U.S. Department of
Transportation

**Federal Railroad
Administration**

Railroad Right-of-Way Incident Analysis Research

Office of Railroad
Policy and Development
Washington, DC 20590

Safety of Highway Railroad Grade Crossings

DOT/FRA/ORD-11/09

Final Report
April 2011

REPORT DOCUMENTATION PAGE			*Form Approved OMB No. 0704-0188*

1. AGENCY USE ONLY (Leave blank)	2. REPORT DATE	3. REPORT TYPE AND DATES COVERED
	April 2011	Technical Report

4. TITLE AND SUBTITLE Railroad Right-of-Way Incident Analysis Research	5. FUNDING NUMBERS RR97-EG136 RR97A1-EG227 RR97A1-FG227
6. AUTHOR(S) Mina Chaudhary, Adrian Hellman, and Tashi Ngamdung	
7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES) U.S. Department of Transportation Research and Innovative Technology Administration John A. Volpe National Transportation Systems Center Cambridge, MA 02142	8.PERFORMING ORGANIZATION REPORT NUMBER DOT-VNTSC-FRA-09-11
9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES) U.S. Department of Transportation Federal Railroad Administration Office of Railroad Policy and Development 1200 New Jersey Avenue, SE Washington, DC 20590	10. SPONSORING/MONITORING AGENCY REPORT NUMBER DOT/FRA/ORD-11/09

11. SUPPLEMENTARY NOTES
Safety of Highway-Railroad Grade Crossings series
Program Manager: Leonard W. Allen III

12a. DISTRIBUTION/AVAILABILITY STATEMENT	12b. DISTRIBUTION CODE
This document is available to the public online through the FRA Web site at http://www.fra.dot.gov.	

13. ABSTRACT (Maximum 200 words)
Locations of railroad right-of-way incidents in this research were identified as hotspots. These can be defined as highway-rail grade crossings or locations along the railroad right-of-way where collision or trespassing risk is unacceptably high and intervention is justified because the potential safety benefits exceed the cost of intervention. This project categorizes the hotspots as grade crossing and trespass incident hotspots. Mathematical models and theories are researched to see which ones may be used in identifying the hotspots. For the analysis of grade crossing incident hotspots, the Transport Canada model is modified to accommodate U.S. data and is applied to a sample of grade crossing incidents from 2003 to 2007 in the *San Joaquin* corridor in California. In analyzing trespass incident hotspots, the theory of cluster analysis, a type of spatial analysis, was researched. It appears that cluster analysis, used in conjunction with a geographic information system platform, would be a beneficial way of analyzing and predicting trespass hotspots.

14. SUBJECT TERMS Hotspots, grade crossing, trespass, accident prediction, severity prediction, risk, cluster analysis, geographic information system platform	15. NUMBER OF PAGES 46
	16. PRICE CODE

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

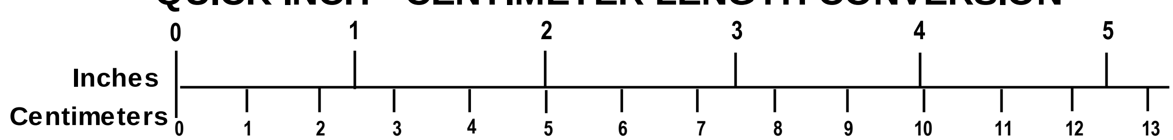
NSN 7540-01-280-5500

Standard Form 298 (Rev. 2-89)
Prescribed by ANSI Std. 239-18
298-102

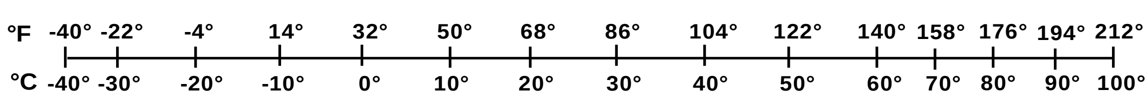
METRIC/ENGLISH CONVERSION FACTORS

ENGLISH TO METRIC			METRIC TO ENGLISH		
LENGTH (APPROXIMATE)			**LENGTH** (APPROXIMATE)		
1 inch (in)	=	2.5 centimeters (cm)	1 millimeter (mm)	=	0.04 inch (in)
1 foot (ft)	=	30 centimeters (cm)	1 centimeter (cm)	=	0.4 inch (in)
1 yard (yd)	=	0.9 meter (m)	1 meter (m)	=	3.3 feet (ft)
1 mile (mi)	=	1.6 kilometers (km)	1 meter (m)	=	1.1 yards (yd)
			1 kilometer (km)	=	0.6 mile (mi)
AREA (APPROXIMATE)			**AREA** (APPROXIMATE)		
1 square inch (sq in, in^2)	=	6.5 square centimeters (cm^2)	1 square centimeter (cm^2)	=	0.16 square inch (sq in, in^2)
1 square foot (sq ft, ft^2)	=	0.09 square meter (m^2)	1 square meter (m^2)	=	1.2 square yards (sq yd, yd^2)
1 square yard (sq yd, yd^2)	=	0.8 square meter (m^2)	1 square kilometer (km^2)	=	0.4 square mile (sq mi, mi^2)
1 square mile (sq mi, mi^2)	=	2.6 square kilometers (km^2)	10,000 square meters (m^2)	=	1 hectare (ha) = 2.5 acres
1 acre = 0.4 hectare (he)	=	4,000 square meters (m^2)			
MASS - WEIGHT (APPROXIMATE)			**MASS - WEIGHT** (APPROXIMATE)		
1 ounce (oz)	=	28 grams (gm)	1 gram (gm)	=	0.036 ounce (oz)
1 pound (lb)	=	0.45 kilogram (kg)	1 kilogram (kg)	=	2.2 pounds (lb)
1 short ton = 2,000 pounds (lb)	=	0.9 tonne (t)	1 tonne (t)	=	1,000 kilograms (kg)
				=	1.1 short tons
VOLUME (APPROXIMATE)			**VOLUME** (APPROXIMATE)		
1 teaspoon (tsp)	=	5 milliliters (ml)	1 milliliter (ml)	=	0.03 fluid ounce (fl oz)
1 tablespoon (tbsp)	=	15 milliliters (ml)	1 liter (l)	=	2.1 pints (pt)
1 fluid ounce (fl oz)	=	30 milliliters (ml)	1 liter (l)	=	1.06 quarts (qt)
1 cup (c)	=	0.24 liter (l)	1 liter (l)	=	0.26 gallon (gal)
1 pint (pt)	=	0.47 liter (l)			
1 quart (qt)	=	0.96 liter (l)			
1 gallon (gal)	=	3.8 liters (l)			
1 cubic foot (cu ft, ft^3)	=	0.03 cubic meter (m^3)	1 cubic meter (m^3)	=	36 cubic feet (cu ft, ft^3)
1 cubic yard (cu yd, yd^3)	=	0.76 cubic meter (m^3)	1 cubic meter (m^3)	=	1.3 cubic yards (cu yd, yd^3)
TEMPERATURE (EXACT)			**TEMPERATURE** (EXACT)		
[(x-32)(5/9)] □F	=	y □C	[(9/5) y + 32] □C	=	x □F

QUICK INCH - CENTIMETER LENGTH CONVERSION



QUICK FAHRENHEIT - CELSIUS TEMPERATURE CONVERSION



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

Updated 6/17/98

Acknowledgments

The U.S. Department of Transportation (USDOT) Federal Railroad Administration (FRA) Office of Research and Development sponsored the work leading to this report. The authors would like to thank Sam Alibrahim P.E., Chief of the Signals, Train Control and Communications Division, FRA; Leonard Allen, Program Manager, Signals, Train Control and Communications Division, FRA; and Gary Carr, Chief of the Track Research Division, FRA for their guidance and direction in developing this report.

The authors also wish to thank Jim Hernandez from the California Public Utilities Commission for his cooperation and the contribution of data that were used for this project. Appreciation is also extended to Mike Grizkewitsch of the FRA Office of Safety and Raquel Wright of the FRA Office of Policy for their time, support, ideas, and exchange of data.

Anya A. Carroll, Domain Expert of Intermodal Surface Transportation Systems, and Marco daSilva, Highway-Rail Grade Crossing Safety and Trespass Prevention Research Program Manager, John A. Volpe National Transportation Systems Center, provided overall direction of the technical information and the report preparation, and also provided invaluable guidance and leadership.

Contents

Illustrations

Tables

Executive Summary

In the past 30 years, especially between 1993 and 2003, great strides have been made in improving safety at highway-rail grade crossings. Collisions at grade crossings declined 41 percent and fatalities declined 48 percent. However, in the past 5 years, they have been falling at a lower rate. The number of trespass fatalities is now higher than the number of grade crossing fatalities. In other words, safety issues associated with both trespass and grade crossing incidents still need to be addressed.

Identifying railroad infrastructure hotspots could determine which crossings in a corridor or community require further examination. Hotspots can be defined as highway-rail grade crossings or other locations along the railroad right-of-way where collision or trespassing risk is unacceptably high and intervention is justified because the potential safety benefits exceed the cost of intervention. Hotspots can be separated into grade crossing incident hotspots and trespass incident hotspots.

Two models have been used in this research; the Transport Canada (TC) and U.S. Department of Transportation (USDOT) accident prediction model were chosen for analysis. TC model was used successfully in Canada and is a best-fit data model that uses negative binomial regression. The USDOT accident prediction model was used on the same data sample and the results of the two models are compared. The results show that the USDOT model predicted the number of accidents closest to the observed number. The TC model overestimated the number of accidents, which may be because of certain assumptions that were made about the data. However, for crossings that had an accident history, the TC model gave a more accurate prediction of incidents. It could theoretically be used in ranking dangerous crossings, thereby helping to determine grade crossing warning upgrades.

1. Introduction

The John A. Volpe National Transportation Systems Center (Volpe Center) provides technical support to the Federal Railroad Administration (FRA) on all aspects of grade crossing research. Significant progress has been made in the past 30 years in improving safety at highway-rail grade crossings. Collisions at grade crossings have declined 41 percent, and fatalities have declined 48 percent between 1993 and 2003. The goal is to continue the downward trend of grade crossing incidents in spite of limited funding. Another objective is addressing the issue of trespass incidents.

In recent years, there have been approximately 35 percent more trespass fatalities than grade crossing fatalities. For the years 2003 through 2006, the annual number of trespass fatalities ranged from 458 to 511, while the annual number of grade crossing fatalities ranged from 334 to 371. Annual trespass fatalities averaged approximately 500 and grade crossing fatalities averaged approximately 350. Figure 1 displays the number of trespass and grade crossing fatalities per year for the past 17 years [1]. Grade crossing fatalities have steadily decreased over the years while trespass fatalities have not. However, it can be seen that grade crossing fatalities in the past 5 years are not declining at the rate they were in previous years. Trespass and grade crossing incidents still need to be addressed.

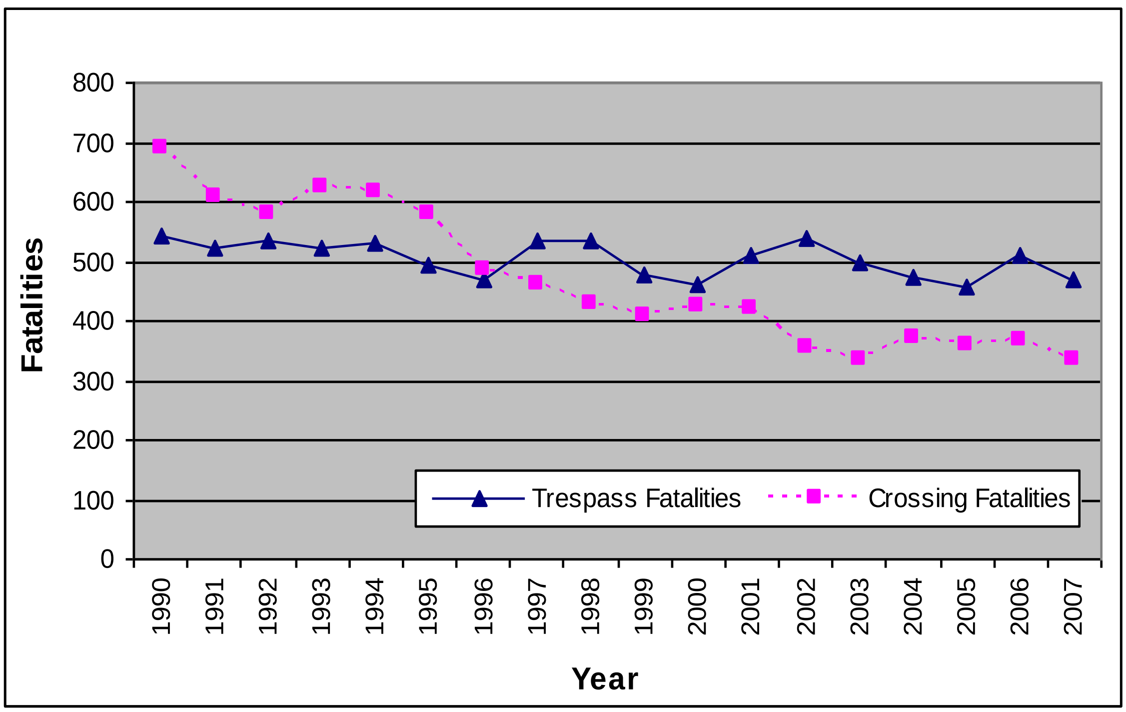


Figure 1. Railroad Trespass/Grade Crossing Fatalities (1990–2007)

While trying to improve safety, the question of how to allocate limited funds for crossing improvements is almost always an issue for local communities or states. Identifying railroad infrastructure hotspots could help determine which crossings in a corridor or community require further examination. Hotspots can be defined as highway-rail grade crossings or locations along the railroad right-of-way where collisions or trespass risk is unacceptably high and intervention is justified because the potential safety benefits exceed the cost of intervention. This research is similar to epidemiology studies of disease outbreaks in medicine. In the railroad environment, this includes the analysis of highway-rail grade crossing clusters exhibiting a high concentration of incidents.

In this document, hotspots are broken down into two categories, grade crossing and trespass incidents. The research and analysis varies for each category. For grade crossing incident hotspots, the TC model was chosen for analysis. It was used successfully in Canada and is a best-fit data model that uses negative binomial regression. It identifies hotspots and hotspot clusters based on user-defined thresholds for frequency and consequences. The data requirements necessary to use this model in the United States have been researched and will be demonstrated on a data sample from California. The USDOT accident prediction model will also be used on this data sample and the results of the two models are compared.

For trespass incident hotspots, cluster analysis was deemed a useful theory that could be used in developing a model that could predict trespass incident hotspots. Cluster analysis uses negative binomial regression and can be normalized for exposure. Currently, trespass data is being examined as raw data. It would be beneficial to be able to identify potential hotspots. While researching trespass incident hotspots, the Volpe Center began working with the CPUC to compare the quality of CPUC trespass and grade crossing incident data with the National Response Center, operated by the U.S. Coast Guard (USCG). This is being performed so the National Response Center may be used as an alternative data source. During the course of this project, the Volpe Center also worked with FRA's Office of Safety to supply trespass data from the National Response Center so it can be mapped on a Geographic Information System (GIS) platform. If mapped on a GIS platform, there is great potential for cluster analysis of trespass incidents. The data can be stratified to conduct demographic studies. For example, male trespass incidents could be examined versus female. There is a great deal of potential for future research in this area.

2. Models Researched

The idea of using the TC model for grade crossing incident hotspot research and using cluster analysis for trespass hotspot research came after a comprehensive literature review was performed. The pertinent models examined are listed below.

2.1 USDOT Accident Prediction Model

According to Hauer and Persaud, the safety of a specific grade crossing is a direct function of the expected number of accidents and the resulting severity over some unit of time [2]. In this context, Hauer and Persaud define *expected* as what would be the average, in the long run, if it was possible for all relevant conditions to remain unchanged. Also, accident frequency is not synonymous with safety, but is instead an indirect measure. In general, two categories of variables, causal factors and accident history, are the best descriptors of grade crossing safety. Again referencing Hauer, the USDOT rail-highway crossing accident prediction formula, first published in 1981, was the first accident prediction model to linearly combine both causal factors and accident history.

The first attempt in the development of the USDOT model was to use linear regression analysis. Although the results were interesting, they were less than satisfactory as compared with previous accident prediction models. The salient findings of Mengert are given below [3]:

- Variables associated with train and vehicle movements (traffic moment) total approximately 90 percent of the predictive force of the regression model.
- Linear regressions with more than eight variables produces inferior results due to issues related to co-linearity, including misdirected signs of variables.
- Linear regression techniques are not necessarily the best for producing accurate accident prediction models. The regression techniques studied did not appear to provide any benefit over pre-existing modeling techniques.

The key finding in this research was that the grade crossing accident probability curve is nonlinear, resembling the hyperbolic tangent function, $\tanh(x) = \dfrac{e^{2x}-1}{e^{2x}+1}$. This assertion formed the basis for the derivation of the USDOT accident prediction formula. To this end, Mengert's approach was to develop an accident prediction model for each warning device class (i.e. crossbucks, flashing lights, and gates with flashing lights). The first step of this process was to build a regression model using only variables associated with the traffic moment.[1] The next step was to express the model as a third degree polynomial. Finally, variables unassociated with the traffic moment, such as the number of main tracks and number of highway lanes, were integrated within the model to produce the accident prediction model or the "comprehensive" model. The most general form of the model, consisting of only traffic moment related variables, is expressed in Equation 2-1 on the next page.

[1] Traffic moment refers to the product of the average daily rail and average vehicle traffic at a highway-rail crossing.

$$a_0 + a_1 \cdot \log_{10}(T+1) + a_2 \cdot \log_{10}(C+1) + a_3 \cdot [\log_{10}(T+1)]^2$$
$$+ a_4 \cdot [\log_{10}(C+1)]^2 + a_5 \cdot \log_{10}(T+1) \cdot \log_{10}(C+1) \quad \text{(Eq. 2-1)}$$

Where

a_i = the regression coefficients
T = train movements
C = vehicle movements

After some manipulation, the more familiar form of the equation, as published by Farr, is arrived at [4].

$$a = K \cdot EI \cdot DT \cdot MS \cdot MT \cdot HP \cdot HL$$

(Eq. 2-2)

Where

a = un-normalized accident prediction (accidents/year at the crossing)
K = constant for initialization of factor values at 1.00
EI = exposure index based on product of highway and train traffic
DT = number of through trains per day during daylight hours
MS = maximum railroad timetable speed
MT = number of main tracks
HP = highway paved factor
HL = number of highway lanes

After normalizing for accident history, this is expressed as

$$B = \frac{T_o}{T_o + T}(a) + \frac{T}{T_o + T}(N/T), T_o = 1/(0.05 + a)$$

(Eq. 2-3)

Where

T_o is derived from the basic un-normalized accident prediction formula, a
N/T = the accident history of the grade crossing, expressed as ratio of the number of accidents in T years
N = the number of accidents recorded for a crossing in T years
T = the number of years, usually 5, but can be any number

To obtain the normalized value, the predicted value, B, is multiplied by the appropriate normalizing constant depending on the type of warning device used (see Table 1 below).

Table 1. Crossing Characteristics Factors

Warning Device Groups	Normalizing Constants
Passive	0.6768
Flashing Lights	0.4605
Gates	0.6039

*Source: 2007 Accident Prediction Formula Normalizing Constants
http://safetydata.fra.dot.gov

2.2 Transport Canada Model

2.2.1 Test USDOT Model with Canadian Data

The goal of the research, funded by TC and performed by the University of Waterloo, was to develop a risk-based methodology for identifying grade crossing hotspots in Canada [5]. The initial focus of this research was to test the applicability of the USDOT accident frequency and severity prediction formulas with Canadian grade crossing data. Grade crossings in the TC database were classified by warning device type (passive, lights, gates), train speed, and levels of traffic exposure. Next, the Chi-Square Goodness-of-Fit Test was used to determine if the differences between the predicted and observed collision values were statistically significant. For the 1,724 reported grade crossing collisions from 1993 to 2001, a poor goodness-of-fit was observed, with the USDOT accident prediction formula overestimating the total by 349 collisions or approximately 20 percent.

Similarly, the differences between the predicted and observed values using the USDOT severity models were found to be statistically significant. As with the accident prediction model, one of the main disadvantages of the USDOT severity prediction model is the tendency to overestimate the fatality and casualty calculations. TC found grade crossing collisions involving fatalities and injuries to be a small subset of the total crossing collisions. As such, they chose to develop a combined fatality and injury severity model with an overall collision severity score based on the number of fatalities, injuries, and property damage cost.

2.2.2 Grade Crossing Accident Prediction Model

Historically, linear and nonlinear regression analysis techniques have been the preferred means for modeling the probability of highway collisions and highway-rail grade crossing collisions. However, in the past 20 years, research has focused on the exponential family of discrete-time stochastic distributions. These distributions are used to model naturally occurring random processes with the following characteristics:

- There are *n* independent repeated trials
- The probability of success, *p*, is constant from trial to trial
- Each trial results in an outcome that may be classified as a success or failure
- Each outcome, is described by a random variable

Grade crossing accidents, which are essentially random events, can be described by the exponential family of distributions. This is especially true for discrete time intervals, such as 1 year or 5 years, in which the characteristics of a grade crossing are relatively constant. Two discrete-time distributions, the Poisson and negative binomial, are employed specifically to describe the distribution of random variables, such as grade crossing collisions, over a given time interval. As such, Poisson and negative binomial regression analyses have been employed extensively for accident frequency modeling of grade crossings.

The Poisson probability function is expressed as:

$$P(n;\lambda t) = \frac{e^{(-\lambda it)}\lambda t^{n}}{n!} \qquad \text{(Eq. 2-4)}$$

Where

$P(n; \lambda t)$ = the probability of n accidents occurring at a grade crossing per unit time t
t = the time interval being studied
λ = average number of collisions per unit time
n = the number of collisions in the time interval t

Now, the expected number of collisions, in time interval t, can be expressed as

$$E(n) = \lambda = e^{\left[\sum_i \beta_i X_i\right]} \qquad \text{(Eq. 2-5)}$$

Where

$E(n)$ = the expected number of collisions at a grade crossing per unit time t
β_i = a vector of unknown regression coefficients that can be estimated by standard maximum likelihood methods.
X_i = grade crossing geometric, spatial/land use, and other relevant attributes that impact accident frequency

A key requirement of the Poisson probability distribution is that the mean and variance are equal in value. This means that the closer they are to each other, the better the goodness-of-fit between the regression model and the observed data. However, under real world conditions, the variance is frequently larger than the mean. This results in a condition known as overdispersion, in which the model coefficients are not accurate.

The negative binomial probability distribution, which in its limiting form, converges to the Poisson distribution, can be used to relax the constraint of the mean and variance being equal. Under this condition, Equation 2-5 can be rewritten as

$$E(n) = \lambda = e^{\left[\sum_i \beta_i X_i + \varepsilon_i\right]} \qquad \text{(Eq. 2-6)}$$

Where e^{ε_i} is called a Gamma-distributed error term (Lee, Nam, and Park, 2005) [6].

The approach taken by TC was to construct a model using Poisson regression analysis, a technique that is widely employed in collision prediction modeling. For the three types of grade crossing warning devices—(i) signs, (ii) signs and flashing lights, and (iii) signs, flashing lights, and gates—a separate regression analysis model was developed. Regression analysis was used on Canadian grade crossing collision data from 1993 to 1996 to develop the models, and data from 1997 to 2000 was used to validate them. Of the nine factors in Table 2 that were tested in the model, a total of five were found to be statistically significant, depending on the model. Many of these factors are reflected in the USDOT accident prediction formula as well.

The expressions for each warning device type are shown below.

Table 2. Factors Tested in Accident Prediction Model Regression Analysis

Warning device type	Road surface width
Train vehicle speed	Traffic exposure (average annual daily traffic (AADT) × number of daily trains)
Road vehicle speed	Road class (arterial/other)
Number of tracks	Road pavement condition (paved/unpaved)
Track angle	

$$E\,(m_S) = e^{[-5.66 + 0.0128 * \text{TSPD} + 0.3791 * \ln(\text{EXPO})]} \qquad \text{(Eq. 2-7)}$$

$$E\,(m_F) = e^{[-9.1620 + 0.0112 * \text{TSPD} + 0.0151 * \text{SW} + 0.6103 * \ln(\text{EXPO})]} \qquad \text{(Eq. 2-8)}$$

$$E\,(m_G) = e^{[-7.2304 + 0.0118 * \text{RSPD} + 0.1912 * \text{TN} + 0.3526 * \ln(\text{EXPO})]} \qquad \text{(Eq. 2-9)}$$

Where

TSPD = Maximum train speed in miles per hour (mph)
RSPD = Road speed in kilometers per hour (km/h)
EXPO = Crossing exposure factor, AADT × Number of daily trains
SW = Road surface width in feet
TN = Number of railroad tracks in both directions

In Equation 2-7, the expression for passive crossings, only train speed and exposure were found to be statistically significant. The expression for flashing lights in Equation 2-8 includes an extra statistically significant variable, road surface width. As in the USDOT model, train speed is not a statistically significant variable at gated grade crossings but road speed is, as shown by Equation 2-9.

As described in the TC report, the Poisson model is constrained such that the mean number of collisions is equal to the variance. However, frequently the collision frequency variance exceeds the mean, indicating a lack of explanation in the underlying Poisson model. This is known as Poisson overdispersion and could result in significant prediction error. The negative binomial regression technique reduces the overdispersion effect by relaxing the Poisson model assumption of the mean equaling the variance. Empirical Bayesian analysis, which calibrates the Poisson model with historical collision data, is another approach to overcoming the overdispersion phenomena and has been used in hotspot modeling.

In light of the overdispersion concern, TC tested the three Poisson models using the Scaled Deviance and Pearson χ^2 tests, where values close to 1.0 indicate overdispersion is low. Although a small amount of overdispersion was measured for each model, it was not considered significant. Moreover, the Chi-Square Goodness-of-Fit Test showed good correspondence between the modeled and the measured values. The Empirical Bayesian technique was used to calibrate the Poisson expressions with 4 years of crossing collision data (1997–2000). However, since grade crossing collisions are low probability events, accident history calibration did not provide appreciable improvement over the stand-alone Poisson regression modeling for the 4-year time period. As such, TC selected the stand-alone Poisson model for grade crossing collision.

2.2.3 Grade Crossing Severity Prediction

As noted previously, TC developed an overall consequence score that is a weighted sum of severity resulting from fatalities, injuries, and property damage, as shown in Equation 2-3. This technique not only facilitates grade crossing hotspot identification but also (i) encompasses all grade crossing collision data, not just collisions for grade crossings with fatalities and injuries, and (ii) addresses co-linearity between fatalities and injuries, which, if not dealt with, results in counterintuitive model variables. It works for all crossings because once a collision has occurred; the consequences are independent of the warning device. TC defined severity in terms of the consequence score below.

$$CS_i = 44 \times NF_i + 1 \times NI_i + 1 \times PD_i \qquad \text{(Eq. 2-10)}$$

Where

NF_i = Number of fatalities
NI_i = Number of injuries
PD_i = Property damage

The first attempt at fitting the consequence score involved Poisson regression analysis on Canadian grade crossing accident data for 1997–2001. The 826 reported accidents were randomly divided into two sets of 413, one for regression modeling and the other for model validation. Of the eight factors in Table 3 that were tested in the model, the four independent variables in Equation 2-11 were found to be statistically significant. However, the Pearson χ^2 and scaled deviance values showed significant overdispersion, an indication of significant prediction error. When TC used negative binomial regression analysis, considerable improvement in overdispersion was observed, as indicated by the Pearson χ^2 and scaled deviance values being close to 1.0. The resulting model is shown below.

Table 3. Factors Tested in Collision Consequence Model Regression Analysis

Train speed	Road surface width
Road speed	AADT
Number of tracks	Number of trains per day
Track Angle	Number of road vehicle occupants

$$E\,(C_q/C) = e^{(0.3426 * PI - 0.2262 * TN + 0.0069 * TA + 0.0250 * TSPD)} \qquad \text{(Eq. 2-11)}$$

Where

$E\,(C_q/C)$ = Expected consequence/collision
PI = Number of persons involved

TN	= Number of railway tracks both directions
TA	= Track angle
TSPD	= Maximum train speed in mph

2.2.4 Hotspot Identification

The final piece of the TC research constituted development of a system to differentiate between hotspot and non-hotspot grade crossings with respect to a specified threshold value. The key was to construct a combined risk index that inherently encompassed both the expected frequency and consequence of a collision. This metric, expressed in terms of risk per year, results from the product of the expected frequency and the collision consequence score in Equation 2-10. By defining the risk index in this manner, crossings that are in one of the following two categories, 1) low expected collision frequency and high collision consequence score, and 2) high expected collision frequency and low collision consequence score, could potentially yield a low combined risk index. Since resources for risk mitigation treatments are finite, this approach provides a mechanism to "filter out" those grade crossings with a low risk index and focus in on the ones that have a high risk index.

Intuitively, the threshold value is a user-defined number and is implicitly dependent on a variety of subjective factors that may be specific to the grade crossings and treatments being evaluated. However, the guiding principle is to determine the level of investment that produces the largest decrease in risk for the smallest capital expenditure. One way to find this balance is by performing a cost-benefit study of the proposed grade crossing treatments. An alternative approach involves defining a low threshold value such that the majority of crossings being evaluated satisfy the hotspot criteria. If the hotspot crossings are considered collectively, it is possible to calculate a total risk index by summing the risk index of all the individual crossings. As such, the benefit associated with a specific treatment can be ascertained from the resulting decrease in the total risk index.

2.3 University of North Carolina Pedestrian Crash Risk Model

Schneider developed an approach to pedestrian accident risk using Poisson and negative binomial regression analysis [7]. The general problem posed by this research, as well as the methodology and results show significant parallels to the FRA grade crossing and railroad right-of-way trespass hotspot program. As such, this work has great promise and presents the opportunity to leverage the results from another transportation mode as a foundation.

In his research, Schneider divided the University of North Carolina (UNC) at Chapel Hill campus roadway network into 38 intersections and 56 segments. In this approach, intersections are the network nodes and the segments are the network links equivalent to lengths of roadway between intersections. Each of the 94 total intersections or segments was then classified by pedestrian crash-risk, exposure, roadway, and land use attributes. For each segment/intersection of the UNC campus roadway network, Schneider used traffic and pedestrian volume maps, field observations, and geographic information systems (GIS) measurements to populate a pedestrian risk database. This database consisted of three exposure, five roadway, and seven land use attributes. Additionally, police reports of car accidents on the UNC campus for the 5-year period from October 1994 to September 1999 were used to develop a representation of pedestrian accident history. During that timeframe, 127 pedestrian accidents were recorded in Chapel Hill, of which 57 occurred on the UNC campus, including one fatality in 1999.

Concurrently, Schneider administered a survey to a randomly selected group of UNC students, faculty, and employees, including pedestrians and drivers, to gather data on the locations of perceived pedestrian and driver risk. They were also asked if they perceived a higher risk for being involved in accidents at night versus during the day and whether they had been involved in a vehicle-pedestrian or pedestrian-vehicle "near miss" during the past month. In addition to the survey, the respondents were requested to fill out two maps. For the first map, they were asked to identify the three locations that they perceived to have the highest risk of pedestrian crashes during daylight. Also, if they perceived that the risk was different at night, they were requested to identify them on the second map.

Using the data from the police recorded accidents and perceived risky locations, Schneider developed a series of pedestrian risk models based on Poisson and negative binomial regression analysis. The results were paired by observed risk and perceived risk and are shown below.

$$E(m_{po}) = e^{[-13.5 + 0.625\beta_1 + 0.672\beta_2 + 0.381\beta_3 - 0.935\beta_4 + 0.0940\beta_5 - 0.159\beta_6 + 0.0810\beta_7 - 0.000841\beta_8 + 0.000996\beta_9 + 0.000783\beta_{10}]} \qquad \text{(Eq. 2-12)}$$

$$E(m_{nbo}) = e^{[-13.5 + 0.625\beta_1 + 0.667\beta_2 + 0.383\beta_3 - 0.932\beta_4 + 0.101\beta_5 - 0.158\beta_6 + 0.0806\beta_7 - 0.000822\beta_8 + 0.000973\beta_9 + 0.000764\beta_{10}]} \qquad \text{(Eq. 2-13)}$$

$$E(m_{pp}) = e^{[-19.9 + 1.05\beta_1 + 1.25\beta_2 + 1.05\beta_3 - 0.797\beta_4 - 1.01\beta_5 + 0.00750\beta_6 + 0.090\beta_7 + 0.000218\beta_8 - 0.000225\beta_9 - 0.000636\beta_{10}]} \qquad \text{(Eq. 2-14)}$$

$$E(m_{nbp}) = e^{[-14.4 + 0.867\beta_1 + 0.871\beta_2 + 0.796\beta_3 - 0.246\beta_4 - 0.697\beta_5 + 0.0159\beta_6 + 0.0643\beta_7 + 0.000316\beta_8 - 0.000348\beta_9 - 0.000644\beta_{10}]} \qquad \text{(Eq. 2-15)}$$

Where

$E(m_{po})$ = the Poisson regression model for police recorded accidents
$E(m_{nbo})$ = the negative binomial regression model for police recorded accidents
$E(m_{pp})$ = the Poisson regression model for perceived accident risk
$E(m_{nbp})$ = the negative binomial regression model for perceived accident risk

β_1 = natural logarithm of a segment/intersection in feet

β_2 = natural logarithm of the estimated daily pedestrian volume

β_3 = natural logarithm of the estimated daily vehicle volume

β_4 = 0 if the location is a segment and 1 if the location is within 50 feet of an intersection

β_5 = 0 if there is no sidewalk, 1 if there is a sidewalk on one side of the street, or 2 if there are complete sidewalks on both sides of the street

β_6 = the number of bus stops per 1,000 linear feet

β_7 = the number of marked crosswalks per 1,000 linear feet

β_8 = the distance the nearest campus library

β_9 = the distance to the nearest of over 30 academic buildings

β_{10} = the distance to the nearest sports stadium

For the observed data, the Poisson model (Eq. 2-12) performed slightly better than the negative binomial regression model for two reasons. First, the significance levels of the parameter coefficient estimates were better. Second, there was no statistical difference between the mean and variance of the police reported incident risk model, indicating a low level of dispersion and a better fit for the Poisson model. Conversely, for perceived risk, the binomial regression model of Equation 2-15 returned a better goodness-of-fit than the Poisson model, which exhibited a statistically significant level of dispersion.

2.4 Summary

Two grade crossing accident risk prediction models were described in this section. In addition, a pedestrian crash risk model for highway-highway intersections and highway segments was presented. The first of the grade crossing accident risk models, the USDOT accident prediction formula, is mathematically rigorous and widely accepted within the grade crossing community. However, the applicability of this model for hotspot identification has not yet been validated. The TC risk prediction model, which is the foundation for this report, employs the more modern techniques of Poisson and negative binomial regression analysis. Although it has not been in service for as many years as the USDOT accident model, it has undergone extensive testing using Canadian grade crossing data. The methodology used by TC to identify grade crossing hotspots should apply to the U.S. grade crossing inventory as well and would be the subject of a future research endeavor.

The approach employed in the UNC research effort was to describe the entire campus roadway network in terms of intersections and road segments. The network accident frequency is then expressed using Poisson and binomial distributions. This methodology provides a framework with the potential for duplication in the modeling of grade crossing and trespass hotspot locations. The parallels between the modeling of roadway network and railroad right-of-way hotspots should be explored, so as to leverage the benefits of previously documented research.

3. Grade Crossing Incident Hotspots

FRA has had a memorandum of understanding with TC in the past. The TC model has been quite successful in Canada and the question was how it would work in the U.S. In comparison to the USDOT accident prediction model, the TC model appears to be simpler and easier to modify. To compare the two models, a sample of grade crossing incident data from U.S. data also needed to be in the same format as the TC model. This data formatting is described below.

3.1 Data Fields and Sources for TC Model

This subsection focuses on the data field requirements that are needed for the TC Model. Like the USDOT Model, the TC Model consists of three separate regression expressions, one for each of the three warning devices (Type S for Passive, F for Flashing Lights, and G for Gates). These three regression expressions are detailed below.

3.1.1 Passive Warning Devices

The accident prediction formula for crossings with passive warning devices is:

$$E\,(m_S) = e^{[-5.66 + 0.0128 * \mathrm{TSPD} + 0.3791 * \ln(\mathrm{EXPO})]} \qquad \text{(Eq. 3-1)}$$

Where: TSPD = maximum train speed (mph)
EXPO = cross product of AADT and number of trains daily

The three data elements required to run the above expression are easily accessible from the FRA National Highway-Rail Crossing Inventory (Crossing Inventory) database. The maximum train speed (TSPD) is located in column AI, titled "MAXTTSPD," and is under the correct variable miles per hour. For EXPO, the AADT is located in column CD, titled "AADT," and the year the AADT collected is available in column DO, titled "AADTYEAR." The number of daily trains is located in column EE, titled "TOTALTRN."

3.1.2 Flashing Lights Warning Devices

The accident prediction formula for crossings with flashing light warning devices is:

$$E\,(m_F) = e^{[-9.1620 + 0.0112 * \mathrm{TSPD} + 0.0151 * \mathrm{SW} + 0.6103 * \ln(\mathrm{EXPO})]} \qquad \text{(Eq. 3-2)}$$

Where: TSPD = maximum train speed (mph)
SW = surface width (feet (ft))
EXPO = cross product of AADT and number of trains daily

The locations of the TSPD and EXPO variables are mentioned in the above paragraph. The surface width (SW) is not available in the FRA Crossing Inventory database, but the database provides the number of traffic lanes crossing the railroad tracks. The number of traffic lanes crossing the railroad tracks is located in column BY, titled "TRAFICLN." The assumption was made that the average width of a traffic lane would be 12 feet; therefore, the number of traffic lanes crossing the railroad tracks was multiplied by 12 feet to get the surface width of each crossing.

3.1.3 Gated Warning Devices

The accident prediction formula for crossings with gated warning devices is:

$$E\,(m_G) = e^{[-7.2304 + 0.0118 * \text{RSPD} + 0.1912 * \text{TN} + 0.3526 * \ln(\text{EXPO})]} \qquad \text{(Eq. 3-3)}$$

Where: RSPD = road speed (kilometers per hour (km/h))
TN = number of railway tracks (both directions)
EXPO = cross product of AADT and number of trains daily

The locations of the EXPO variables are mentioned in the above paragraphs. The number of railroad tracks (TN) is not easily accessible from the FRA Crossing Inventory. They are organized into main track, other track, and SEPIND. SEPIND is the number of tracks another railroad operates. To obtain the number of railroad tracks (TN), the sum of the following fields is taken: number of main track, other track and SEPIND. The RSPD variable is located in column EQ, titled "HWYSPEED." The value is in miles per hour, but the accident prediction formula for crossings with gated warning devices in the TC model calls for this value in km/h. To convert to kilometers per hour, the mph value is multiplied by 1.609344. (If the RSPD of the crossings has a value of 0 mph, an assumption of 25 mph or 40.23 km/h was made.)

3.2 Application of Models on a Sample from California

The crossings on the *San Joaquin* high-speed rail corridor that run from Port Chicago to Bakersfield were used to test the TC accident prediction model and to determine whether it is more reliable or accurate than the USDOT accident prediction model. The *San Joaquin* corridor was chosen over other corridors (namely the three corridors that the Volpe Center has researched in the past: North Carolina sealed corridor, *Chicago-St. Louis* high-speed corridor, and *San Joaquin* corridor) because it encompasses crossings with a greater variety of warning devices present. Also, other methods were used to validate crossing inventory data along this corridor in the previous Volpe study.

The portion of the *San Joaquin* corridor used in this study consists of 229 public at-grade and 36 private at-grade crossings with the total of 265 at-grade crossings to be analyzed in this report. These crossings are broken down by warning device type in Table 4.

Table 4. Warning Device Type for Public and Private Crossings for Sample

	Gates	Flashing Lights	Passive	Total
Public Crossings	220	6	3	229
Private Crossings	1	1	34	36

3.3 Results

3.3.1 Crossing Incident History

Five-year incident data from 2003 to 2007 was compiled from the FRA RAIRS database for the 265 at-grade crossings on the *San Joaquin* corridor to validate and compare the USDOT accident prediction model and the TC accident prediction model. There were a total of 77 incidents for the 5-year period with an average of 15.4 incidents per year along the corridor. Of the 265 at-grade crossings along the corridor, only 56 crossings had at least one incident during the 5-year period. Table 5 below summarizes the incidents at those crossings by year.

Table 5. Incidents at Highway-Rail Grade Crossing for Sample, 2003–2007

	Public Crossing	Private Crossing	Total
2003	13	1	14
2004	20	1	21
2005	10	0	10
2006	15	1	16
2007	15	1	16
Total	73	4	77

3.3.2 Risk Calculation of the Models

USDOT Accident Prediction Model
The USDOT accident prediction model was applied to the 265 highway-rail grade crossings to calculate the expected number of incidents at each crossing. The expected number of accidents was generated by retrieving data fields from the FRA Crossing Inventory database. Of the 265 crossings, 34 crossings lacked required data to calculate the expected number of accidents. On average, there were 14.22 predicted accidents per year for the 231 crossings that were reviewed along the corridor.

TC Model
The appropriate expressions were applied to the 265 at-grade crossings located along the corridor to calculate the expected number of incidents. The majority of variables required to calculate the expected number of incidents was obtained from the FRA Crossing Inventory database. However, some of the variables were not easily accessible so certain assumptions were made. Section 3.1 Data Fields and Sources for TC Model above described those assumptions. Of the 265 at-grade crossings, five crossings lacked required data to calculate the expected number of accidents. For the 260 crossings that were used in the calculation, on average there were 26.07 predicted accidents per year.

3.3.3 Comparison of the Models

The USDOT accident prediction model and the TC model were compared against observed accidents to determine the reliability and accuracy of each model. Three different methods were used for this comparison. First, overall raw incident data was examined. This was the observed average number of incidents for 265 grade crossings, taken from the RAIRS database over 5 years, 2003–2007. Next, for each individual crossing, incident data for the same years, 2003–2007, was reviewed. And last, the Chi-Square Goodness-of-Fit Test was used to compare the accuracy of the models.

First, the raw accident data was examined to determine which model predicted the number of incidents closest to the observed number that is in the FRA Railroad Accident Incident Reporting System – Highway-Rail Grade Crossing (RAIRS) database for the period 2003–2007. To compare the predicted number of incidents from the USDOT model and the TC model, the same number of crossings needed to be used. For this analysis, it was 229 crossings; the original 265 crossings could not be used because of missing data that would be required to calculate the risk. The USDOT model predicted 14.22 incidents per year; the TC model predicted 23.01 incidents per year. The average number of incidents per year for the entire 265 crossings on the corridor, based on the observed data, was 15.40. However, when only the same 229 crossings used in the models were included, the average number of annual incidents decreased to 14.80. The results are summarized in Table 6 below.

Based on this comparison, the TC Model was found to overestimate the average number of incidents per year along the *San Joaquin* corridor. Some of the assumptions that were made in the calculations for the TC model (see Section 3.1 Data Fields and Sources for TC Model) may have led to this overestimation. Also, the observed data values showed that the 36 crossings with the missing data fields contained a small amount of risk, as indicated by the smaller risk value for the 229 crossing data set. Other than bringing the USDOT model into closer alignment with the observed data, this did not significantly impact the inferences drawn from this analysis.

Table 6. Risk Results Summary

Model	Number of Crossings Analyzed	Average Annual Incidents
Observed	265	15.40*
Observed	229	14.80*
Expected (TC Model)	229	23.01
Expected (USDOT Model)	229	14.22

*Average of actual incidents over the 5 years

Next, individual crossings were examined to determine how well the models predicted incidents for each crossing compared with the observed number of incidents. The top 50 percent of observed incidents at each crossing was plotted with expected number of incidents from each model. As can be seen from Figure 2 on the next page, the predicted number of incidents from the TC model is a closer fit to the observed number of incidents than that of the USDOT model. For crossings that had an incident history, the TC model gave a more accurate prediction of incidents. However, both models employ historical national data to predict accidents at specific crossings. Although they are accurate on the national level, resolution may be influenced by local variations when examining a relatively small number of grade crossings or a specific corridor.

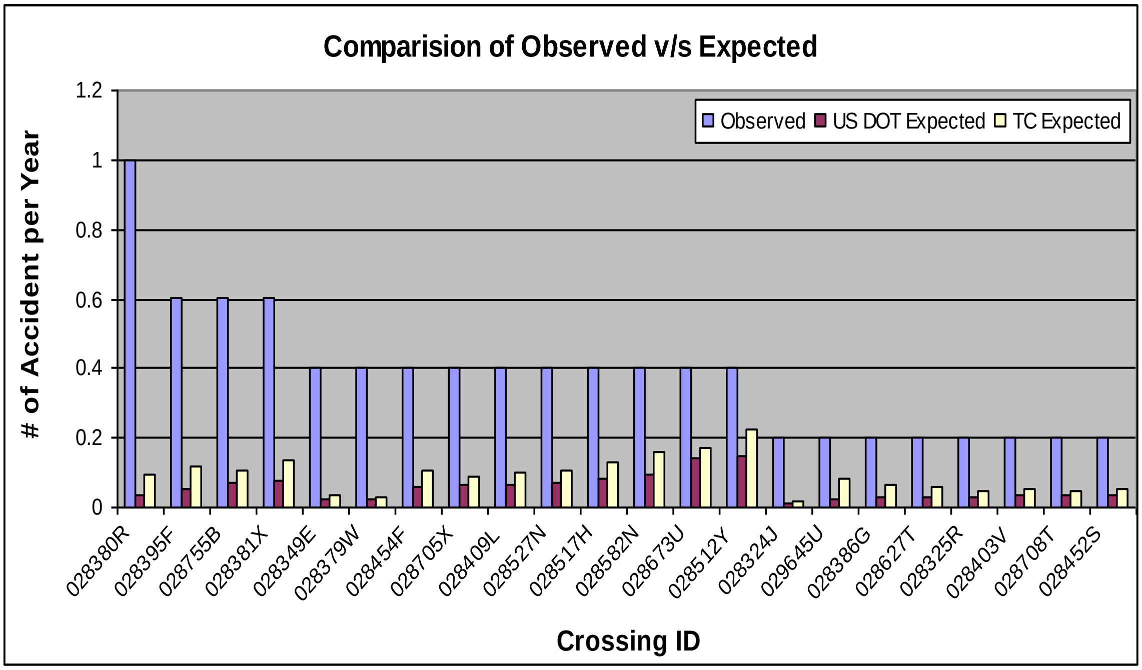


Figure 2. Comparison of Observed vs. Expected Incidents for Top 50 Percent of Incidents

Last, the χ^2 was used to determine which model was a better fit to the observed data. To calculate the χ^2, each model was evaluated by comparing the observed frequency of collisions at grade crossings versus estimated frequency. Initially, the number of collisions was divided into six bins, but since each bin should have at least five crossings with collisions greater than zero for expected frequency, the bins were combined and reduced to two bins [8]. The observed frequency is the number of crossings with the average collision per year for the 5-year time period from 2003 to 2007 for that bin. The estimated frequency is the number of crossings with collisions that were calculated using the TC or USDOT model. Table 7 presents the observed and estimated collision frequencies and their chi-square value.

Table 7. Observed vs. Estimated Frequency of Collisions

Number of Collisions (y)	Observed Frequency of Crossings with y	Estimated Frequency of Crossings with y Collisions		χ^2 for TC Model	χ^2 for U.S. Model
		USDOT Model	TC Model		
0–0.1	175	198	147	2.672	5.333
>0.1	54	31	82	17.065	9.561
Total	**229**	**229**	**229**	**19.736**	**14.894**

The χ^2 was calculated using the following expression:

$$X^2 = \sum \frac{(Observed - Expected)^2}{Expected} \qquad \text{(Eq. 3-4)}$$

A low chi-square value (not statistically significant) suggests a good match between observed and expected results. For either model to be a good fit, the calculated χ^2 value should be less than the critical value at a 5 percent level. Using the critical value for a chi-square distribution table with v = 1 degree of freedom, the critical value was found to be $\chi^2_{0.05} = 3.841$. Since both of the models' χ^2 values are greater than 3.841, there is insufficient information to determine which model is a better fit to the observed collisions. For the 0–0.1 category, the TC model performed better than the USDOT model. But for y > 0.1, the USDOT model performed better than the TC model.

Overall, both models yielded poor χ^2 results. This could be because grade crossing collisions are rare events, and as such, large sample sizes are needed to achieve a reasonable population distribution to obtain more accurate results.

3.4 Summary

Based on the above test cases, the TC model performed better on crossings with some accident history. It is also easier to use, in terms of manipulation, and risk calculation, than the USDOT model. The TC model may be a better model in assessing risk to rank dangerous crossings so that in a cost benefit analysis, the appropriate warning upgrades could be selected. However, this finding is based on using the *San Joaquin* corridor as a sample and a bigger sample would be needed to support this result. The USDOT accident prediction model still performed better when all crossings in the sample were analyzed.

4. Trespass Incident Hotspots

4.1 Cluster Analysis

Cluster analysis is a way of grouping similar objects into respective categories. The goal is to organize observed data into meaningful structures. Cluster analysis is primarily used when there is no theory to explain the facts. It can be used to discover data structures without providing explanations. Cluster analysis is an exploratory data analysis tool that sorts different objects into groups such that the degree of association between two objects is maximal if they belong to the same group and minimal otherwise.

In the railroad environment, cluster analysis could be used to identify hotspots. The concept is similar to epidemiology studies of disease outbreaks in medicine. If trespass incidents were plotted on rail lines and grade crossings, it would be easier to see where there was an unacceptable concentration or cluster of incidents. Furthermore, demographic studies could be performed. For example, compare the hotspots for men versus women; hotspots for teenagers; or urban versus rural. This information could be useful for communities in targeting safety efforts to certain social groups.

4.2 Spatial Testing

Schneider used four different quantitative techniques to evaluate the spatial distribution of police-reported pedestrian crash locations on the UNC Chapel Hill campus [7]. The first two techniques fall under the category of intra-distributional spatial testing, meaning that tests are performed on each spatial distribution separately. The latter two, known as inter-distributional tests, are designed to examine the relationship between points in a specific distribution.

The first of the intra-distributional tests, Ripley's K-statistic, is used to measure clustering by comparing the number of points within a radius, d_s, to the expected number for a spatially random distribution. If the sum of the number of points within d_s around each point is greater than expected for a random pattern with the same radius, then the points have a tendency to be clustered. The K-statistic is expressed as

$$K(d_s) = \left[\frac{A}{N^2}\right]\left[\sum_{ip}\sum_{iq} I(d_{ipiq})\right] \qquad \text{(Eq. 4-1)}$$

Where

A = the total study area in square feet
N = the incident sample size
i_p = the location of the specific incident under study
i_q = nearby incidents of the same type
$I(d_{ipiq})$ = the number of events, i_q, within distance d_s of each event i_p, summed for all events, i_p.

From the above formula, the physical meaning of the K-statistic is not readily apparent. To overcome this challenge, Schneider illustrates a transformation of the K-statistic into what is known as an L-function, as shown below.

$$L(d_s) = \sqrt{\left[\frac{K(d_s)}{\pi}\right]} - d_s \qquad \text{(Eq. 4-2)}$$

In this form, values of $L(d_s)$ greater than zero show a propensity for clustering, whereas values less than zero demonstrate randomness for the radius of distance d_s.

The other intra-distributional test, nearest neighbor analysis, is a tool for discerning locations within a spatial distribution that are nearer to each other than what would be found in a random distribution. This phenomenon, known as clustering, occurs when the mean random distance between the points is less than the minimum random distance calculated from the standard error of a random distribution. This is illustrated in Equation 4-3, below.

$$D_{\min} = 0.5\sqrt{\left(\frac{A}{N}\right)} - t\left[\frac{0.26136}{\sqrt{\frac{N^2}{A}}}\right] \qquad \text{(Eq. 4-3)}$$

Where

A = the total study area in square feet

N = the incident sample size

t = the probability level in the Student's t-distribution

$\left[\frac{0.26136}{\sqrt{\frac{N^2}{A}}}\right]$ = the standard error distance of a random distribution

Inter-distributional testing is used to determine the relationship between points in a distribution. The first test, the widely known chi-square test, can be employed to ascertain if the actual and expected incident locations have a similar spatial distribution. This involves dividing the railroad right-of-way into grade crossings and the segments between them and classifying them according to the number of reported incidents. Here, a segment is defined as the distance between two grade crossings. This can be expressed as:

$$\chi^2 = \sum_{i=1}^{n} \frac{(O_{ir} - O_{ie})^2}{O_{ie}} \qquad \text{(Eq. 4-4)}$$

Where

i = the number of reported incidents on a segment (0, 1, 2, 3…)

O_i = the number of grade crossings/segments with 0, 1, 2, 3, etc., incidents

O_{ie} = the expected number grade crossings/segments with 0, 1, 2, 3, etc., incidents

The second Inter-distributional testing technique is the G-function, which is a measure of the fraction of a specific incident within any given distance of another specific incident type. This function is shown in Equation 4-5 below:

$$\hat{G}(w) = \frac{\#(w_i \leq w)}{n} \qquad \text{(Eq. 4-5)}$$

Where

= "number of incidents"

w_i = the distance from a reported incident, r_1, to the nearest reported incident, r_2

w = a radius around all reported incidents in the study area

n = the total number of incidents in the study area

By plotting the cumulative probability distribution (CDF) of $\hat{G}(w)$ against x, the distance between two nearby events, it is possible to determine if the two events are clustered. If the CDF increases at a high rate for low values of x and then levels off, then the two points are considered to be clustered. Likewise, if the distribution function is flat for low values of x and then accelerates, the events are not clustered and therefore random.

4.3 Data Requirements

This research seeks to determine if areas of high trespass activity are indeed clustered into hotspots or if they are random. This requires resolving the exact geographical coordinates of trespass incidents on a segment of the railroad right of way, as well as the length and endpoint locations of the segment. FRA does not maintain the location of trespass incidents at a resolution less than the county level. Other sources, such as the USCG National Response Center and State and local governments like the California Public Utilities Commission (CPUC) do maintain this information. However, the key is obtaining and verifying the quality of the data. The next step after spatial analysis is to develop a regression analysis model of the expected pedestrian risk locations. Theoretically, the results of one analysis should validate the other and a test of this nature would be a good indicator of the legitimacy of the regression analysis modeling.

Two potential sources of geographical coordinates of segments are U.S. Class I railroads, which maintain detailed track charts of their entire infrastructure, and, as above, State and local government agencies. Also, the FRA Office of Safety is in the process of developing a highly detailed map of the entire U.S. railroad network in GIS format.

Table 8 illustrates a potential list of data requirements for regression modeling development. This list is by no means final, but it represents the types of variables that could be used in the regression modeling. As indicated, some of the data is available from the FRA Rail Accident and Incident Reporting System (RAIRS) database. However, many are not and will have to be obtained from multiple disparate sources, including the ones discussed previously in this section. For the spatial/land use category, potential data sources include the U.S. Census Bureau, which maintains detailed databases that can be exported.

Table 8. Potential Data Requirements

Dependent	Exposure	Rail ROW	Spatial/Land Use
Number of Trespass Incidents/Segment	Segment length	Number of main tracks (RAIRS)	Distance from crossing, bridge, highway, etc.
	Estimated trespassers/day	Number of switch tracks (RAIRS)	Distance from park, beach, etc.
	Estimated trains/day	Maximum Timetable speed (RAIRS	Urban or Rural Location (population density)
	Mix of freight/ passenger trains (RAIRS)	Mainline or Switchyard (RAIRS)	Industrial or Residential
	Number of switch trains/day (RAIRS)		

4.4 Collaboration with the FRA Office of Safety

In March 2008, the Volpe Center established ties with FRA's Office of Safety, which is independently investigating approaches that target trespass hotspot locations for enforcement and/or education. The intent is to effectively coordinate and share research by leveraging the strengths of both organizations and in the process ensure that duplication of effort is minimized. This effort involves two tasks, as described below.

1. The objective of this task was twofold. The first was to verify the benefits offered by current GIS software tools by mapping USCG National Response Center generated data onto the FRA GIS rail network described in Section 4.3. The second was to compare the quality of the two databases, identify areas of disagreement, and seek a resolution.

   By means of the USCG National Response Center online database, the Volpe Center created a database of trespass incidents for California for the years 2003–2007. The USCG National Response Center database contains many of the same data fields as the FRA RAIRS. Although the USCG database is incomplete, it does contain some data fields not found in RAIRS, such as latitude/longitude, railroad subdivision name, milepost number, and sometimes the nearest station. This information is critical to geolocating the exact position of a trespass incident. In the future, it is anticipated that the latitude and longitude data will be added to the RAIRS database.

   For the years 2003–2007, the Volpe Center built a database of trespass incidents in California from USCG National Response Center master database. The California data contains 477 trespass incidents, as compared with the 686 trespass incidents found in RAIRS for the same years. After receiving guidance from FRA, the Volpe Center reworked the USCG National Response Center data structure into a format that was compliant with the FRA GIS rail network, and transmitted it to the FRA Office of Safety in May 2008. In September 2008, FRA transmitted the California trespass data back to the Volpe Center mapped onto the FRA GIS rail network. This map is shown in Figure 3.

2. The second task involved comparing the quality of CPUC trespass incident data against USCG National Response Center data between 2001 and 2007. In July 2008, the FRA Office of Safety transmitted the CPUC trespass to the Volpe Center. This map is shown in Figure 4.

Figure 3. USCG National Response Center Trespass Data, 2003–2007

Figure 4. CPUC Trespass Data, 2001–2007

4.5 Summary

Spatial analysis testing of pedestrian-vehicle accidents in the highway domain has yielded beneficial results and shows promise for the identification of railroad trespass hotspots. This approach will improve the ability of researchers to discern hotspot clusters from random events. As with much of this research, the quality of the railroad trespass incident data will have significant impact on the validity of the test results. This holds true for the regression model development as well. As this is new research, the complete set of hotspot data requirements and potential data sources have not been thoroughly identified. The Volpe Center has begun to explore how this information can be gathered from other sources, including the U.S. Census Bureau, the USCG National Response Center, and the CPUC. To this end, the Volpe Center is collaborating with FRA's Office of Safety to rationalize the information in these sources.

5. Conclusions

Preliminary results of testing the TC accident prediction model with U.S. grade crossing inventory and accident history data have been positive. In the sample used, the *San Joaquin* corridor, the TC model gave a more accurate accident prediction for crossings with accident history. But looking at all crossings, including those with no accident history, the USDOT model gave a more accurate accident forecast than the TC model. However, it should be noted that the TC model may not have performed well when looking at all crossings because of some of the assumptions that were made in this project's adaptation of the TC model to U.S. data. Some of the data that was used in Canada for the TC model is not available in the U.S. crossing inventory database. These assumptions may have led to overestimation.

Further analysis is still required to assess the portability of the TC model in its present form to U.S. data. Theoretically, there should be little distinction between U.S. and Canadian data since the railroads and highways in both countries are constructed using North American design standards and practices. However, differences in the distribution of grade crossing inventory and accident history data could yield disparate accident prediction and consequence models for each country. Ideally, the TC model approach should be duplicated with U.S. grade crossing inventory data and accident history to develop U.S. specific models. As mentioned earlier, the TC model makes it easier to manipulate, run, and calculate the risks than the USDOT grade crossing model. This could make it more desirable for states or communities looking to customize the model to fit their environments. It could be used in assessing risk to rank dangerous crossings. Then, in a cost benefit analysis, the appropriate warning upgrades could be selected.

Currently, FRA is mapping trespass incidents from raw data. Although this is valuable, a prediction model would have even more benefits. The development of a trespass hotspot prediction model would allow researchers to evaluate the impact of various treatment options prior to actually implementing them. Also, if the hotspots could be mapped on a GIS platform, more possibilities for trespass incident predictions could exist. More detailed demographic studies could also be performed.

However, the randomness of many trespass incidents may preclude the feasibility of constructing a reliable prediction model. Spatial analysis, which has shown promise in the evaluation of vehicle collisions with pedestrians, is a viable alternative approach and is worth pursuing.

As with most research, finding good data sources can be difficult, but they are the key to a good analysis. In its current state, the FRA trespass incident database does not provide the necessary location resolution to perform spatial analysis. This information may reside in other databases, including the USCG National Response Center and those maintained by individual State and local agencies. It is important to reconcile the structural and reporting differences among disparate databases, as well as validate the quality of the data sources. This is imperative to obtaining a good sample size.

Another lesson learned is that by looking at research literature for different modes, it is possible to come up with potential tools that can be applied in the railroad environment. The theory of cluster analysis was found by looking at work involving the study of pedestrian incidents involving cars on a university campus. The UNC research broke the campus roadway network into intersections and road segments, but the same theory could be applied to trespass incidents in the railroad infrastructure.

6. References

[1] da Silva, M. (2008). *Railroad Infrastructure Trespass Detection Performance Guidelines.* Report No. DOT/FRA/ORD-11/01. Washington, DC: Federal Railroad Administration.

[2] Hauer, E., and Persaud, B. N. (1987). *How to Estimate the Safety of Rail-Highway Grade Crossings and the Safety Effects of Warning Devices*. Transportation Research Record 1114, pp. 131 – 140. National Research Council, Washington, DC.

[3] Mengert, P. (1979). *Rail-Highway Crossing Hazard Prediction Results*. Washington DC: Federal Railroad Administration, U.S. Department of Transportation.

[4] Farr, E. H. (1987). *Summary of the DOT Rail-Highway Crossing Resource Allocation Procedure-Revised.* Washington DC: Federal Railroad Administration, U.S. Department of Transportation.

[5] Saccomanno, F., Fu, L., Ren, C., and Miranda, L. (2003). *Identifying Highway-Railway Grade Crossing Black Spots: Phase 1*. Montreal, Canada: Transport Canada Development Centre.

[6] Lee, J., Nam, D., and Park, D. (2005). *Analyzing the Relationship between Grade Crossing Elements and Accidents*. Journal of the Eastern Asia Society for Transportation Studies. Volume 6, pp.3658-3668.

[7] Schneider, R.J. (2001). *Development of a Proactive Approach to Pedestrian Safety Planning*. Chapel Hill, NC: University of North Carolina.

[8] Walpole, R., and Myers, R. (1989). *Probability and Statistics for Engineers and Scientists*, Fourth Edition. New York, New York: Macmillan Publishing Company.

Appendix A
Observed versus Estimated Accident

CROSSING	Observed Average per Year (5 Yrs)	PRED ACCID/YR	TC Model Results
028380R	1	0.03402	0.09363
028395F	0.6	0.05483	0.11952
028755B	0.6	0.07027	0.10437
028381X	0.6	0.07444	0.1354
028349E	0.4	0.02591	0.03823
028379W	0.4	0.02608	0.03182
028454F	0.4	0.06117	0.10802
028705X	0.4	0.06423	0.0913
028409L	0.4	0.06686	0.09932
028527N	0.4	0.06809	0.10348
028517H	0.4	0.08295	0.13111
028582N	0.4	0.09711	0.15679
028673U	0.4	0.14032	0.17011
028512Y	0.4	0.14682	0.22388
028324J	0.2	0.01412	0.01539
029645U	0.2	0.02153	0.08345
028386G	0.2	0.02968	0.06656
028627T	0.2	0.03098	0.05724
028325R	0.2	0.03132	0.04842
028403V	0.2	0.03501	0.05482
028708T	0.2	0.03619	0.04712
028452S	0.2	0.0384	0.05107
028310B	0.2	0.03889	0.07174
028459P	0.2	0.04018	0.05392
029697L	0.2	0.04084	0.07283
028623R	0.2	0.04412	0.06091
028687C	0.2	0.04835	0.06668
028591M	0.2	0.04919	0.0694
028343N	0.2	0.05121	0.08647
028456U	0.2	0.0539	0.07668
028556Y	0.2	0.05865	0.08568
028397U	0.2	0.06558	0.09699
028767V	0.2	0.06757	0.14222
028707L	0.2	0.06762	0.09967
029604P	0.2	0.06953	0.11095
028478U	0.2	0.0803	0.12363
028688J	0.2	0.08093	0.14968
029608S	0.2	0.08366	0.09854
028739S	0.2	0.08477	0.13068
028732U	0.2	0.08527	0.1316
028583V	0.2	0.08559	0.13476
028780J	0.2	0.08572	0.15892
028752F	0.2	0.0865	0.13389
028682T	0.2	0.09019	0.14076
028410F	0.2	0.10544	0.12199
028585J	0.2	0.10863	0.17933
028528V	0.2	0.1116	0.18708
029654T	0.2	0.11321	0.1416
028706E	0.2	0.11946	0.14026

CROSSING	Observed Average per Year (5 Yrs)	PRED ACCID/YR	TC Model Results
028781R	0.2	0.13212	0.18989
028580A	0.2	0.14331	0.17785
028427J	0.2	0.16047	0.20179
028573P	0.2	0.19654	0.18478
029598N	0.2	0.19804	0.25511
028328L	0	0.01591	0.01776
028787G	0	0.02117	0.02456
029638J	0	0.0213	0.02661
028632P	0	0.02218	0.02671
028633W	0	0.02244	0.02708
028789V	0	0.02304	0.02719
028645R	0	0.02317	0.02814
028326X	0	0.02331	0.02808
029637C	0	0.02441	0.03132
028635K	0	0.02454	0.03015
029770G	0	0.02462	0.03165
028327E	0	0.02554	0.03132
028383L	0	0.02554	0.03132
028709A	0	0.0266	0.03258
028728E	0	0.02716	0.0396
028631H	0	0.02723	0.03416
028660T	0	0.02738	0.03373
028308A	0	0.0278	0.03467
028440X	0	0.0278	0.03467
028322V	0	0.0278	0.04197
028385A	0	0.0278	0.04197
028453Y	0	0.0278	0.04197
028401G	0	0.02795	0.04185
028791W	0	0.02798	0.04154
028753M	0	0.02809	0.03478
028674B	0	0.0282	0.03494
028671F	0	0.0282	0.04231
028670Y	0	0.0282	0.05122
028716K	0	0.02847	0.03534
028316S	0	0.02859	0.03585
028643C	0	0.02974	0.03796
028321N	0	0.03003	0.03803
028348X	0	0.03011	0.0378
028384T	0	0.03101	0.04785
028438W	0	0.03132	0.04
028455M	0	0.03132	0.04
028441E	0	0.03132	0.04842
028638F	0	0.03137	0.07182
028309G	0	0.03187	0.05335
028344V	0	0.03325	0.04257

028639M	0	0.03366	0.05332
028637Y	0	0.03373	0.06472
028408E	0	0.03377	0.0525
028595P	0	0.03411	0.05417

CROSSING	**Observed Average per Year (5 Yrs)**	**PRED ACCID/YR**	**TC Model Results**
028431Y	0	0.0344	0.04478
028784L	0	0.03511	0.04503
028461R	0	0.03524	0.04607
028782X	0	0.03577	0.04604
028434U	0	0.0364	0.04789
028457B	0	0.0364	0.04789
028433M	0	0.03667	0.04832
029685S	0	0.03709	0.07582
029768F	0	0.03787	0.05303
028717S	0	0.03795	0.04989
028783E	0	0.03834	0.05003
028786A	0	0.03838	0.0501
028317Y	0	0.0384	0.06183
029693J	0	0.03847	0.06544
029639R	0	0.03887	0.05541
028788N	0	0.03889	0.0509
028315K	0	0.04035	0.05419
028329T	0	0.04068	0.05473
028640G	0	0.04084	0.09452
028734H	0	0.04086	0.0545
028715D	0	0.04102	0.06628
028306L	0	0.04148	0.05602
029573T	0	0.04153	0.06667
028662G	0	0.04175	0.05592
028437P	0	0.04179	0.08285
028721G	0	0.04233	0.05685
029698T	0	0.0428	0.06221
028628A	0	0.04315	0.07181
028458H	0	0.04327	0.05892
028598K	0	0.04366	0.06015
028618U	0	0.04366	0.06015
028607G	0	0.04366	0.08817
028775M	0	0.04396	0.05796
028730F	0	0.04396	0.05949
029677A	0	0.04454	0.07799
028393S	0	0.04463	0.06115
028446N	0	0.04463	0.06115
028442L	0	0.04463	0.07403
029641S	0	0.04469	0.06617
028406R	0	0.04545	0.09076
028445G	0	0.04709	0.06521
028334P	0	0.04766	0.08009
028439D	0	0.04821	0.06707
028449J	0	0.04821	0.06707
028341A	0	0.04835	0.11833
028714W	0	0.04879	0.06741
028300V	0	0.05028	0.08541
028570U	0	0.05035	0.08639
028650M	0	0.05048	0.10493

CROSSING	**Observed Average per Year (5 Yrs)**	**PRED ACCID/YR**	**TC Model Results**
028726R	0	0.05059	0.10318
028736W	0	0.05135	0.10504
029614V	0	0.05136	0.07717
029616J	0	0.05136	0.07717
028606A	0	0.05211	0.09002
028367C	0	0.05211	0.08839
028785T	0	0.05257	0.07305
028330M	0	0.05262	0.07449
028656D	0	0.0531	0.07606
028689R	0	0.05312	0.07464
028624X	0	0.05354	0.07681
028647E	0	0.05391	0.11351
028626L	0	0.05456	0.11517
028733B	0	0.05463	0.07719
028302J	0	0.05472	0.2977
029651X	0	0.05553	0.08475
028422A	0	0.05583	0.07922
029607K	0	0.05589	0.08541
029603H	0	0.05589	0.15157
028462X	0	0.05595	0.08017
028430S	0	0.05643	0.08105
028779P	0	0.05701	0.11801
028400A	0	0.05706	0.09846
028686V	0	0.05815	0.08318
028710U	0	0.05858	0.08392
028675H	0	0.05899	0.08462
028392K	0	0.05912	0.18401
028778H	0	0.0592	0.08281
028744N	0	0.05934	0.18495
029613N	0	0.05968	0.0924
028619B	0	0.0602	0.0884
028394Y	0	0.06042	0.13807
028729L	0	0.06129	0.0886
029647H	0	0.06215	0.097
029650R	0	0.06223	0.09715
028416W	0	0.06256	0.09079
028432F	0	0.06299	0.11083
028323C	0	0.06369	0.09365
029660W	0	0.06398	0.12037
028719F	0	0.06536	0.09569
028724C	0	0.06557	0.09605
028773Y	0	0.0657	0.09383
029643F	0	0.06677	0.183
029649W	0	0.0668	0.12803

029599V	0	0.06721	0.11887
028748R	0	0.06779	0.09998
028601R	0	0.07076	0.10728

CROSSING	Observed Average per Year (5 Yrs)	PRED ACCID/YR	TC Model Results
028735P	0	0.07294	0.15999
028464L	0	0.07353	0.16305
028337K	0	0.07459	0.1429
028391D	0	0.07477	0.11836
028578Y	0	0.07648	0.11777
028423G	0	0.08076	0.12456
028390W	0	0.08191	0.13203
028620V	0	0.08207	0.12816
029606D	0	0.08754	0.10404
028429X	0	0.09028	0.14235
029578C	0	0.09733	0.15437
028746C	0	0.09968	0.15869
028704R	0	0.09985	0.13698
029617R	0	0.09999	0.12202
028584C	0	0.10182	0.16595
028428R	0	0.10873	0.17788
028424N	0	0.10873	0.17789
029732X	0	0.11365	0.20645
029773C	0	0.12163	0.40776
028743G	0	0.12237	0.24813
028273B	0	0.13336	0.16165
028551P	0	0.13796	0.16992
028569A	0	0.13917	0.17171
028553D	0	0.13967	0.14546
028574W	0	0.14544	0.18103
028558M	0	0.14735	0.18387
028425V	0	0.14792	0.18302
028672M	0	0.15188	0.22646
028577S	0	0.15251	0.19162
028669E	0	0.15686	0.23538
028549N	0	0.17131	0.22248
028539H	0	0.20046	0.32518
028667R	0	0.25551	1.14501
029766S	0.4	#VALUE!	0.11346
028745V	0.2	#VALUE!	0.12724
029709D	0	#VALUE!	0.01702
028653H	0	#VALUE!	0.04472
029115E	0	#VALUE!	0.0597
029656G	0	#VALUE!	0.0597
028594H	0	#VALUE!	0.07268
028596W	0	#VALUE!	0.07268
028597D	0	#VALUE!	0.07268
028605T	0	#VALUE!	0.07268
028629G	0	#VALUE!	0.07268
028636S	0	#VALUE!	0.07268
028655W	0	#VALUE!	0.07268

028376B	0	0.07097	0.10467
028725J	0	0.07186	0.15714

028718Y	0	#VALUE!	0.07523
028737D	0	#VALUE!	0.07523
028738K	0	#VALUE!	0.07523

CROSSING	Observed Average per Year (5 Yrs)	PRED ACCID/YR	TC Model Results
028364G	0	#VALUE!	0.07584
028795Y	0	#VALUE!	0.07645
028774F	0	#VALUE!	0.07881
029122P	0	#VALUE!	0.11219
029096C	0	#VALUE!	0.11346
029743K	0	#VALUE!	0.11346
029767Y	0	#VALUE!	0.11346
028622J	0	#VALUE!	0.12293
028652B	0	#VALUE!	0.12293
028345C	0	#VALUE!	0.12618
028740L	0	#VALUE!	0.12724
028772S	0	#VALUE!	0.13329
028776U	0	#VALUE!	0.13329
028370K	0	#VALUE!	0.18156
029678G	0	#VALUE!	0.27162
029717V	0	#VALUE!	#NUM!
029718C	0	#VALUE!	#NUM!
029719J	0	#VALUE!	#NUM!
029755E	0	0.00232	#NUM!
028372Y	0	0.00247	#NUM!

Appendix B
Chi-Square Goodness-of-Fit

Number of Collisions	Observed Collisions	Expected Collisions from TC Model	Expected Collisions from U.S. Model	χ^2 for TC Model	χ^2 for U.S. Model
0 - 0.1	175	198	147	2.672	5.333
0.1 - 0.2	0	29	70	29.000	70.000
0.2 - 0.3	40	2	9	722.000	106.778
0.3 - 0.4	0	0	1	-	-
0.4 - 0.5	10	0	1	-	-
> 0.5	4	0	1	-	-
Total	**229**	**229**	**229**	**753.672**	**182.111**

Number of Collisions	Observed Collisions	Expected Collisions from TC Model	Expected Collisions from U.S. Model	χ^2 for TC Model	χ^2 for U.S. Model
0 - 0.1	175	198	147	2.672	5.333
0.1 - 0.3	40	31	79	2.613	29.165
>0.3	14	0	3	-	-
Total	**229**	**229**	**229**	**5.285**	**34.498**

Number of Collisions	Observed Collisions	Expected Collisions from TC Model	Expected Collisions from U.S. Model	χ^2 for TC Model	χ^2 for U.S. Model
0 - 0.1	175	198	147	2.672	5.333
>0.1	54	31	82	17.065	9.561
Total	**229**	**229**	**229**	**19.736**	**14.894**

Abbreviations and Acronyms

AADT	average annual daily traffic
CPUC	California Public Utilities Commission
FRA	Federal Railroad Administration
GIS	geographic information system
km/h	kilometers per hour
mph	miles per hour
RSPD	road speed
TC	Transport Canada
UNC	University of North Carolina
USCG	U.S. Coast Guard
USDOT	U.S. Department of Transportation
Volpe Center	John A. Volpe National Transportation Systems Center

Glossary of Statistical Terms

Bayes' Theorem: relates the conditional and marginal probabilities of two random events. For example, the probability of Event A occurring, given that Event B has occurred.

Chi-squared test: a statistical hypothesis test where the test statistic has a chi-square distribution when the null hypothesis is true.

Colinearity: indicates that a set of points are on a single straight line.

Cluster analysis: the classification of objects into different groups.

Dispersion: the variability or spread in a variable or probability distribution. For example, if all the data points in the sample are identical, then the dispersion would be zero. If the data points differ greatly from one another, then there would be overdispersion.

Empirical Bayes method: a method that uses empirical data to evaluate the conditional probability distributions that arise from Bayes' theorem.

G-function: an inter-distributional spatial test that is a measure of the fraction of a specific incident within any given distance of another specific incident type.

Goodness-of-fit: measures how well a statistical model fits the observations by the discrepancy between observed values and the values expected under the model.

Inter-distributional spatial testing: tests are designed to examine the relationship between points in a specific distribution.

Intra-distributional spatial testing: tests are performed on each spatial distribution separately.

Linear regression: a form of regression analysis where the relationship between one or more independent variables and a dependent variable is modeled by a least squares function, the linear regression equation. This equation is a linear combination of one or more model parameters, which are the regression coefficients. For example, a linear regression equation with one independent variable represents a straight line.

Mean: the expected value of a random variable.

Nearest neighbor analysis: an intra-distributional spatial test that looks for locations within a spatial distribution that are nearer to each other than what would be found in a random distribution.

Negative binomial probability distribution: a discrete probability distribution. For example, it can be used if one wants to know how many times a coin will be tossed to get k number of heads.

Negative binomial regression: an extension of the Poisson regression model that allows the variance of the process to differ from the mean.

Nonlinear regression: a form of regression analysis where observational data are modeled by a function which is a nonlinear combination of the model parameters and depends on one or more independent variables. Exponential and logarithmic functions are examples of nonlinear regression.

Normalize: to make something more normal; to remove statistical error in repeated measured data

Poisson probability distribution: a discrete probability distribution that expresses the probability of a number of events occurring in a fixed period of time if these events occur with a known average rate and independently of the time since the last event. For example, an event could be modeled as a Poisson distribution would be how many customers a cashier rings up in an hour.

Poisson regression: assumes the response variable Y has a Poisson distribution and the logarithm of its expected value can be modeled by a linear combination of unknown parameters. For example, it can be used when the outcome variable is comprised of counts, often rare events.

Regression analysis: techniques for the modeling and analysis of numerical data consisting of values of a dependent variable and of one or more independent variables. The dependent variable in the regression equation is modeled as a function of the independent variables, constants, and a random variable. For example, regression can be used for forecasting data in a time-series.

Ripley's K-function: an intra-distributional spatial test that compares the pattern of the data to that produced by a homogeneous Poisson point process, where cases are considered events.

Statistically significant: if a result is unlikely to have occurred by chance.

Spatial analysis: methods to study entities using their topological, geometric, or geographic properties.

Variance: a measure of statistical dispersion; averaging the squared distance of its possible values from the mean.