ELSEVIER

Accident Analysis and Prevention 38 (2006) 346–356

ACCIDENT
ANALYSIS
&
PREVENTION

www.elsevier.com/locate/aap

Accident prediction model for railway-highway interfaces

Jutaek Oh[a,*], Simon P. Washington[b,1], Doohee Nam[a,2]

[a] *The Korea Transport Institute, Ilsan, Koyang-city, Kyeonggi-do 411-701, South Korea*
[b] *Department of Civil and Environmental Engineering, Arizona State University, Tempe, AZ 85287-5306, USA*

Received 15 April 2005; received in revised form 29 September 2005; accepted 14 October 2005

Abstract

Considerable past research has explored relationships between vehicle accidents and geometric design and operation of road sections, but relatively little research has examined factors that contribute to accidents at railway-highway crossings. Between 1998 and 2002 in Korea, about 95% of railway accidents occurred at highway-rail grade crossings, resulting in 402 accidents, of which about 20% resulted in fatalities. These statistics suggest that efforts to reduce crashes at these locations may significantly reduce crash costs.

The objective of this paper is to examine factors associated with railroad crossing crashes. Various statistical models are used to examine the relationships between crossing accidents and features of crossings. The paper also compares accident models developed in the United States and the safety effects of crossing elements obtained using Korea data.

Crashes were observed to increase with total traffic volume and average daily train volumes. The proximity of crossings to commercial areas and the distance of the train detector from crossings are associated with larger numbers of accidents, as is the time duration between the activation of warning signals and gates. The unique contributions of the paper are the application of the gamma probability model to deal with underdispersion and the insights obtained regarding railroad crossing related vehicle crashes.

Keywords: Railroad crossings; Poisson; Gamma; Negative binomial; International comparison

1. Introduction

The relationships between vehicle accidents and geometric design of road sections, such as horizontal curvature, vertical grade, lane width, and shoulder width, have been extensively studied using regression models (Abbess et al., 1981; Hauer et al., 1988; Persaud and Dzbik, 1993; Kulmala, 1995; Poch and Mannering, 1996; Lord, 2000; Ivan et al., 2000; Lyon et al., 2003; Miaou and Lord, 2003; Oh et al., 2003). A similarly large literature exists for research on roadway intersections. Relatively little research, however, has been carried out to understand and identify factors that contribute to accidents on railway-highway crossings (RHXs). Because of the lack of detailed information on crossing element data and the failure of selecting appropriate tools for analyzing the data, statistical models explaining the relationships between roadway geometry, grade crossing characteristics, and crossing accident frequencies have rarely been developed. Therefore, research gaps remain regarding the identification of factors associated with crashes at RHXs.

This paper seeks to fill some of the knowledge gaps regarding complex relationships between crashes involving motor vehicles and/or trains through the application of statistical models to crash data. The contrasts and tradeoffs between various probability models are discussed with the main focus to provide the greatest insight into RHX related crashes, especially in Korea.

In Korea, about 95% of railroad accidents are associated with RHXs, and a total of 402 accidents occurred at these locations from 1998 to 2002, of which about 20% resulted in fatalities. These statistics suggest that efforts are needed to develop effective countermeasures to reduce crossing accidents. However, considering the fact that about 1800 crossings are located in Korea, 80 accidents per year suggest that for a period of 5 years most of the crossings observed zero accidents.

Statistical models are used to examine the relationships between crossing accidents and features of crossings. However, many past studies illuminating the numerous problems with linear regression models (Joshua and Garber, 1990; Miaou and Lum, 1993) have led to the adoption of more appropriate

[*] Corresponding author. Tel.: +82 31 910 3174; fax: +82 31 910 3235.
E-mail addresses: jutaek@koti.re.kr (J. Oh), Simon.Washington@asu.edu (S.P. Washington), doohee@koti.re.kr (D. Nam).
[1] Tel.: +1 480 965 2220, fax: +1 480 965 0557.
[2] Tel.: +82 31 910 3092, fax: +82 31 910 3235.

0001-4575/$ – see front matter © 2005 Elsevier Ltd. All rights reserved.
doi:10.1016/j.aap.2005.10.004

regression models such as Poisson regression which is used to model data that are Poisson distributed, and negative binomial (NB) model which is used to model data that have gamma distributed Poisson means across crash sites—allowing for additional dispersion (variance) of the crash data. Although the Poisson and NB regression models possess desirable distributional properties to describe motor vehicle accidents, these models are not without limitations. One problem that often arises with crash data is the problem of 'excess' zeroes (see Lord et al., 2004; Washington et al., 2003 for detailed discussions), which often leads to dispersion above that described by even the negative binomial model. 'Excess' does not mean 'too many' in the absolute sense, it is a relative comparison that merely suggests that the Poisson and/or negative binomial distributions predict fewer zeroes than present in the data. As discussed in Lord et al. (2004), the observance of a preponderance of zero crashes results from low exposure (i.e. train frequency and/or traffic volumes), high heterogeneity in crashes, observation periods that are relatively small, and/or under-reporting of crashes, and not necessarily a 'dual state' process which underlies the 'zero-inflated' model. Thus, the motivation to fit zero-inflated probability models accounting for excess zeroes often arises from the need to find better fitting models which from a statistical standpoint is justified; unfortunately, however, the zero-inflated model comes also with "excess theoretical baggage" that lacks theoretical appeal (see Lord et al., 2004).

Another problem not often observed with crash data is underdispersion—where the variance of the data is less than the expected variance under an assumed probability model (e.g. the Poisson). One manifestation might be "too few zeroes", but this is not a formal description. Underdispersion is a phenomenon which has been less convenient to model directly than overdispersion mainly because it is less common observed. Winkelman's gamma probability count model offers an approach for modeling underdispersed (or overdispersed) count data (Winkelmann and Zimmermann, 1995), and therefore may offer an alternative to the zero-inflated family of models for modeling overdispersed data as well as provide a tool for modeling underdispersion.

The paper also compares the results of accident models developed in the United States and the models developed here. The examination of international differences may illuminate differences in design, driver behavior, or both that may lead to different crash processes and generate ideas for further research.

The remainder of the paper is organized as follows. First, previously estimated RHX accident prediction methods are examined to both understand relationships and to identify potential areas for model improvement. Second, probability models are described. Third, the relative merits of probability models are contrasted and evaluated. Then, the modeling results are compared with modeling results from the US. The final section concludes the study.

2. Literature review

Safety levels at highway-rail crossings continue to be of major concern despite of improved design and application practices. This speaks directly to the need to re-examine both accident prediction methods and application practices at railway-highway crossings. With respect to previously developed accident prediction methods, the Peabody Dimmick Formula, the New Hampshire Index and the National Cooperative Highway Research Program (NCHRP) Hazard Index, all lack descriptive capabilities due to their limited number of explanatory variables. The US Department of Transportation's (USDOT) Accident Prediction Formula, which is most widely used, also has limitations related to the complexity of the three-stage formula and its decline in accident prediction model accuracy over time.

2.1. Peabody Dimmick Formula

The Peabody Dimmick Formula, one of the earliest railway-highway crossing accident prediction model, was developed in 1941 using accident data from rural railway-highway crossings in 29 states in US. This formula is primarily used for resource allocation at railway-highway crossings. The formula uses AADT (Average Annual Daily Traffic), P (0, 1) is the presence of warning devices, and average daily train traffic, and is given by:

$$A_5 = \frac{1.28(V^{0.17} \times T^{0.151})}{P^{0.0171} + K} \tag{1}$$

where A_5 is the expected number of accidents in 5 years; V is the average annual daily traffic (AADT); T is the average daily train traffic; P is the protection coefficient indicative of warning devices present; K is the additional parameter (Federal Highway Administration, 1986).

Clearly, the Peabody Dimmick Formula is derived from a stochastic model where $V \times T$ represents the interaction of motor vehicle traffic and trains, and coefficients were derived from a data fitting procedure. As such, the coefficients are based on data that are 20 years old.

2.2. New Hampshire Index

The next evolutionary step in railway-highway crossing accident prediction models was the New Hampshire Index (Austin and Carson, 2002). The New Hampshire Index is given as:

$$\mathrm{HI} = V \times T \times P_{\mathrm{f}} \tag{2}$$

where HI is the hazard Index; V is the average annual daily traffic (AADT); T is the average daily train traffic; P_{f} is the protection factor indicative of warning devices present (Federal Highway Administration, 1986).

Note the similarity in variables used to predict railway-highway crossing accident as compared to the Peabody Dimmick Formula. The protection factor varies from state to state, and this has raised concern over its validity in accurately predicting railway-highway crossing accidents.

2.3. NCHRP Hazard Index

Following the development of the New Hampshire Index, the National Cooperative Highway Research Program (NCHRP) Hazard Index was developed in 1964 with a joint effort between the American Association of State Highway Officials (ASSHTO) and the Association of American Railroads (AAR) (Austin and Carson, 2002).

The program provides for the installation of automatic warning devices at crossings on a statewide basis according to a hazard index which is calculated using ADT, number of train movements, latest 5-year accident experience, and a protection factor based on the type of existing warning devices at the crossing. Based on the program, the DOT in the state of Connecticut designs and constructs between five and six crossings per year (Federal Railway Administration, 1996). Since the start of the program, 120 crossings have received safety improvements. The NCHRP Hazard Index resembles the basic formula for the New Hampshire Index, and is given by:

$$\mathrm{EA} = A \times B \times \mathrm{CTD} \tag{3}$$

where EA is the expected accident frequency; A is the vehicles per day factor provided in tabular format as a function of vehicles per day; B is the protection factor indicative of warning devices present; CTD is the current trains per day (Schoppert et al., 1968).

2.4. USDOT accident prediction equations

The US Department of Transportation (USDOT) accident equations were developed in the early 1980s and sought to address earlier model limitations (Austin and Carson, 2002). The equations were used to assist in the assessment of grade crossing hazards. The USDOT equations consist of a basic equation that predicts the number of crossing collisions and equations that predict the probabilities that collisions will result in fatality or personal injuries. The basic USDOT accident prediction equations consist of an equation for each of the three categories of warning devices: passive devices only, flashing lights, and automatic gates (see Table 1). The three equations generally utilize two highway, three railroad and two combination railway-highway factors.

Various other collision prediction or hazard index formulas utilize greater or fewer numbers of factors, many that are identical or similar to the USDOT equation factors. It may be noted that the results of the various formulas when applied to a group of crossings generally rank crossings in the same order though the predicted number of collisions or hazard index at a particular grade crossing as calculated by the various formulas differs. The Federal Highway Administration's *Railroad-Highway Grade Crossing Handbook* provides information concerning some of the various formulas (Federal Highway Administration, 1986).

$$a = K \times \mathrm{EI} \times \mathrm{MT} \times \mathrm{DT} \times \mathrm{HP} \times \mathrm{HL} \times \mathrm{MS} \times \mathrm{HT} \tag{4}$$

where K is the formula constant; EI is the exposure index; MT is the main tracks; DT is the day through trains; HP is the highway paved; HL is the highway lanes; MS is the max train speed; HT is the highway type.

The USDOT equation factors are based on crossing characteristics that are identified in the national crossing inventory database. The basic two USDOT equation highway factors are based on the number of highway lanes and whether or not the highway is paved (a third highway factor, the type of highway: urban or rural arterial, collector, local road, etc., was removed from the equation when it was updated in 1987). The three railroad factors are based on the number of main tracks, number of daylight through trains per day, and the maximum authorized train speed. The two combination railroad-highway factors are the types of warning device (passive, flashing lights only, or gates), and exposure index. The exposure index in turn is based on the product of the annual average number of highway vehicles per day (AADT) and the average total number of train movements per day.

The full USDOT formula includes an adjustment based on recent collision experience, typically the most recent 5-year

Table 1
1987 USDOT accident prediction equations

Accident prediction equation factors	Warning device category		
	Passive	Lights	Gates
K	0.006938	0.0003351	0.0005745
EI	$((ct+0.2)/0.2)^{0.370}$	$((ct+0.2)/0.2)^{0.4106}$	$((ct+0.2)/0.2)^{0.2942}$
MT	1.0	$\mathrm{e}^{0.1917\mathrm{mt}}$	$\mathrm{e}^{0.1512\mathrm{mt}}$
DT	$((d+0.2)/0.2)^{0.178}$	$((d+0.2)/0.2)^{0.1131}$	$((d+0.2)/0.2)^{0.1781}$
HP	$\mathrm{e}^{-0.5766*(\mathrm{hp}-1)}$	1.0	1.0
HL	1.0	$\mathrm{e}^{-0.1826(\mathrm{h1}\text{-}1)}$	$\mathrm{e}^{-0.1420(\mathrm{h1}-1)}$
HT	Factor estimated when equations were revised in 1987		
MS	$\mathrm{e}^{0.0077\mathrm{ms}}$	1.0	1.0

a is the annual collisions unadjusted for accident experience; c is the average number of highway vehicles per day; t is the average total train movements per day; d is the average number of through train per day during daylight; K is the formula constant factor; MT is the number of main tracks; DT is the day thru trains; HT is the highway pavement (yes = 1, no = 2); HL is the number of highway lanes; MS is the maximum timetable speed (mph); HT is the highway type factor (defined as urban and rural, 1 = interstate, . . ., 6 = local) (Schoppert et al., 1968).

period for which complete collision information is available. The numeric value of each of these factors is calculated using the relationship given in Table 1.

3. Statistical methodology

Given the variables shown in Table 2, including the response variable ‘crashes’, various statistical models are considered for revealing relationships in the data. Because of the random, discrete, non-negative nature of accident data, multiple regression models were not considered to be appropriate. The Poisson regression is usually a good modeling starting point, since crash data often are approximately Poisson distributed. When data are observed with overdispersion (crash mean less than crash variance), some modifications to the standard Poisson regression are available. The most commonly applied (and described in the literature) variations include the negative binomial model and the zero-inflated models including both the zero-inflated Poisson (ZIP) and zero-inflated negative binomial (ZINB). A less common model that can deal with overdispersion and underdispersion (crash mean greater than crash variance) is the gamma probability count model. The general forms of these regression models for count data, their limitations, and brief descriptions of their estimation procedures are now presented. In all models ‘subjects’ are crossings (RHXs), and events are crashes.

3.1. The Poisson model

Integer count data are often approximated well by the Poisson distribution. In the Poisson regression model, the expected number of crashes follows a Poisson distribution, where the expected crash count for the ith crossing $\hat{y}_i$, $i=1, \ldots, N$ is a function of covariates X_{ij}, $i=1, \ldots, N$, $j=1, \ldots, M$ so that:

$$y_i \sim Poi(\lambda_i)$$

$$\hat{\lambda}_i = \exp(\beta_0 X_{i0} + \beta_1 X_{i1} + \beta_2 X_{i2} + \ldots + \beta_M X_{iM}) = \exp\left(\sum_{j=1}^{M} \beta_j X_{ij}\right) \quad (5)$$

where the β_j’s are the estimated regression coefficients across covariates $j=1, \ldots, M$ (for the slope intercept model the first covariate is a vector of 1’s) averaged across crossings $i=1, \ldots, N$.

Because the Poisson regression model is heteroscedastic, the model coefficients for (5) are typically estimated via maximum likelihood methods. The likelihood function for the Poisson regression model (Eq. (5)) is given as (with j subscripts removed for simplicity):

$$L(\beta) = \prod_i \frac{\exp[-\exp(\beta X_i)][\exp(\beta X_i)]^{y_i}}{y_i!} \quad (6)$$

The maximum value possible of the likelihood for a given data set occurs if the model fits the data exactly, resulting in a value of 0. In practice, however, the value 0 is a theoretical lower bound, since integer outcomes are observed while the model predicts non-integer outcomes.

If the mean of the crash counts is not equal to the variance (after accommodating a reasonable degree of sampling variability), then the data are said to be over- or underdispersed. In practice, overdispersion is the most commonly observed condition (variance ≫ mean) with respect to crash data, where the ‘extra’ variation is thought to arise from unobserved differences across sites (Washington et al., 2003; Lord et al., 2004).

3.2. The negative binomial model

The negative binomial regression is a commonly applied alternative statistical model to deal with overdispersed data. The negative binomial model takes the relationship between the expected number of accidents occurring at the ith element and the M parameters, $X_{i1}, X_{i2}, \ldots, X_{iM}$, as follows:

$$y_i \sim Poi(\lambda_i)$$

$$\hat{\lambda}_i = \exp(\beta_0 X_{i0} + \beta_1 X_{i1} + \beta_2 X_{i2} + \ldots + \beta_M X_{iM} + \varepsilon_i) = \exp\left(\sum_{j=1}^{M} \beta_j X_{ij} + \varepsilon_i\right) \quad (7)$$

where $\exp(\varepsilon_i)$ is distributed as gamma with mean 1 and variance α^2. The negative binomial distribution arises as a consequence of gamma heterogeneity in Poisson means, hence the name. The effect of the error term in the negative binomial regression model allows for overdispersion of the variance, such that:

$$\mathrm{Var}(y_i) = E(y_i) + \alpha E(y_i)^2 \quad (8)$$

where α is the overdispersion parameter.

If overdispersion, α, equals 0, the negative binomial reduces to the Poisson model. The larger the value of α, the more variability there is in the data over and above that associated with the mean $\hat{\lambda}_i$. As is the case for the Poisson regression model, the coefficients β_j are estimated by maximizing the log likelihood $\log_e L(\beta)$.

3.3. The gamma model

Underdispersion – when the crash mean is greater than the crash variance – is a phenomenon which has been much less convenient to model directly. The gamma model proposed by Winkelman (1995) provides an approach for count data with under or overdispersion. More detailed discussion can also be found in Cameron and Trivedi (1998).

The gamma probability model for count data is given as:

$$\Pr[y_i = j] = \mathrm{Gam}(\alpha j, \lambda_i) - \mathrm{Gam}(\alpha j + \alpha, \lambda_i) \quad (9)$$

Table 2
Summary statistics of key variables

Variables	Mean	Median	Minimum	Maximum	Frequency
Number of accidents	0.33	0	0	1	162
Number of tracks					
1					100 (61.73%)
2					62 (38.27%)
Average daily traffic (ADT) (vehicles per day)	4617	671	10	61199	162
Average daily railway traffic (trains per day)	70	57	32	203	162
Road pavement					
0 (Paved)					156 (96.30%)
1 (Not paved)					6 (3.70%)
Number of lanes	1.6	1	1	6	162
Road width (m[a])	5.9	1	2	37.5	162
Grade crossing width (m)	5.9	4.5	2.5	39.7	162
Presence of adjacent residential area					
0 (No)					109 (67.28%)
1 (Yes)					53 (32.72%)
Presence of adjacent commercial area					
0 (No)					13 (8.02%)
1 (Yes)					149 (91.98%)
Presence of agricultural area					
0 (No)					37 (22.84%)
1 (Yes)					125 (77.16%)
Presence of adjacent school					
0 (No)					146 (90.12%)
1 (Yes)					16 (9.88%)
Distance to nearest school (m)	3.2	0	0	200	162
Presence of speed hump					
0 (No)					134 (82.72%)
1 (Yes)					28 (17.28%)
Presence of rumble strip					
0 (No)					146 (90.12%)
1 (Yes)					16(9.88%)
Presence of reflecting mirror					
0(No)					107 (66.05%)
1 (Yes)					55 (33.95%)
Distance of train detector (m)	824.5	919.5	0	1329	162
Warning time difference (s)	25.5	20	0	232	162
Crossing angle (°)	6.6	0	0	70	162
Types of barrier					
Duplex semaphore barrier					41 (25.31%)
Simplex semaphore barrier					90 (55.56%)
Vertical-lifting barrier					31 (19.14%)
Types of crossing gate					
Pole gate					19 (11.73%)
Normal gate					143 (88.27%)
Types of interception					
One-way interception					68 (41.98%)
Two-way interception					94 (58.02%)
Types of warning device					
Normal warning device					123 (75.93%)
Suspended warning device					39 (24.07%)
Presence of train moving direction indicator					
0 (No)					103 (63.38%)
1 (Yes)					59 (36.42%)

Table 2 (*Continued*)

Variables	Mean	Median	Minimum	Maximum	Frequency
Control types					
Track circuit					49 (30.25%)
Controller					30 (18.52%)
Others					81 (50.00%)
Presence of a guide					
0 No)					126 (77.78%)
1 (Yes)					36 (22.22%)
Presence of adjacent intersection					
0 (No)					88(54.32%)
1 (Yes)					74 (45.68%)
Presence of traffic signal					
0 (No)					108 (66.67%)
1 (Yes)					54 (33.33%)
Sight distance of grade crossing (m)					
Left-side	85.6	70	9	997	162
Right-side	82.6	65	10	991	162

[a] 1 m = 3.28083 feet.

where:

$$\lambda_i = \exp(\beta' X_i)$$

$$\mathrm{Gam}(\alpha j, \lambda_i) = 1, \quad \text{if} \quad j = 0, \quad \text{or} \quad \frac{1}{\Gamma(\alpha j)} \int_0^{\lambda_i} u^{\alpha j - 1} \mathrm{e}^{-u} \mathrm{d}u$$

$$\text{if} \quad j > 0, j = 0,\ 1,\ \ldots$$

The dispersion parameter is again α; there is underdispersion if $\alpha > 1$, overdispersion if $\alpha < 1$, and equidispersion if $\alpha = 1$, which reduces the gamma probability to the Poisson model. Due to the relative scarcity of the gamma probability count model in the transportation literature, additional model details are provided. The conditional mean function is given by:

$$E[y_i|X_i] = \sum_{j=1}^{\infty} j\mathrm{Gam}(\alpha j, \lambda_i) \tag{10}$$

and the cumulative distribution function is:

$$\begin{aligned} F(T|\alpha, \lambda_i) &= \int_0^T \frac{\lambda_i^{\alpha j}}{\Gamma(\alpha j)} u^{\alpha j - 1} \mathrm{e}^{-\lambda_i u} \mathrm{d}u, \alpha > 0,\ \lambda_i > 0 \\ &= \frac{1}{\Gamma(\alpha j)} \int_0^{\lambda_i T} u^{\alpha j - 1} \mathrm{e}^{-u} \mathrm{d}u,\ j = 0,\ 1, \ldots \\ &= \mathrm{Gam}(\alpha j, \lambda_i T) \end{aligned} \tag{11}$$

3.4. Zero-inflated Poisson model

Zero altered count models, such as the zero-inflated Poisson (ZIP) and zero-inflated negative binomial ZINB models have seen recent attention in crash analysis literature (Miaou, 1994; Shankar et al., 1997). However, as pointed out in Washington et al. (2003) and with greater emphasis in Lord et al. (2004), zero-inflated models may offer improved fit and perhaps better predictive performance, but these models lack theoretical appeal with respect to crash data in most circumstances. So, if statistical fit is the main objective of modelling, then zero-inflated models can often outperform Poisson and negative binomial models. However, if agreement with underlying model theory is paramount (in addition to statistical fit), then alternatives might be sought.

Zero-inflated models are theorized to account for “excess zeroes”—zeroes observed in the data above and beyond the number of zeroes predicted by Poisson or negative binomial models. A troubling assumption (of the zero-inflated theory with respect to crash data) is that excess zeroes may be present because certain crash locations can be considered to be virtually safe—in a zero accident state (Lord et al. (2004) demonstrate that the excess zeroes are not likely to be caused by an underlying zero state but instead by high heterogeneity in crash counts, low exposure, or small spatial or temporal measurement scales). The remaining ‘non-zero’ locations are theorized to follow a normal count process for accident frequency in which non-negative integers (i.e., including zero) are possible accident frequency outcomes over a specified time period. The zip models can be thought of two stage models, where the first stage is a splitting model (e.g. binomial) between two states (zero or count), and the second stage is the count model—Poisson.

The zero-inflated Poisson (ZIP) assumes that the events, $Y = (Y_1, Y_2, \ldots, Y_n)$, are independent and

$$Y_i = 0 \text{ with probability } p_i + (1 - p_i)\mathrm{e}^{-\lambda}$$

$$Y_i = y \text{ with probability} (1 - p_i)\mathrm{e}^{-\lambda i} \frac{\lambda_i^y}{y!} \quad y = 1,\ 2,\ \ldots \tag{12}$$

To test the goodness of fit of a zero-inflated model, Vuong (1989) proposed the following test statistic for non-nested models:

$$m_i = \log\left(\frac{f_1(y_i/x_i)}{f_2(y_i/x_i)}\right) \tag{13}$$

where $f_1(y_i/x_i)$ is the probability density function of the zero-inflated model and $f_2(y_i/x_i)$ is the probability density function of the Poisson or negative binomial distribution. Then Vuong's statistic for testing the non-nested hypothesis of zero-inflated model versus traditional model is (Greene, 1997):

$$v = \frac{\sqrt{n}\left[((1/n))\sum_{i=1}^{n} m_i\right]}{\sqrt{((1/n))\sum_{i=1}^{n}(m_i - \bar{m})^2}} = \frac{\sqrt{n}(\bar{m})}{S_m} \tag{14}$$

where $\bar{m}$ is the mean, S_m is standard deviation, and n is a sample size.

4. Data analysis

4.1. Data description

With consideration given to variables applied in past models and data availability, data were obtained for estimating RHX crash models. Accident records at a total of 162 crossings, which constitute about 10% of railway-highway crossings in Korea, were first sampled randomly from all crash locations. Detailed records of accidents by approach for the 5-year period from 1998 to 2002 were extracted from the Korean National Railroad accident database.

A total of 56 explanatory variables were explored during model development. Some of the main explanatory variables include *Average Daily Traffic*, *Average Daily Train Volumes*, *number of tracks*, *number of roadway lanes*, *road widths*, *crossing widths*, *traffic control devices*, *presence of flashing lights*, *warning times*, *stop signs*, *crossing warning sings*, *speed humps*, *road grade*, *crossing angles*, *roadway speed limits*, *train detector types*, *train detector distances*, and *sight distances*. These railway-highway crossing data available were collected over a period from December 2003 to February 2004 through site visits and investigations.

Several indicator variables were also developed. An *adjacent intersection* variable indicates the possible effect of approaching and departing traffic from an adjacent intersection and is equal to 1 if an adjacent intersection is within 50 m from the crossing under consideration and 0 otherwise. A *land use* variable is created to account for additional 'distractions' and 'complexity' associated with nearby residents, shopping, employees, and schools related activities. It is expected that introduction of these variables may yield significant results as observed in other studies (Laffey, 1999; Long, 2003). Table 2 summarizes the main variables considered in the study.

4.2. Results of the analysis

Table 3
Poisson estimation of railway-crossing accidents ($\alpha = 0.1$)

Variables	Coefficient	*t*-Ratio
Constant	−5.406	−5.94
ADT	0.460	5.09
Presence of commercial area	0.975	2.53
Train detector distance	0.0016	2.61
Presence of track circuit controller	−0.917	−2.72
Presence of guide	−0.613	−1.56
Presence of speed hump	−1.063	−2.50
Log-*L*	−98.35	
G^2	102.80	
R^2	0.27	
R^2	0.30	

The modeling starting point is a Poisson regression model with functional form guided by prior research regarding RHX crashes (and as reflected in the 56 potential explanatory variables). The correlation matrix between variables was estimated to avoid multicollinearity between explanatory variables. A 90% confidence interval was used as an acceptable cutoff for statistical significance. The results of the Poisson model are shown in Table 3. In the Poisson model, *roadway traffic flows, presence of commercial areas near RHX crossings, presence of speed hump*, *track circuit detection system*, *presence of guide at RHX*, *train detector distance from RHX crossings* were found to be statistically significant.

As mentioned in Section 3 of this paper, the regression-based test for overdispersion can determine the appropriateness of negative binomial regression models over Poisson regression models, and Vuong's test statistic is suitable for testing the appropriateness between the zero-inflation model and the traditional Poisson. This decision rule was adopted in selecting the proper econometric method for the RHX crash models.

Because overdispersion has been observed in prior studies, a negative binomial regression model and zero-inflated Poisson model were estimated and compared to the Poisson. The Poisson model without zero inflation accounts sufficiently well for the number of observed zeros in the data, and in fact underdispersion was observed. The negative binomial regression model and zero-inflated model, therefore, are not well suited for the RHX data.

Suspecting possible underdispersion, a gamma probability count model was fit to the data (see Table 4). The results

Table 4
Gamma estimation of railway-crossing accidents ($\alpha = 0.1$)

Variables	Coefficient	*t*-Ratio
Constant	−3.438	−3.41
ADT	0.230	3.01
Average daily railway traffic	0.004	1.66
Presence of commercial area	0.651	2.27
Train detector distance	0.001	2.24
Time duration between the activation of warning signals and gates	0.004	1.82
Presence of speed hump	−1.58	−1.84
α (Dispersion parameter)	2.062	2.72
Log-*L*	−98.69	
Restricted log likelihood	−100.52	
Chi-squared	3.67	

in Table 4 revealed that for any subset of the independent variables, the RHX crash data exhibit underdispersion, where $\alpha = 2.062$ (refer to Winkelmann (1997) and Cameron and Trivedi (1998) for detailed discussions for underdispersion) the model with six covariates and a slope intercept term. The log likelihood ratio test, comparing underdispersion versus equidispersion, favors the former. Thus, the gamma probability model for count data appears to offer considerably good statistical fit with both significant variables and accounts for the mild underdispersion.

In the RHX crash model, *road width*, *crossing width*, *number of tracks*, *existence of nearby schools*, *gate control types*, *crossing angle*, *presence of adjacent intersections*, *crossing shape*, *rumble strip*, *reflecting mirror*, etc. were not statistically significant. Although road and crossing width and number of tracks are thought to be important, the effects of these factors are likely to be captured by the correlated variables, traffic flow and train volumes.

Consistent with the concept of exposure to risk, an aggregate relationship between accident frequency and functions of roadway traffic flows (entering and leaving traffic flows) were explored. Expectedly, the total roadway traffic volume was found to increase annual RHX crashes significantly ($t = 3.01$), similar to findings by others (Austin and Carson, 2002; Hauer et al., 1988; Lyon et al., 2003; Oh et al., 2003, 2004). Not only do higher roadway traffic volumes increase aggregate exposure to risk, but increased roadway traffic volumes at crossings restrict the mobility of motor vehicle drivers at crossings, raising driver frustration levels, leading to greater risk taking on approaches to crossings.

The model results also indicate that crossing accidents increase ($t = 1.66$) as daily train volumes increase, again as expected. Similar to roadway traffic flows, crash opportunities increases with increasing train volumes.

The presence of commercial areas near RHX crossings is associated with higher accidents ($t = 2.27$), which agrees with expectation because relatively more complicated traffic activities occur in commercial areas and because drivers in commercial areas may be relatively unfamiliar with the crossing (compared with residential and agricultural areas). There are likely to be a larger percentage of trucks in these areas also, perhaps reducing visibility for motorists.

Many past studies (Berg et al., 1982; Hopkins, 1981; Mather, 1991) indicate that grade crossing safety increases with provision of reasonable and consistent warning times. This study also supports these findings. In this study, the distance of train detector from crossings was associated with an increase in accidents ($t = 2.24$) and the time duration between the activation of warning signals and the activation of gates increases the probability of accidents ($t = 1.82$). Warning times refer to the time between device activation and arrival of a train at the crossing and are usually controlled by the train detection device. In Korea, less than 5% of crossings are using constant warning time systems. In other words, whenever a train is detected in front of the crossing, warning signals are activated regardless of speeds of the train resulting in inconsistent warning times and potentially excessive delays. Excessively long warning times also have negative impacts on crossing safety and traffic operations because frequent users of a crossings become aware that signals flash ‘too long’ in advance of a train’s arrival and thus proceed through the crossing when the warning device is activated. The drivers’ expectation, then, is that a train crossing is not imminent. By eliminating unnecessarily long and inconsistent warning times, the rate of disobedience towards crossing signals would be expected to decline.

Installation of an appropriate train detection system may also improve crossing safety. Train detection types are important for the crossing safety. Table 3 shows that the track circuit detection system is associated with a decrease in accident probability ($t = -2.72$), while the controller detection system is associated with greater likelihood of crashes. This is explained by the fact that the track circuit system uses lower frequency for train detection than the controller system. Thus, the track circuit system is cheaper to use, while the controller system provides efficient operations if crossings are closely spaced (i.e. 2–3 crossings within 1 km), which is common in Korea. However, the controller system sometimes suffers from detection errors when higher speed trains come because the detection length of the system is relatively short. Therefore, installing an appropriate train detection system should be considered carefully because there are tradeoffs in cost, performance, and safety of various systems. Other train detection systems were found to have insignificant effects RHX crossing safety.

The presence of speed hump decreases the probability of crossing accidents ($t = -1.84$), which agreed with expectations. Because speed humps force motorists to reduce speeds prior to RHX crossings, for a given reaction time (e.g. to the presence of a train) shorter stopping distances are required. Furthermore, perhaps motorists are discouraged from ‘racing’ to beat a train, resulting in fewer illegal crossings.

4.3. Model comparison

Table 5 compares the findings of this study with those from the previous studies. As discussed in the literature review, the Peabody Dimmick Formula, the New Hampshire Index, and the NCHRP model are simple to use, but they lack descriptive capabilities and some their coefficients are derived from fairly dated data. Though much more comprehensive and descriptive than the other models discussed, the USDOT accident formula is also not without limitation. As discussed by Austin and Carson (2002), the complexity of three-stage formula makes it difficult to identify design improvement activities that will be most effective from improving crossing safety.

It is difficult to compare directly the findings (regarding RHX factors) in this study to the USDOT accident formula, mainly because all crossings in Korea are equipped with gates, flashing lights, and warning bells. Looking for commonalities, however, both studies observed that *number of trains* and *ADT* are significant factors associated with RHX crossing accident frequencies.

Surprisingly, crossing and roadway factors found to be significant predictors are different across models. In the USDOT

Table 5
Accident prediction model comparison

Independent variable	Peabody Dimmick Formula	New Hampshire Index	NCHRP Hazard Index	USDOT prediction formula	Gamma model
Constant	○			○	○
Traffic characteristics					
Number of day through trains				○	
Total number of trains	○	○	○	○	○
Maximum timetable speed				○	
Number of main tracks				○	
Number of traffic lanes				○	
AADT in both directions	○	○	○	○	○
Commercial Area					○
Railway Characteristics					
Highway paved or gravel				○	
Highway type factor				○	
Crossing Characteristics				○	
Pavement marking					
Probability of a stop sign				○	
Presence of Speed humps					○
Probability of a gate/gates				○	
Probability of flashing lights				○	
Time between warning signal and gate activation					○
Train detector distance					○
Protection coefficient/warning devices	○	○	○		

accident formula, *type of highway surface*, *presence of stop signs*, and *pavement markings* were found to be significant factors affecting accident frequency, while they were insignificant in the gamma probability model. In contrast, the gamma probability count model includes the *presence of speed humps*, *train detector distance*, the *time duration between warning signal activation and the gate activation*, and *proximity to commercial area* as significant factors in the model.

The differences between models might indicate several things. First, it might indicate that unobserved underlying factors affecting safety are proxied by various easily measured factors, such as *presence of speed humps*, *stop signs*, and *pavement markings*. Second, it may indicate that true differences exist between Korean and US RHX crash experience, perhaps due to differences in driver behavior, design factors, or both. Third, it may indicate that that random fluctuation in crashes in combination with relatively low crash counts pose significant modeling challenges.

5. Summary and conclusion

This paper presents an empirical analysis of RHX accidents in Korea. Various statistical models were estimated to illuminate structure in data with attention paid to appropriate modeling techniques. The modeling results are compared and contrasted to prior US research findings. The study intended to address and explore three main issues:

1. Safety levels at RHX crossings are a major concern for transportation system operators, emphasized by a high proportion of fatalities and serious injuries. Relatively little research has been carried out to understand and identify factors that contribute to accidents on the railway-highway crossings, and only few attempts have been made to quantify the safety effects of geometric factors on crossing safety.
2. The nature of crashes between trains and motor vehicles is fundamentally different than crashes between motor vehicles, since trains operate on fixed guideways, trains have virtually no stopping ability, RHX crossings are relatively rare for motorists (compared with potential vehicle to vehicle conflicts), traffic control measures are unique at RHX crossings, and crashes that result are generally more severe. As a consequence of these differences, existing models for intersection or road segment accidents are not reliable for providing sufficient insight.
3. Due to sometimes stark differences in driving behavior, compliance to traffic laws, and design and operational differences across countries, potential exists for substantially different relationships between RHX crossing features and safety. As a result, different countermeasures may have differing effectiveness across countries.

After careful application and assessment of statistical models, accompanied by a detailed examination of RHX crossing accident models between Korea and the US, the following general conclusions are drawn.

The gamma probability model was the most appropriate statistical model (of the ones fitted) for the RHX crossing accident data, which exhibited slight underdispersion with respect to Poisson distributed counts. The gamma probability count model is a flexible model that can accommodate under and overdispersion, and is relatively new in the transportation safety literature. Significant explanatory factors (and their direction of association) include average *daily traffic volume* (+), *daily train volumes* (+), *proximity of commercial area* (+), *distance of train detector from crossing* (+), *time duration between the activation of warning signals* and t*he activation of gates* (+), *and presence of speed hump* (−) (see Table 4).

Examination of the modeling results suggests that driver warning and operations associated with passing trains are associated with crash risk. Thus, fruitful research (and crash mitigation investments) may include examination (or application) of intelligent transportation system technologies associated with driver warning devices, such as devices that detect and warn approaching vehicles, trains, or both. More careful research is needed into these and other technologies and devices that offer focus on warning of impending conflicts between motor vehicles and trains.

Of the four railway-highway crossing accident prediction models examined, the Peabody Dimmick Formula, the New Hampshire Index, and the NCHRP model are simple to apply, but generally lack adequate descriptive capabilities. The USDOT accident formula is relatively more comprehensive and descriptive, but the complexity of three-stage formula makes it difficult to use. Perhaps more important, the relationships between the RHX crossing and safety may differ between Korea and the US, diminishing the utility of US based models.

With respect to suggestions for future railroad safety research, there is a need to include additional driver, geometric, and operational features in an analysis. For example, congestion levels (driver fatigue, frustration levels, etc.) at the time of the crash may be important, as might be the age and gender (proxy for risk aversion and driving ability) of the motor vehicle driver.

Limitations in observational cross sectional studies cannot be overcome with the conduct of additional ‘similar’ studies. So, the results of this study should be used to inform the design of before–after studies, which in general will yield more reliable results regarding countermeasure effectiveness.

Acknowledgements

The authors wish to express their gratitude to the Korean National Railroad and Dr. Dong Joo Park of University of Seoul for their assistance in acquiring data.

References

Abbess, C., Jarett, D., Wright, C.C., 1981. Accidents at Blackspots: estimating the effectiveness of remedial treatment, with special reference to the “regression-to-mean” effect. Traffic Eng. Control 22, 535–542.

Austin, R., Carson, J., 2002. An alternative accident prediction model for highway-rail interfaces. Accid. Anal. Prev. 34, 31–42.

Berg, W., Knoblauch, K., Hucke, W., 1982. Causal factors in railroad-highway grade crossing accidents. Transport. Res. Record, 847.

Cameron, A., Trivedi, P., 1998. The Analysis of Count Data. Cambridge University Press, New York.

Federal Railroad Administration, 1996. Highway-Rail Grade Crossing Safety Research. Office of Research and Development, Washington, DC.

Federal Highway Administration, 1986. Railroad-Highway Grade Crossing Handbook, second ed. FHWA-TS-86-215, Springfield, Virginia: NTIS.

Greene, W.H., 1997. Econometric Analysis, third ed. Prentice Hall.

Hauer, E., Ng, J.C.N., Lovell, J., 1988. Estimation of safety at signalized intersections. Transport. Res. Record 1185, 48–61.

Hopkins, J.B., 1981. Technological Innovations in Grade Crossing Protection Systems. Report TSC-FRA-71-3, FRA, U.S. Department of Transportation, Cambridge, Mass.

Ivan, J.N., Wang, C., Bernardo, N.R., 2000. Explaining two-lane highway crash rates using land use and hourly exposure. Accid. Anal. Prev. 32, 787–795.

Joshua, S.C., Garber, N.J., 1990. Estimating truck accident rate and involvements using linear and Poisson regression models. Transport. Planning Technol. 15 (1), 41–58.

Kulmala, R., 1995. Safety at Rural Three- and Four-Arm Junctions: Development and Applications of Accident Prediction Models. VTT Publications 233, Technical Research Centre of Finland, Espoo.

Laffey, S., 1999. Grade Crossings of Northeastern Illinois: An Analysis of the FRA Grade Crossing and Grade Crossing Accident Inventories, and an Analysis of the Potential Impacts from the Horn Sounding Requirement of the Swift Rail Development Act, working paper, Chicago Area Transportation Study, Chicago, IL.

Long, G., 2003. Easy-to-apply solution to a persistent safety problem: clearance time for railroad-preempted traffic signals. Transport. Res. Record 1856, 239–247.

Lord, D., 2000. The prediction of accidents on digital networks: characteristics and issues related to the application of accident prediction models. Ph.D. Dissertation, Department of Civil Engineering, University of Toronto, Toronto.

Lord, D., Washington, S., Ivan, J., 2004. Poisson, Poisson-gamma, and zero-inflated regression models of motor vehicle crashes: balancing statistical fit and theory. Accident Analysis and Prevention, Pergamon Press/Elsevier Science, Forthcoming, 2004.

Lyon, C., Oh, J., Persaud, B.N., Washington, S.P., Bared, J., 2003. Empirical investigation of the IHSDM accident prediction algorithm for rural intersections. Transport. Res. Record 1840, 78–86.

Mather, R.A., 1991. Seven years of illumination at railroad-highway crossings. Transport. Res. Record 1316, 54–58.

Miaou, S.P., 1994. The relationship between truck accidents and geometric design of road sections: Poisson versus negative binomial regressions. Accid. Anal. Prev. 26, 471–482.

Miaou, S.P., Lum, H., 1993. Modeling vehicle accidents and highway geometric design relationships. Accid. Anal. Prev. 25, 689–709.

Miaou, S.-P., Lord, D., 2003. Modeling traffic crash-flow relationships for intersections: dispersion parameter, functional form, and Bayes’ versus empirical Bayes’. Transport. Res. Record 1840, 31–40.

Oh, J., Lyon, C., Washington, S.P., Persaud, B.N., Bared, J., 2003. Validation of the FHWA crash models for rural intersections: lessons learned. Transport. Res. Record 1840, 41–49.

Oh, J., Washington, S.P., Choi, K., 2004. Development of accident prediction models for rural highway intersections. In: Proceedings of the Transportation Research Board Annual Meeting, Washington, DC.

Persaud, B.N., Dzbik, L., 1993. Accident prediction models for freeways. Transport. Res. Record 1401, 55–60.

Poch, M., Mannering, F.L., 1996. Negative binomial analysis of intersection-accident frequency. J. Transport. Eng. 122 (2), 105–113.

Schoppert, D.W., Dan, W., Hoyt, Alan M. Voorhees and Associates, 1968. Factors influencing safety at highway-rail grade crossings (NCHRP Report 50). NAS-NRC Publication.

Shankar, V., Milton, J., Mannering, F., 1997. Modeling accident frequencies as zero-altered probability processes: an empirical inquiry. Accid. Anal. Prev. 29 (6), 829–837.

Vuong, Q.H., 1989. Likelihood ratio tests for model selection and non-nested hypotheses. Econometrica 57 (2), 307–333.

Washington, S., Karlaftis, M., Mannering, F., 2003. Statistical and Econometric Methods for Transportation Data Analysis. Chapman Hall/CRC, Boca Raton, FL.

Winkelmann, R., 1997. Economic Analysis of Counter Data, second ed. Springer-Verlag, Heidelberg.

Winkelmann, R., Zimmermann, K., 1995. Recent developments in count data modeling: theory and applications. J. Econ. Surveys 9, 1–24.