the HSM diagnosis, but it focuses on aspects specific to grade crossings and can be completed without some of the details specified in the HSM. In addition, the second step in the diagnosis includes data from the site condition, but it does not have a defined method for trend discovery. In the micro analysis, the tree structure provides such method, and serves as a visualization tool not included in the HSM diagnosis or in the Railroad-Highway Grade Crossing Handbook. Moreover, a dynamic tree structure is in its development process and will serve for more complex analysis of not only a single location, but also for corridor and regional analysis. The tree structure will be complemented by a temporal analysis of past accident frequencies to enhance the understanding and predictions of future crashes.

Lastly, the procedure in the Railroad-Highway Grade Crossing Handbook is aimed at reviewing the crossing and its environment, identifying the nature of the problem and recommending alternative improvements. This diagnostic team method could be categorized into three areas: traffic operation, traffic control devices and administration. A questionnaire is used during the field investigation, and three study points are used to inspect the surrounding features of a certain crossing. The first point is where the driver first obtain the information of a crossing ahead. The data collected at this point are concerned with the visibility of the crossing, effectiveness of advance warning signs and signals, etc. The second point is where a safe stop could be made by the driver before approaching the crossing. The data collected at this point include obstruction of view of train approach, availability of information for proper stop or go decisions by the driver, etc. The third point is located 15 feet from the nearest rail. Data collected at this point include sight distance down the tracks, pavement markings, etc. The investigation result could be used together with the accident frequency, accident type and accident circumstances to evaluate a certain crossing.

The nature itself of the micro analysis is different from that in the Handbook, since it is an evaluation of the accident trends before a field visit is scheduled, and it is expected to enhance the information available in the preparation of such field visits.

Thus, the idea of a micro analysis falls between a very detailed analysis of the crash site prior to a field visit, currently included in the HSM, and the current macro analysis techniques, where local trends are not identified. The authors believe that the tree structure to visualize trends at a single location, and for corridor and regional analysis can be very useful, and even more in a future implementation that automates the process of sorting the attributes to maximize trend discovery. Also, a new temporal analysis of accidents will be added to the micro analysis in the near future.

3.2. Dataset from FRA

According to the FRA online inventory data, there were 1459 accidents in Illinois from 2003 to 2012. Table 3.1 shows a summary of the accidents based on different vehicle type and accident type. It also contains the dollar amount in property damages caused by different type of accidents.

The number of accidents per year are shown in Figure 3.3, including fatal and non-fatal accidents in the same time period and indicate a downward trend in the total number of accidents and in the non-fatal accidents. However, fatal accidents did not show any clear trend and remained at a similar level over time.

In addition, Figure 3.4 shows the ratio of motorized vehicle accidents, pedestrian accidents and other accidents by each year. The number of motorized vehicle accidents account for the majority of all accidents and had a downward trend, similar to the total number of accidents, while the number of pedestrian accidents and other accidents did not show any particular trend.

Table 3.1. Accident summary by year and type

Year of Accidents	Type of Accidents	No. of Accidents	Non-Fatal	Fatal	Fatalities	Prop. Damage
2003	Total	169	141	28	30	652601
	Motorized Vehicles	150	132	18	20	560001
	Others	7	5	2	2	92600
	Pedestrians	12	4	8	8	0
2004	Total	178	153	25	28	636137
	Motorized Vehicles	156	141	15	18	584537
	Others	9	7	2	2	51600
	Pedestrians	13	5	8	8	0
2005	Total	180	152	28	31	955250
	Motorized Vehicles	150	133	17	20	818350
	Others	15	13	2	2	136900
	Pedestrians	15	6	9	9	0
2006	Total	174	157	17	25	990000
	Motorized Vehicles	147	136	11	19	869500
	Others	16	12	4	4	115500
	Pedestrians	11	9	2	2	5000
2007	Total	160	134	26	29	1095148
	Motorized Vehicles	135	122	13	16	1065848
	Others	10	8	2	2	29300
	Pedestrians	15	4	11	11	0
2008	Total	153	127	26	27	896000
	Motorized Vehicles	126	111	15	16	832300
	Others	12	11	1	1	63700
	Pedestrians	15	5	10	10	0
2009	Total	106	91	15	18	664580
	Motorized Vehicles	85	79	6	9	630780
	Others	12	9	3	3	33800
	Pedestrians	9	3	6	6	0
2010	Total	128	102	26	28	544985
	Motorized Vehicles	99	88	11	11	460485
	Others	10	8	2	2	84500
	Pedestrians	19	6	13	15	0
2011	Total	103	86	17	19	569638
	Motorized Vehicles	84	76	8	10	567338
	Others	4	3	1	1	2300
	Pedestrians	15	7	8	8	0
2012	Total	108	85	23	26	429106
	Motorized Vehicles	85	77	8	11	388181
	Others	10	7	3	3	40925
	Pedestrians	13	1	12	12	0
2003-2012	Grand Total	1459	1228	231	261	7433445
	Motorized Vehicles Total	1217	1095	122	150	6777320
	Others total	105	83	22	22	651125
	Pedestrians Total	137	50	87	89	5000

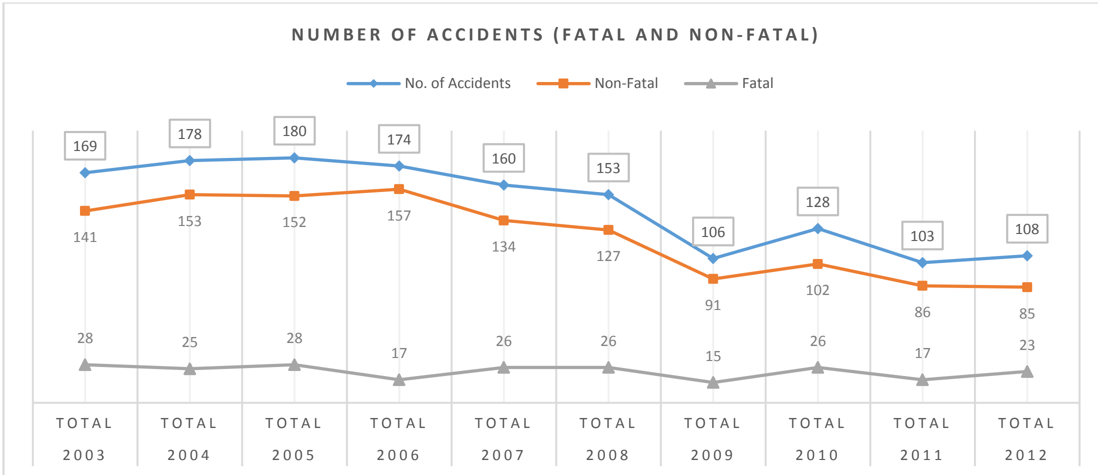


Figure 3.3. Accidents from 2003 to 2012 by type

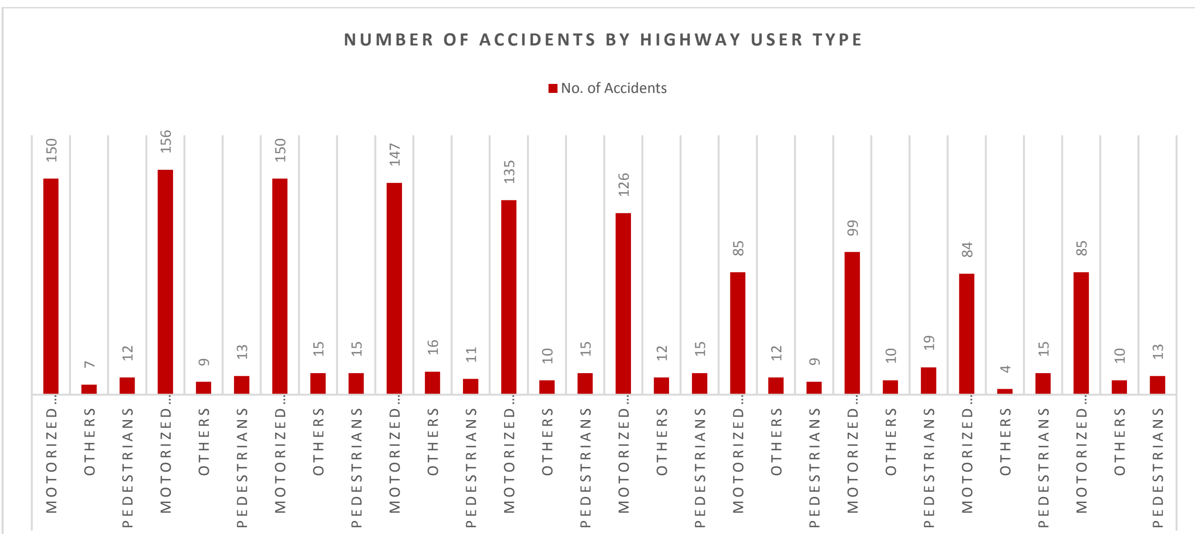


Figure 3.4. Accidents from 2003 to 2012 by type

In 2012, Illinois had 14842 open at grade railroad crossings, including public, private, and pedestrian only crossings. A summary of the accident frequency at these crossings by crossing type is shown in Table 3.2, from which the focus will be given to public crossings. This is because public agencies have direct jurisdiction over this type of crossings and also these are typically the locations with highest crossing volumes.

Table 3.2. Accident frequency of different types of crossings

	Type of Crossings			
Accident Frequency From 2003 to 2012	Pedestrian Only Crossing	Private Crossing	Public Crossing	Grand Total
9	0	0	1	1
8	0	2	0	2
7	0	1	1	2
6	0	4	0	4
5	0	1	4	5
4	0	2	11	13
3	0	3	31	34
2	2	190	136	157
1	12	90	765	867
0	400	3930	9427	13757
Grand Total	414	4052	10376	14842

Thus, the micro analysis will be conducted for public crossings with high accident frequency, which for the purposes of this study include locations with four or more accidents from 2003 to 2012. These crossings were identified and are shown Table 3.3 below and analyzed in the following sections.

Table 3.3. Public Crossings with 4 or more accidents from 2003 and 2012

Accident Frequency	Number of crossings	Crossing Inventory Number
9	1	173887G
7	1	608311K
5	4	175042V 176909P 372131E 386411X
4	11	079493L 174001M 283190L 289554E 291378J 294341E 328516E 608310D 608846J 609011A 724578T

3.3. Micro Analysis for Crossings with More than 4 Accidents

The first crossing to be analyzed is the crossing with highest accident frequency, with nine accidents in the period between 2003 and 2012. Crossing 173887G is located at the crossing of UP-NW line and N

Nagle Avenue in Chicago, as shown in the satellite image in Figure 3.5. The right side of image 3.4 shows a sketch of the crossing with a summary of the location of the accident, the direction of cars and trains, as well as their speeds. This sketch is part of the micro analysis and may help spotting possible trends at this crossing, for example the occurrence of four accidents involving southbound vehicles with trains in the southbound direction and on track #3.

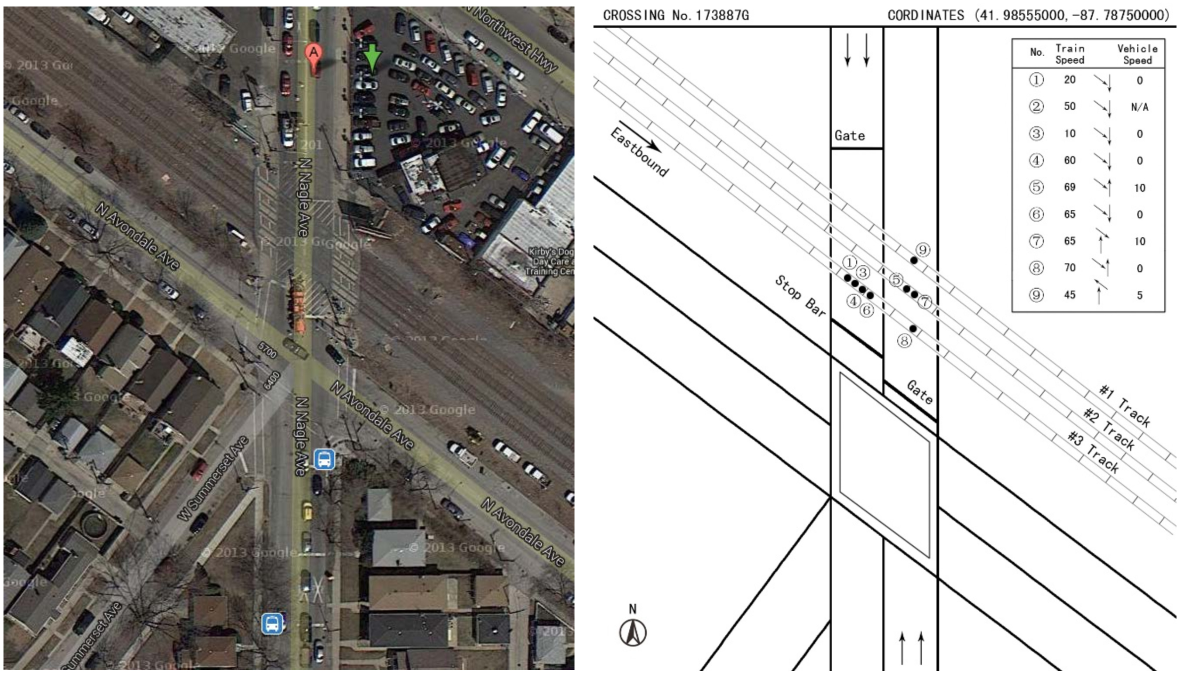


Figure 3.5. Image and sketch of Crossing 173887G

The tree structure also provides a useful representation of the accidents, as shown in Figure 3.6. Out of the nine accidents, one of them involved a pedestrian and eight involved vehicles. Out of the accidents involving a vehicle, in four occasions the user drove around or through the gate and in the remaining four they were recorded as having other actions. All four accidents recorded as driving around or through the gate involved vehicles going northbound, while all the 4 other accidents involved vehicles going southbound and hitting by the trains going eastbound (they all happened on track #3). Thus, it is worthwhile exploring the possibility of vehicles waiting for the nearby signal or in long queues and not being able to prevent the accidents while occupying track #3. Also, it is noted that in Figure 3.6, the tree branches with more prominent trends are highlighted red for visualization purposes. This trends could not be observed in a macro analysis and may be useful for evaluating the risk at the crossing and studying potential safety improvements.

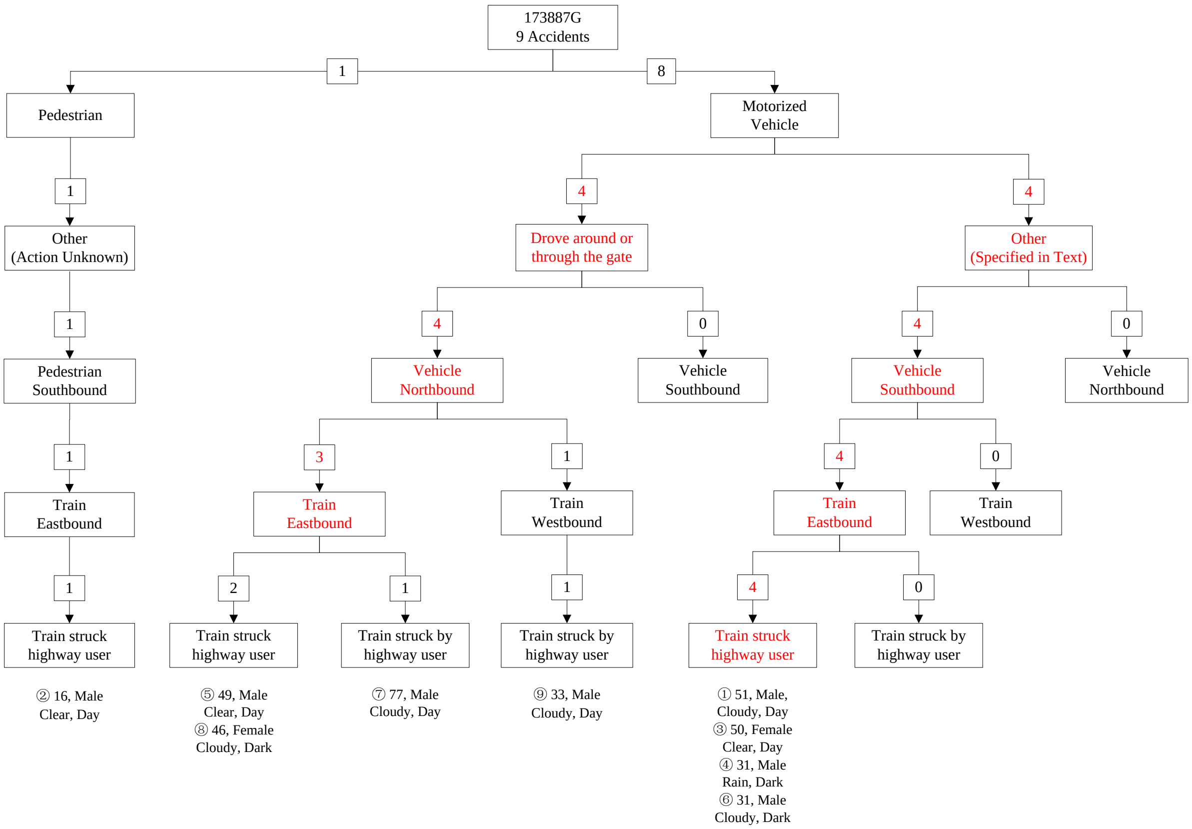


Figure 3.6. Tree diagram for accidents at Crossing 173887G

The second crossing analyzed was 608311K. This crossing is located on W 119th Street in Blue Island and had 7 accidents over a period of 10 years. The aerial image and the sketch of the crossing are shown in Figure 3.7 where it can be seen that accidents were concentrated in tracks 1 and 2.

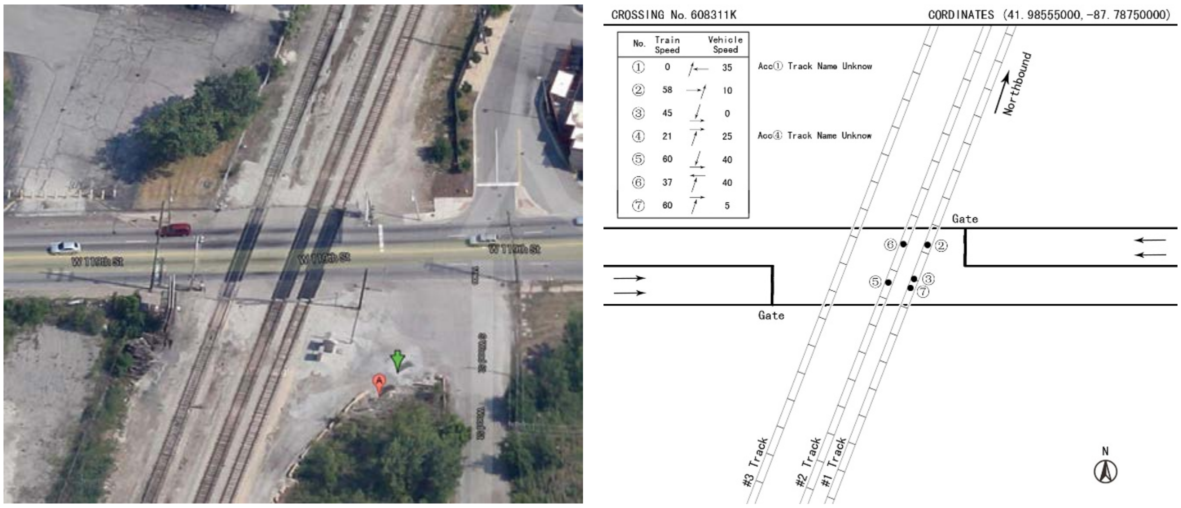


Figure 3.7. Image and Sketch of Crossing 608311K

The seven accidents at this crossing involved vehicles, 5 of which drove around or through the gate, 1 stopped on crossing, and 1 was recorded as other. At this crossing the tree structure shows no clear trend from those 5 accidents where drivers went around or through the gates. This is an example of a location without a clear trend, and where accidents may be the result of higher exposure due to high train and vehicle volume (aadt is 21100 and train volume is 66 trains per day).

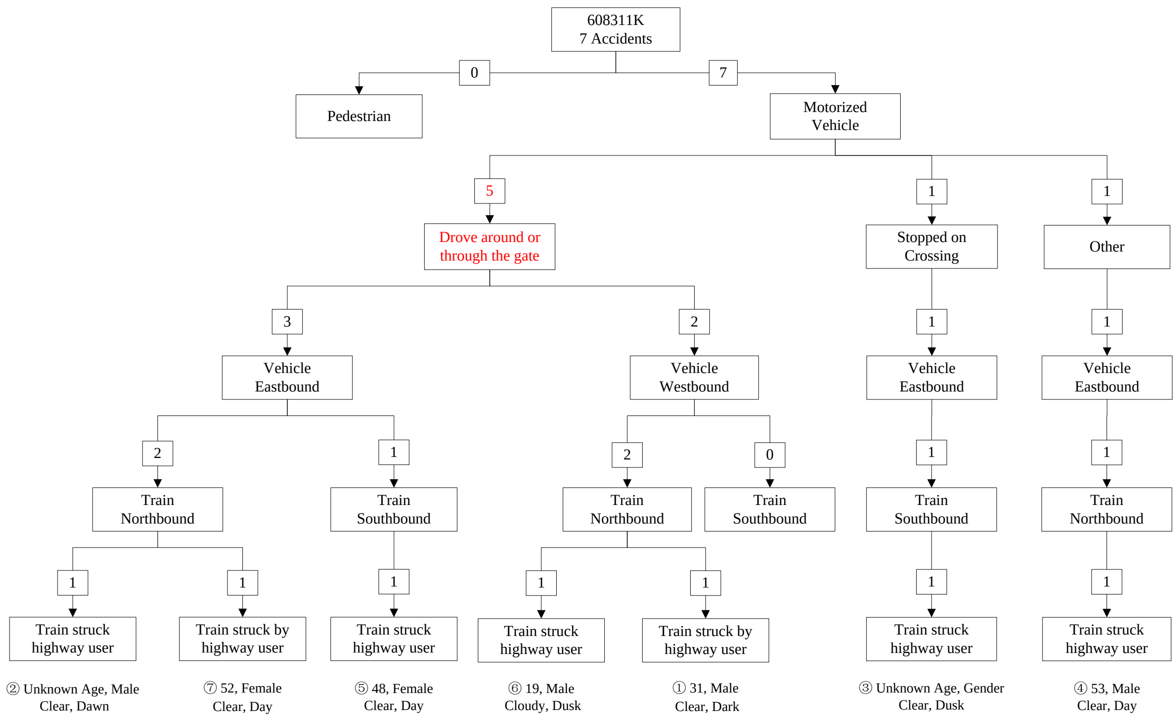


Figure 3.8. Tree diagram for accidents at Crossing 608311K

The third analyzed crossing is 175042V, which is located at the crossing of a UP rail line and E Lincoln Highway in Dekalb. Five accidents happened at this crossing over 10 years, as it is shown in the sketch in Figure 3.9.

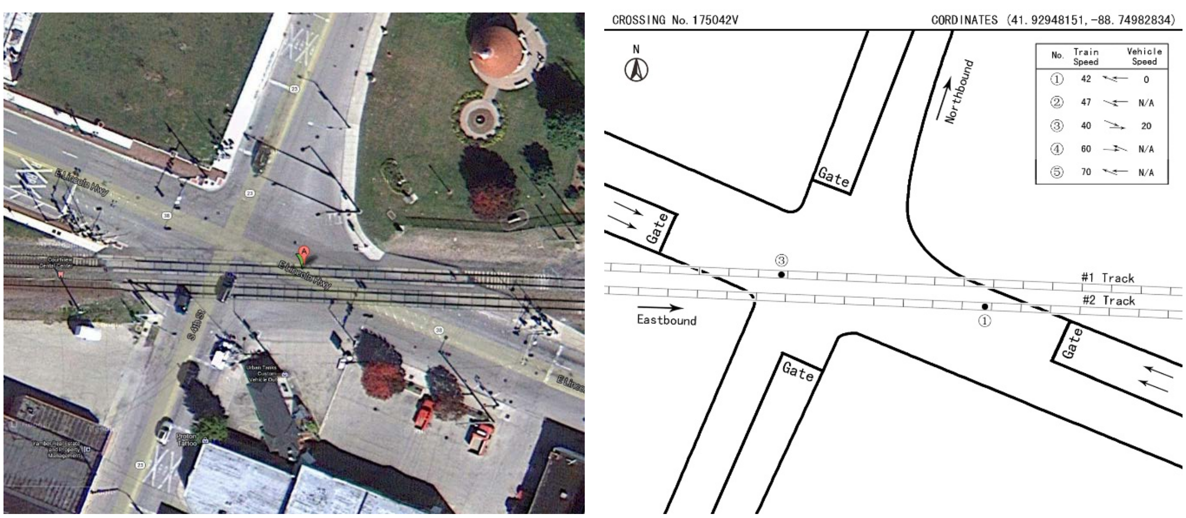


Figure 3.9. Image and Sketch of Crossing 175042V

In these 5 accidents, three of them were pedestrian accidents and the other two involved vehicles. This is a highly unusual frequency of pedestrian accidents, and based on the location of the crossing, an area with relatively high pedestrian traffic near the Northern Illinois University campus. From the tree structure (Figure 3.10), it is also noted that the train and vehicles (and one of the pedestrian cases) were traveling in the same direction. The combination of a small crossing angle (Figure 3.9), users and trains in the same direction, and nighttime conditions (red branches in the tree structure), seem to be an area worth further exploration in the causes of these accidents. Visibility at this crossing should probably be looked into based on the results of the micro analysis.

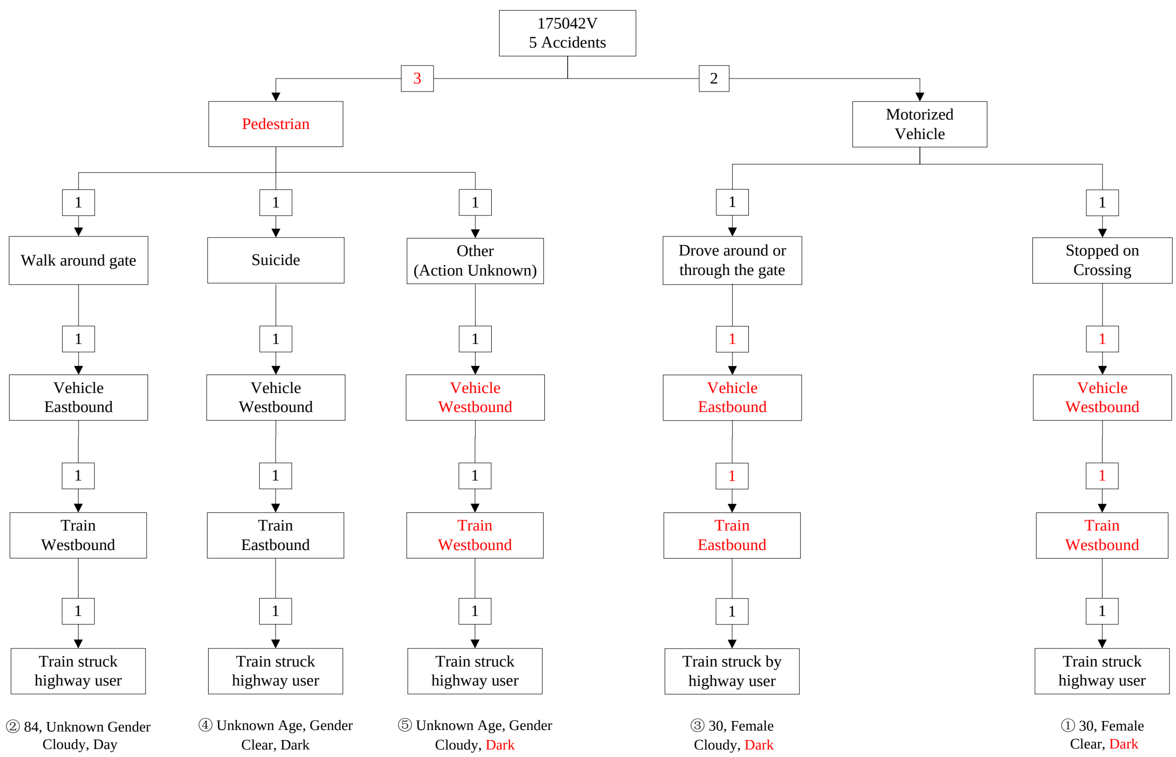


Figure 3.10. Tree diagram for accidents at Crossing 175042V

The forth crossing analyzed is 176909P, which is located at the crossing of a UPME rail line and S Mt Prospect Rd in Mt Prospect. Five accidents happened at this crossing over 10 years (Figure 3.11).

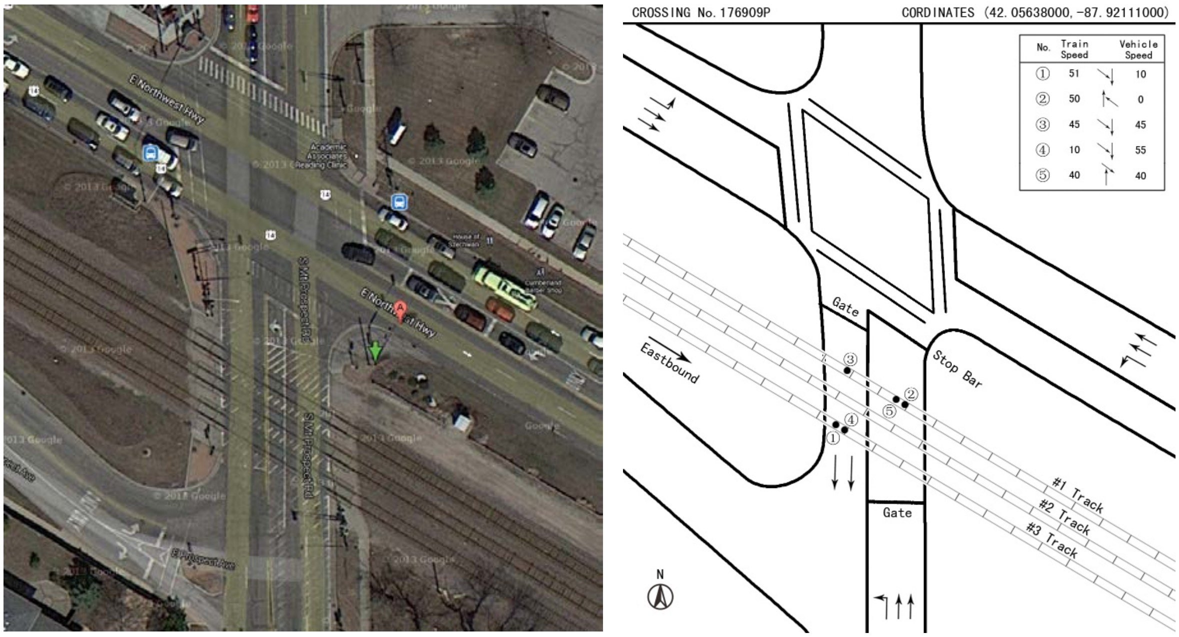


Figure 3.11. Image and Sketch of Crossing 176909P

At this crossing all five accidents involved vehicles, four of them with users driving around or through the gates and one of them stopping before the gates. From Figure 3.12, it is seen that in three occasions southbound vehicles were struck by eastbound trains. Even though it is not clear the reason for these accidents, it is worth exploring the causes of this combination of train-vehicle directions to evaluate safety improvements.

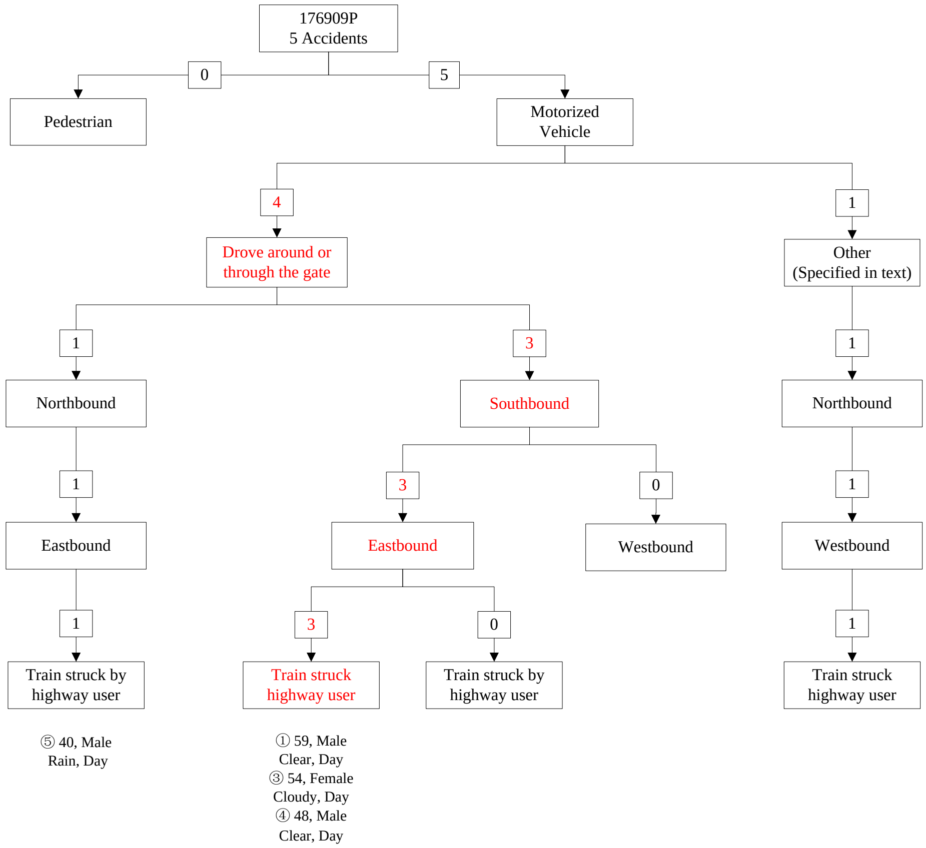


Figure 3.12. Tree diagram for accidents at Crossing 176909P

The fifth analyzed crossing is 372131E, which is located at the crossing of a Metra line and Grand Avenue in Elmwood Park. Five accidents happened at this crossing over 10 years, as seen in Figure 3.13.

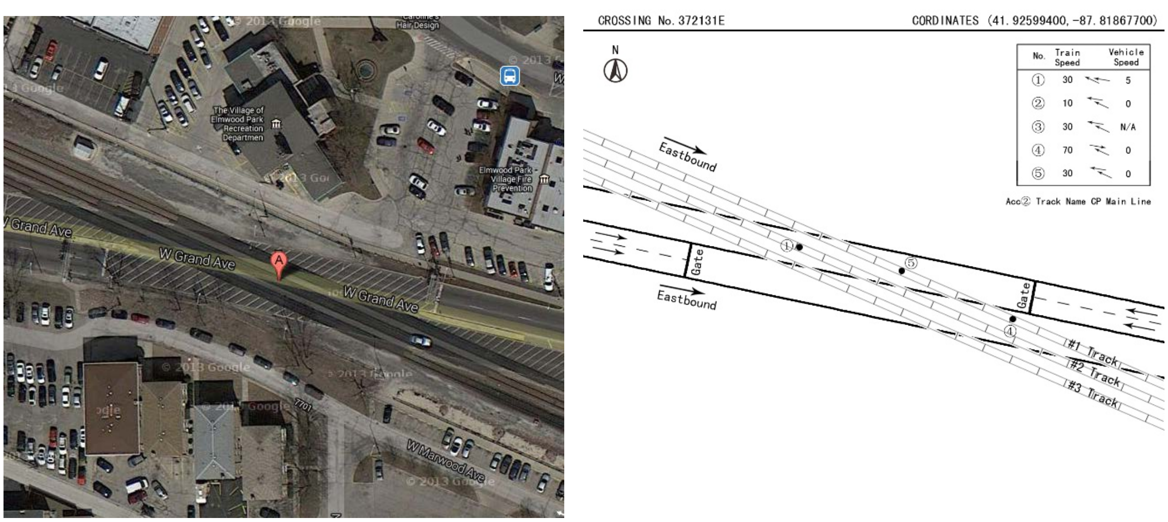


Figure 3.13 Image and Sketch of Crossing 372131E

In these 5 accidents, 1 of them was a pedestrian suicide and the remaining four involved vehicles. Unlike other previous crossings, users in accidents did not drive around or through the gate. The main trend spotted in the tree structure (Figure 3.14) is vehicles and trains traveling in the same direction when the accident happen, which combined with the narrow crossing angle may be an indicator of visibility issues.

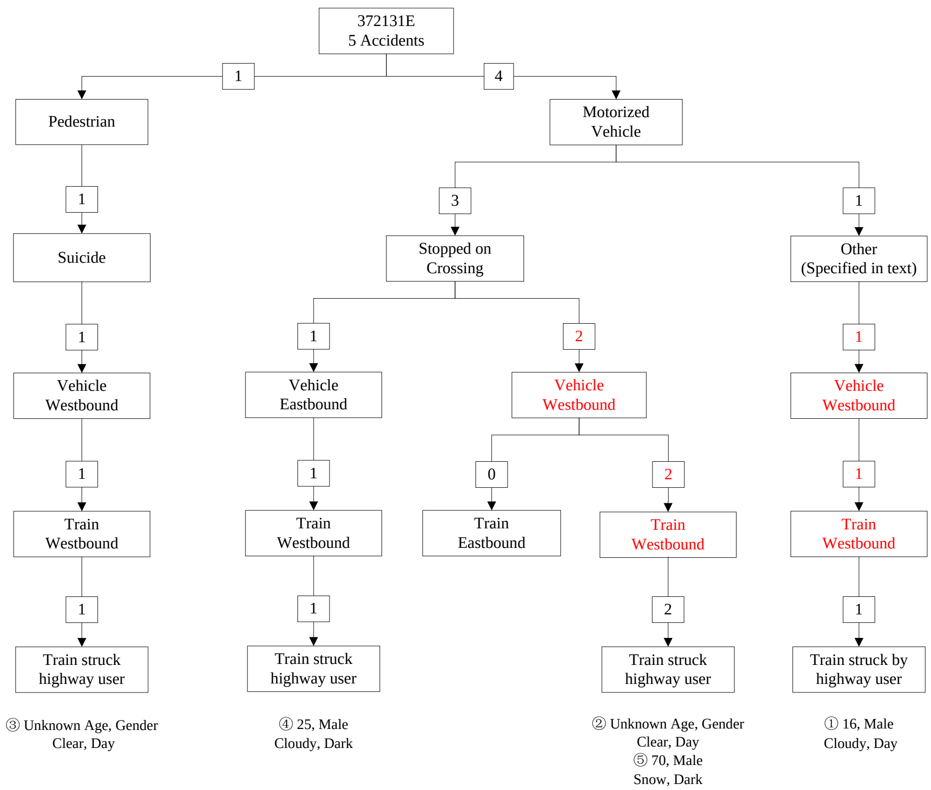


Figure 3.14. Tree diagram for accidents at Crossing 372131E

In addition, it was observed that the roadway has raised medians right before the crossing, preventing users from driving around the gates, as it is seen in Figure 3.15. This treatment is thought to be the significant factor for the different user behavior at this location compared to previous crossings.

Figure 3.15. Westbound (left) and Eastbound (right) Gates at Crossing 372131E

The sixth analyzed crossing is 386411X, which is located at the crossing of a Metra line and Grand Avenue in Elmwood Park. Five accidents happened at this crossing over 10 years. An image and the sketch of this crossing are shown in Figure 3.16.

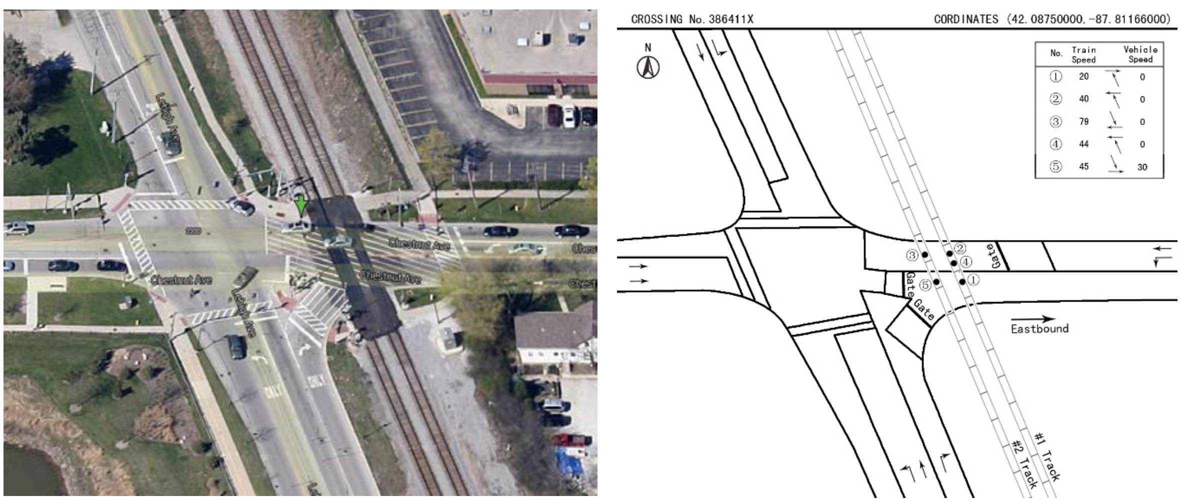


Figure 3.16. Image and Sketch of Crossing 386411X

All five accidents at this crossing involved vehicles. It was also noted that this crossing is equipped with pedestrian gates, also reducing the risk of such events. The tree structure (Figure 3.17) shows that the most common characteristic of these crossings is the advanced age of the drivers, with four of them older than 80 years old and one of them older than 60, combined with three accidents during nighttime. This finding prompted for an exploration of the surroundings of the crossing, where it was noted that the area has a high number of assisted living facilities (Figure 3.18), explaining the predominance of older drivers in accidents at the crossing. Neither of these trends could have been observed by a macro model, and results from this simple micro approach enhance the information available to inspectors before site visits are conducted.

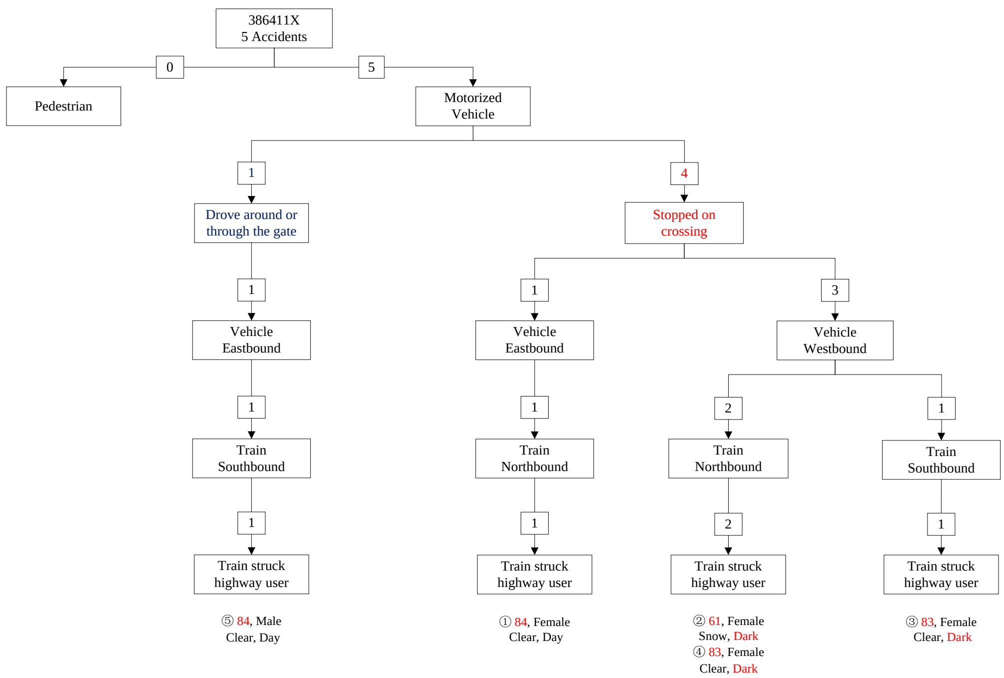


Figure 3.17. Tree diagram for accidents at Crossing 386411X

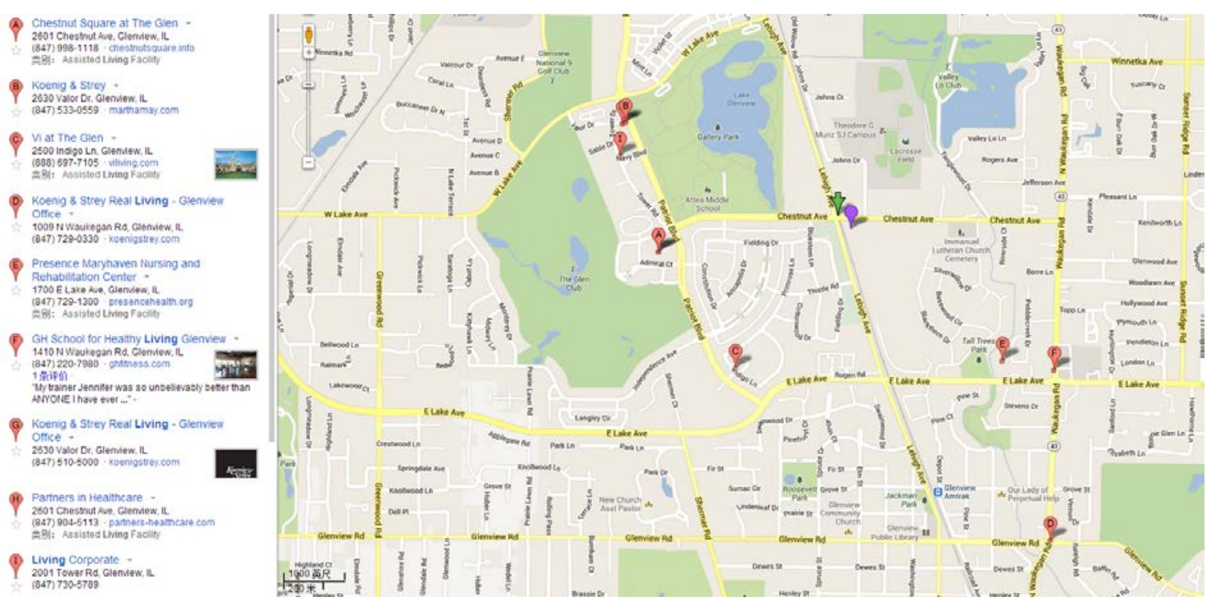


Figure 3.18. Assisted Living Facilities around Crossing 386411X

In summary, based on the previous analysis, the locations with more than four accidents from 2003 to 2012 can be grouped in two main types of crossings, as shown in Figures 3.19 and 3.20: 1) crossings with a nearby intersection with a short distance between the stop bar and the tracks, and 2) crossings with a small crossing angle (smaller than 30 degrees).

Even though there is no certainty about the causes of the accidents at these locations, the micro analysis has provided pointers to trends that could be further investigated to analyze the risk at these crossings. These trends would not have been flagged by a macro analysis, and may help assessing a more accurate measure of risk at these locations.

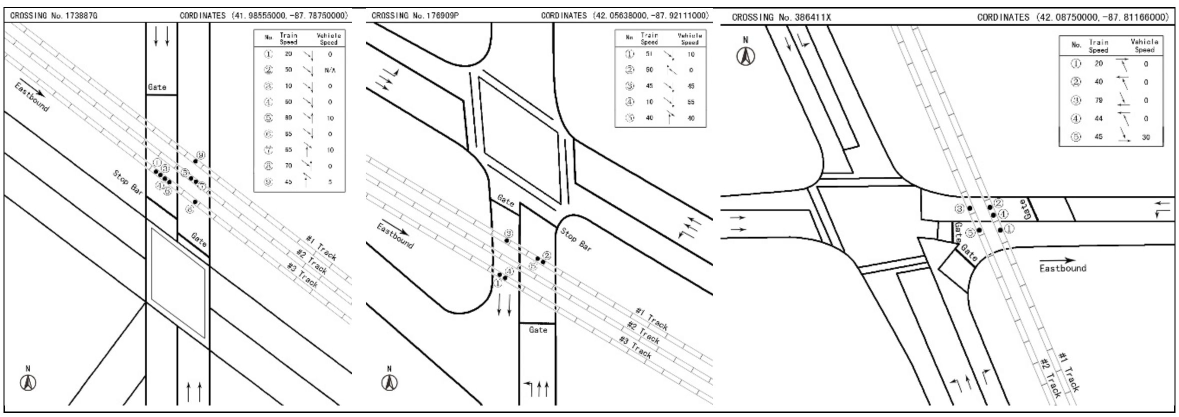


Figure 3.19. Type 1 Crossing with High Accident Frequency

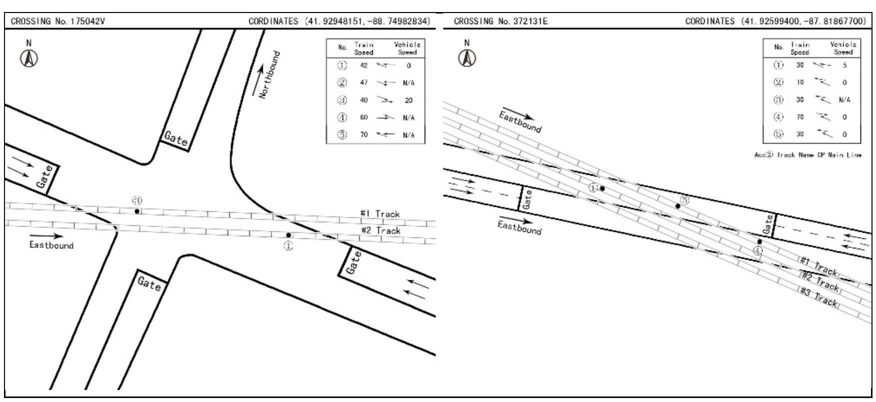


Figure 3.20. Type 2 Crossing with high Accident Frequency

3.4. Micro Analysis for Crossings with Four Accidents

The micro analysis was also conducted for the eleven locations with four accidents between 2003 and 2012. With fewer accidents, trends were not always clear to observe in the sketch or the tree structure, however two of the cases with the most apparent trends are shown in this section for illustration purposes:

1. Crossing 328516E, located on the east side of Decatur, IL, near the intersection of N Brush College and East Faries Parkway, has an AADT of 11500 and an average of only eight trans per day moving at slow speeds. This crossing is equipped with cantilever flashing lights and had no gates. As shown in Figure 3.21 all accidents happened at low vehicle and train speeds, with vehicles traveling southbound and trains traveling eastbound. This trend can be clearly seen in the tree structure in Figure 3.22. The geometry of the crossing does not pose difficulties in terms of the approaching angle, and the visibility seems clear of obstacles.

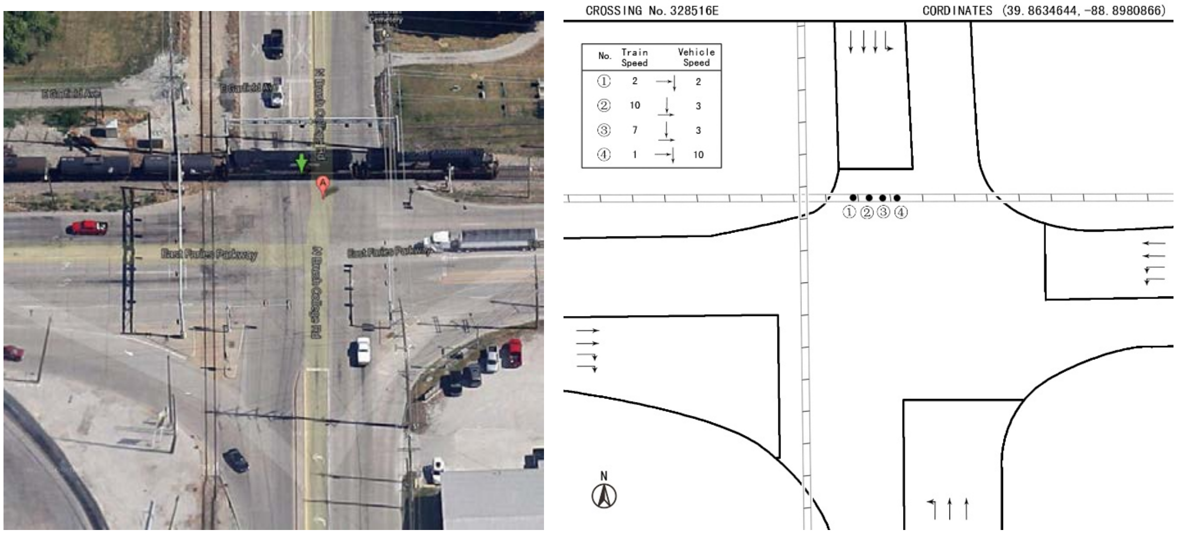


Figure 3.21. Image and Sketch of Crossing 328516E

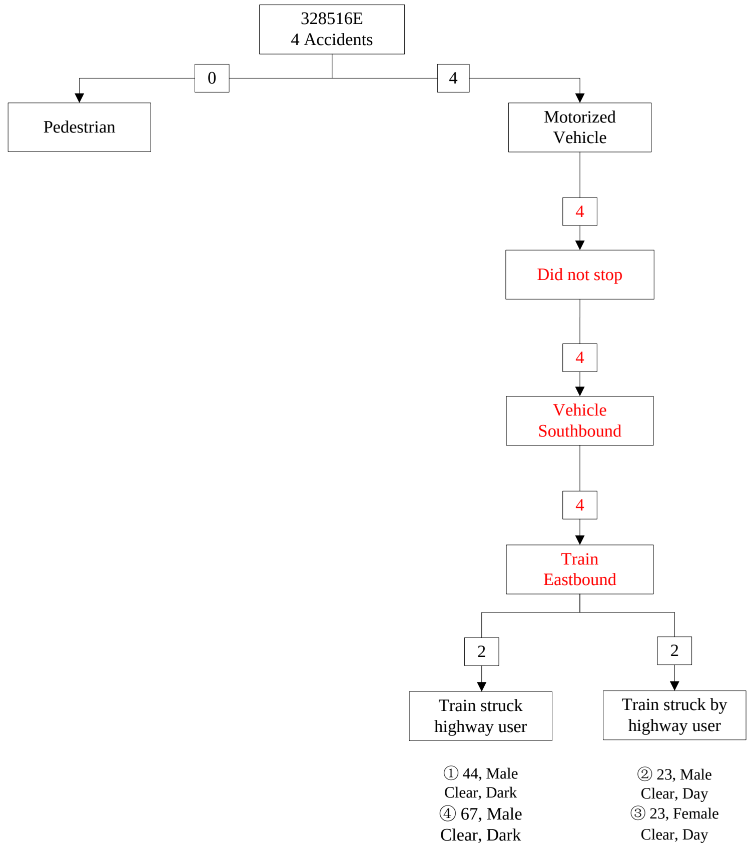


Figure 3.22. Tree diagram for accidents at Crossing 328516E

2. Crossing 608310D is located in Morgan Park, IL, just south of Chicago, near the intersection of 115th St and Paulina St. The roadway has an AADT of 6200 and an average of 66 daily trains. This crossing is equipped with gates and it is near a signalized intersection. In all four accidents, vehicles went around or through the gate and vehicles and trains were traveling in the same direction. In three accidents, vehicles and trains were traveling westbound, and in the remaining accident both train and vehicle were traveling eastbound, as it can be seen in Figures 3.23 and 3.24. Visibility can be an issue at this location, particularly for westbound vehicles, given the structure on the right hand side of approaching vehicles. Westbound trains will not be visible to the drivers until they are very close to the crossing. It is also noted that train speeds were high, further increasing the difficulty to see trains before arriving to the crossing, and in all cases the train struck the vehicles, making this scenario a likely one.

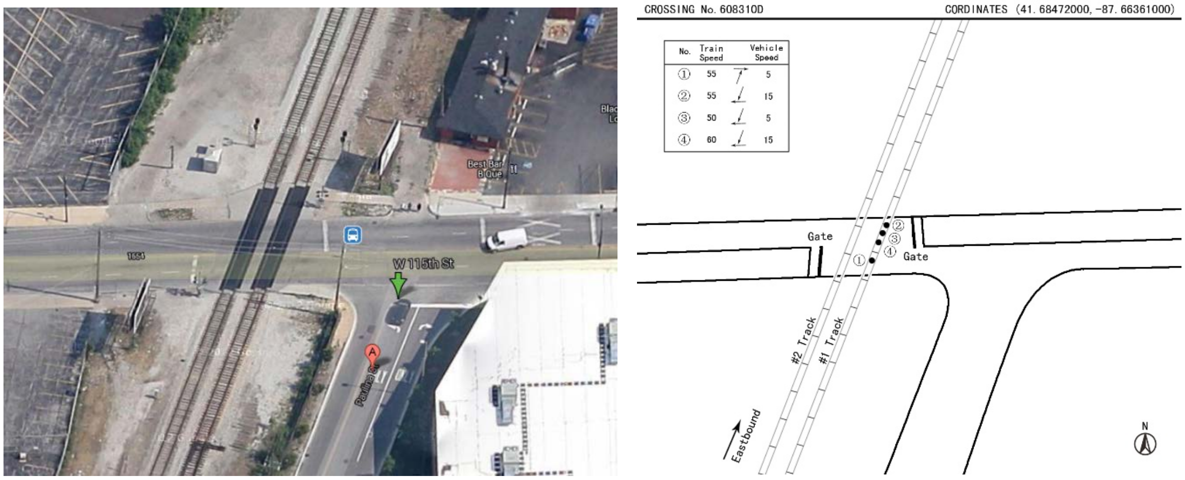


Figure 3.23. Image and Sketch of Crossing 608310D

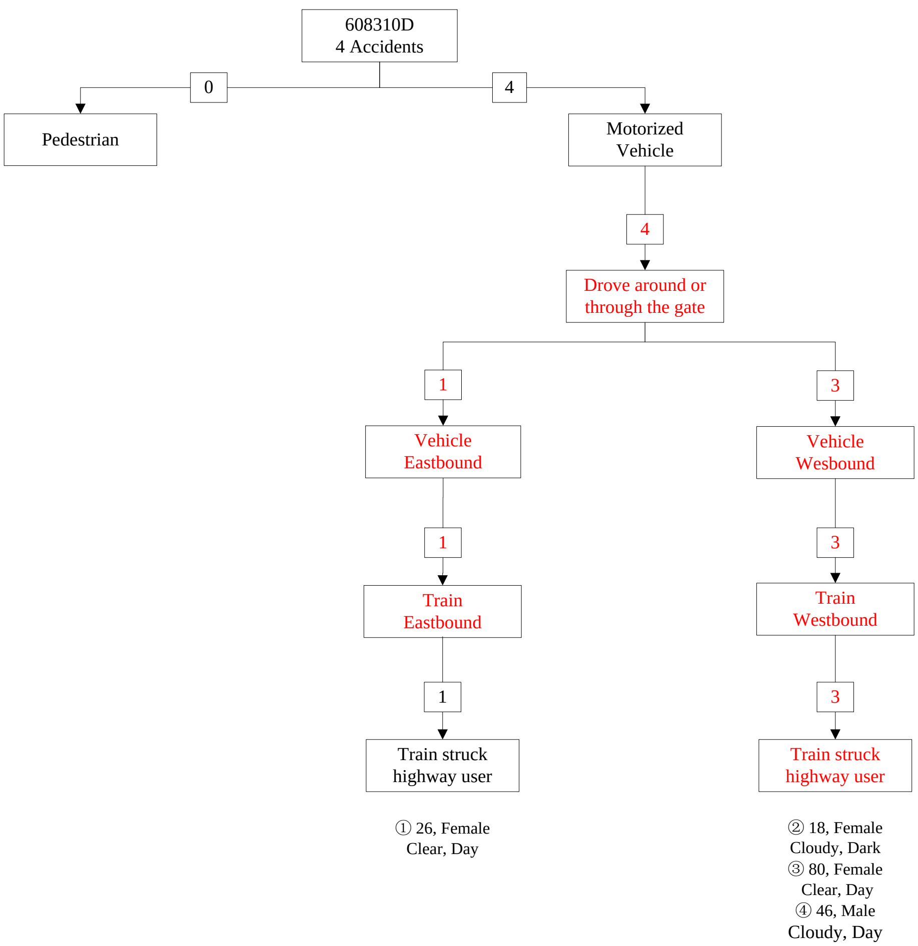


Figure 3.24. Tree diagram for accidents at Crossing 608310D

3.5. Applications of Micro Analysis

The implementation of the proposed micro analysis for high accident locations was described and exemplified in this section above. Trends can be easily visualized in the sketch and the tree diagram, helping the analyst in determining potential contributing factors in the accidents in ways that a macro model would not be able to do so. However, results from the micro analysis can be put to further use by incorporating such results into a macro model and the risk assessment for the crossings.

The addition of results from the micro analysis into a macro model can be thought for a corridor analysis or an analysis of a region, where variables in the most significant trends from the micro analysis can be added to the macro model to improve predictions. Also, a different method could add correction factors developed based on the micro analysis to adjust the predictions of an already existing macro model.

In addition, the micro analysis itself is a tool that can be used by diagnostic teams prior to field visits, enhancing their preparation and knowledge, potentially improving results of such visits.

4. Conclusions and Future Work

This report presented the exploration of methods to analyze accidents at grade crossing at both macro and micro scales. The macro models introduced here showed potential to improve the current state-of-practice using the US DOT accident prediction formula, opening a window for further study and the development of final models for prediction. The zero-inflated negative binomial model had the best fit to the data and its predictions were more accurate than those from the US DOT formula. In an example using data from Illinois, the overall accident frequency predicted by the ZINB model ranged between 1 and 1.6 times the observed frequency, whereas the US DOT formula found values between 1.8 and 3.6 times the observed accidents.

In terms of the micro analysis, a proposed methodology aiming at the analysis of single crossings with high accident frequency was presented. The micro analysis showed that it may be useful to identify trends and contributing factors not considered in macro models, providing information that can be incorporated and used for macro analysis and for diagnostic teams prior to site visits. The procedure to conduct the micro analysis is simple and it uses information from individual accidents. The micro analysis is still in development and additional features to improve it include a dynamic tree diagram, a probabilistic analysis of the accident frequency, and the extension of the methodology to corridor and regional analysis.

References

1. American Association of State Highway and Transportation Officials. Highway Safety Manual. Washington D.C., 2010.

2. Anderson, T.K. Kernel Density Estimation and K-means Clustering to Profile Road Accident Hotspots. Accident Analysis and Prevention. No. 41, 2009, pp. 359-364.

3. Federal Railroad Administration. Accident Prediction and Resource Allocation Procedure Normalizing Constants. 2010.

4. Leibowitz, H. W. Grade Crossing Accidents and Human Factors Engineering: How a Discipline Combining Technology and Behavioral Science can Help Reduce Traffic Fatalities. American Scientist, Vol. 73, No. 6, 1985, 558-562.

5. Mengert, P. Rail-Highway Crossing Hazard Prediction – Research Results. U.S. Department of Transportation. Report No. FRA-RRS-30-02. Washington D.C. 1980.

6. Miranda-Moreno, L. F., Fu, L., Saccomanno, F.F., Labbe, A. Alternative Risk Models for Ranking Locations for Safety Improvements. Transportation Research Record, No. 1908, Washington, D.C., 2005, pp. 1-8.

7. Persaud, B.N., Retting, R.A., Garder, P.E., Lord, D. Safety Effects of Roundabout Conversions in the U.S. – Empirical Bayes Observational Before-after Study. Transportation Research Record 1751. TRB, National Research Council, Washington, D.C. 2001, pp.1-8.

8. Saccomanno, F.F., Fu, L., Miranda-Moreno, L. F. Risk-Based Model for Identifying Highway-Rail Grade Crossing Blackspots. Transportation Research Record, No. 1862, Washington, D.C., 2004, pp. 127-135.

9. U.S. DOT and Federal Highway Administration. Railroad-highway Grade Crossing Handbook. Washington D.C., 2007

10. Washington, S., Oh, J. Bayesian Methodology Incorporating Expert Judgment for Ranking Countermeasure Effectiveness under Uncertainty: Example Applied to at Grade Railroad Crossings in Korea. Accident Analysis and Prevention. No. 38, 2006, pp. 234-247.