

MODELS FOR RANKING RAILWAY CROSSINGS FOR SAFETY IMPROVEMENT

Sandra Kasalica ^{1*}, Marko Obradović ², Aleksandar Blagojević ¹, Dušan Jeremić ¹, Milivoje Vuković ³

¹ Academy of Technical and Art Applied Studies Belgrade,
department High Railway School, Belgrade, Serbia

² Faculty of Mathematics, University of Belgrade, Belgrade, Serbia

³ Infrastructure of Serbian Railways, Belgrade, Serbia

Received: 20 August 2020

Accepted: 11 November 2020

First online: 15 November 2020

Original scientific paper

Abstract: *Analysis of high-risk locations, accident frequency and severity for railway crossing is necessary in order to improve the safety and consequently diminish the number of accidents and their severity. In order to extract the necessary parameters that quantify the risk associated with railway crossings in Serbia, we have carefully analyzed available statistical models commonly used in this kind of studies. A zero-inflated Poisson model and a multinomial logistic model were used for the assessment of accident frequency and accident severity respectively. In order to quantitatively evaluate the risk, a well known measure – total risk was modified and a new measure for risk – empirical risk was introduced. The road sign warning device ($p = 2.76 \cdot 10^{-9}$), exposure to traffic ($p = 4.3 \cdot 10^{-7}$), and maximum train speed at a given crossing ($p = 1.36 \cdot 10^{-5}$) were significantly associated with probability of accident frequency and significantly influenced the expected total number of fatalities or injuries caused by traffic accidents.*

Keywords: *railway crossings, high-risk locations, accidents, regression models*

1. Introduction

The identification of the high-risk railway crossings in Serbia is of great importance because, to our knowledge, no such study has been performed in the past. For example, from 2007 to 2011, 312 accidents occurred at 2,138 railway crossings in Serbia. These accidents resulted in 59 fatalities and 130 injuries (Statistics on accidents at Serbian railways 2011). Currently, more than 74% of the 2,138 railway crossings in Serbia are of passive control type (St. Andrew's cross and Stop sign). From 2004 to 2012, only 22 railway crossings were equipped with an

sandra.kasalica@gmail.com (S. Kasalica)*, relzzup@gmail.com (M. Obradović),
aleksandar.blagojevic23@gmail.com (A. Blagojević), jeremicd@gmail.com (D. Jeremić),
milivoje.vukovic1962@gmail.com (M. Vuković),

active control type (flashing lights with half gates). Along with a control type, other railway crossing factors (e.g., train frequency, train speed, daily road traffic, sight triangle, crossing width and angle) might increase the likelihood of accident occurring at railway crossings. Therefore, investigations of the risk factors that may be associated with accidents at railway crossings are vital in order to identify the crossings for future safety improvement.

Statistical regression models are formulated to express the expected accident count of an entity as a function of its traits. Because of non-negative and count data nature of accident frequency for a time period, the Poisson or negative binomial (NB) models and their variants are developed to model crash or accident frequency, or both, at spatial locations on highways (Miaou, 1994; Persaud et al., 1999; Lord et al., 2005). The NB was verified by (Austin & Carson, 2002). Saccomanno et al. (2004) used Poisson and NB distribution in their risk based statistical models which were also used for identifying the high-risk railway crossings. Extensions of these two models are a zero-inflated Poisson (ZIP) regression model and a zero-inflated negative binomial (ZINB) regression model which have also been utilized for modeling accident data on railway crossings (Miranda-Moreno & Fu, 2006). Three alternative models - the NB model, the heterogeneous negative binomial (HNB) model, and the Poisson lognormal model and two ranking criteria - marginal and posterior mean of accident frequency are considered in a study by Miranda-Moreno et al. (2005) for identification of the high-risk railway crossings. Similarly, Miranda-Moreno et al. (2009) proposed a Bayesian multinomial model to estimate the severity levels of each individual involved in an accident. The generalized logit model was used to explore the key factors that may be responsible for different degrees of accident severity at railway crossings (Huet et al., 2010). A ZIP model was used to describe the relationship between the extra zero count fatality or injury data and explanatory variables on railway crossings in Taiwan (Hu et al., 2011). Rovšek et al. (2014) identified the key risk factors of traffic accident injury severity on Slovenian roads using a non-parametric classification tree. Recently, Washington et al. (2014) applied a quintile regression model for identifying black spots. Moreover, they proposed a more complex formula for modeling a number of crashes based on equivalent property damage crashes.

This paper is the first attempt to analyze accident data using count data and multinomial regression in Serbia and countries in our region. Due to the fact that we have gathered the unique set of the data, we believe that the analyses will serve not only to identify unsafe railway crossings but also additionally verify the methodology used.

Four types of regression models are considered for accident frequency and empirical risk: Poisson, NB, ZIP and ZINB. The analysis of accident severity was performed using a multinomial logit model. The high-risk locations were determined using total risk analysis (Saccomanno et al., 2003). Finding out that variables of accident frequency and accident severity are slightly correlated, we also introduced a new risk measure - empirical risk. The final high-risk location was created using both of these methods.

2. Data source and description

2.1. Inventory data set

The data supporting this research came from two sources; (1) the Serbian railway crossing inventory database (2007-2011) (SRCID) and (2) Accident database of Serbian railway crossings (2007-2011) (ADSRC). The second one is the first database of this kind in Serbia, and we created it for the purpose of this very research.

SRCID contains the characteristics of each railway crossing and its traffic conditions. On the territory of the Republic of Serbia, the railway network has the total length of about 4,000 km, out of which 276 km are multiple tracks and 934 km are electrified. There are 2,138 railway crossings in total. All these crossings have various warning devices.

Certain numbers of crossings were found to be poorly specified. Namely, some attributes associated with railway and highway features and traffic exposure with regard to the number of daily trains and average annual daily traffic (AADT) were missing. In the present study, we have excluded the crossings with incomplete data. In order to avoid possible selection bias, we have carefully analyzed distributions of the excluded crossings and found no significant statistical grouping (see Appendix). The final set was compiled by merging SRCID and ADSRC databases and it consisted of 745 railway crossings.

There were 17 independent variables considered in this study for modeling purpose and they were derived from the SRCID (Table 1).

They can be classified as follows:

- railway characteristics: railway category, maximal train speed at a given crossing and number of tracks.
- traffic volume: Exposure (EXPO) at a given crossing is defined as the geometric mean of number of trains per day and average annual daily traffic volume (AADT).
- crossing characteristics: crossing surface type, crossing width, sight triangle and crossing angle.
- road characteristics: road category mainline, road category regional, road category rural and local, road category farm and non-categorized and road category street.
- warning devices: road signs, flashing lights, full gates and half gates.

Table 1. Independent variables and their characteristics

Variable	Short name	Description	Coding / Unit
x_1	KATPRM	Railway category	Mainlines = 1; Others = 0
x_2	EXPO ^b	Sqrt [AADT · daily trains]	vehicles/day
x_3	MBRZ ^b	Maximal train speed at a given crossing	km/h
x_4	BRKOLB	Number of tracks	Single track = 1; Multiple tracks = 0
x_5	VRKOLA	Crossing surface type	asphalt, concrete panels and rubber panels = 1; cobblestone, wood planks and gravel = 0
x_6	KPM	Road category mainline	Indicator

Models for ranking railway crossings for safety improvement

x ₇	KPR	Road category regional	Indicator
x ₈	KPSL	Road category rural and local	Indicator
x ₉	KPPNP	Road category farm and non-categorized	Indicator
x ₁₀	KPU	Road category street	Indicator
x ₁₁	SIRPPB	Crossing width	6m or less = 1; more than 6m = 0
x ₁₂	TRPRP	Sight triangle	exist = 1; does not exist = 0
x ₁₃	UGUKR	Crossing angle	From 60° to 90° = 1; less than 60° = 0
x ₁₄	VOSIG	Warning devices road signs	Indicator
x ₁₅	VOSV	Warning devices flashing lights	Indicator
x ₁₆	VOBR	Warning devices full gates	Indicator
x ₁₇	VROSP	Warning devices half gates	Indicator

Note: All variables are categorical except EXPO and MBRZ which are numerical;
a The mainline includes reference and intermediate lines and others are supplementary lines.

b In order to get more convenient coefficients for the models, the observed values for maximum train speed at a given crossing and daily traffic volume were divided by ten.

2.2. Accident occurrence data

The available historical accident data-set for modeling accidents at railway crossings were collected from 2007-2011 (5 years of accident information). The data-set provides the information about the time, location and conditions of accident for 2,138 railway crossings, but we observed 745 crossings.

The Accident database of Serbian railway crossings (2007-2011), contains four types of information:

- basic accident data: including the accident reference number, the date and the time of accident, location and cause of accident.
- involved road vehicle driver, vehicle and train data: including information on road vehicle driver action at time of collision (e.g., ignored warning devices, drove through gates, failed to stop), gender and age, visibility, vehicle type and train type. It should be noted that our data lacked the information about the train operator.
- accident type: a road vehicle was hit by a train or a train was hit by a road vehicle.
- data on severity of consequences: including information on the number of fatalities, serious injuries and a property damage-level for each accident.

In this paper, we considered three dependent variables: accident frequency, accident severity and empirical risk. The accident frequency is the number of accidents that took place at a given time period. It is a countable variable that, in our observations, takes values from 0 to 5. The frequency of these values is given in

Table 2. It represents the number of accidents that took place at observed 745 crossings in the period from 2007 through 2011. In this period of time, at 514 (69%) crossings there were no accidents, and at the remaining 231 (31%) crossings there were 312 accidents in total.

Accident severity is defined as an average impact per accident. The average impact is a weighted average of deaths and injuries in each accident. In this paper, the accident severity is characterized as equivalent fatality. For example, Saccomanno et al. (2004) equalized one fatality to 44 injuries to yield a crossing collision consequence score (CS) for the purpose of further crash severity modeling. In Taiwan, one injury from a highway accident has been equalized to 0.37 fatalities, or conversely, a fatality has been treated as 2.72 injuries (Hu et al., 2010). Similarly, we equalized three injuries as an equivalent of one death. The factor 3 was chosen out of practical considerations and matches the regulations for personal compensation stated in the Regulation on personal compensation (Official Gazette of the Republic of Serbia, no. 34/2010). Therefore, the formula for accident severity becomes:

$$\text{Accident severity} = (3 \times \text{fatalities} + \text{injuries}) / \text{accident frequency} \quad (1)$$

This dependent variable is further categorized into three levels: 0 (0 accident severities), 1 (less than 3 accident severities), and 2 (3 or more accident severities). In other words, accident severity is 0 if there were no injuries or fatalities, it takes value 1 if there were less than 3 injuries (or 1 fatality) per accident, and it takes value 2 if there were 3 or more injuries (or 1 fatality) per accident. The frequency of these values is given in Table 2.

The empirical risk is defined as a weighted sum of number of fatalities and the number of injuries. Once again, three injuries were considered an equivalent of one fatality. Consequently, empirical risk is defined as follows:

$$\text{Empirical risk} = (3 \times \text{fatalities}) + \text{injuries} \quad (2)$$

The observed empirical risk frequency is shown in Table 2. In our sample, the empirical risk takes values from 0 to 14.

Table 2. Observed accident frequency, accident severity frequency and empirical risk frequency of $Y=y$

Accident frequency level	Observed frequency	Accident severity level	Observed frequency	Empirical risk level	Observed frequency
$y = 0$	514	$y = 0$	633	$y = 0$	633
$y = 1$	180	$y = 1$	72	$y = 1$	45
$y = 2$	35	$y = 2$	40	$y = 2$	18
$y = 3$	6			$y = 3$	31
$y = 4$	6			$y = 4$	3
$y \geq 5$	4			$y \geq 5$	15

3. Developed models regarding crossing accidents

3.1. Accident frequency model

The review of the prior research for the accident frequency modeling helped us find the most suitable model. Four types of regression models are considered: Poisson, NB, ZIP and ZINB. In our analysis, we included 17 independent variables shown in Table 1. The outcome variable was accident frequency at railway crossings in the period from 2007 to 2011, (Table 2).

For the comparison of two non-nested models, the Vuong test was used (Washington et al., 2003; Miranda-Moreno & Fu, 2006). Results of three Vuong's tests are presented in Table 3.

Table 3. The comparison of models with Vuong's statistic

First model	Second model	Value of $ V $	p value	Better model
Poisson	NB	$ V = 3.60$	$p = 1.70 \cdot 10^{-6}$	NB
NB	ZIP	$ V = 5.84$	$p = 2.56 \cdot 10^{-9}$	ZIP
ZIP	ZINB	$ V = 2.15$	$p = 0.016$	ZIP

The ZIP model was chosen for modeling accident frequencies ($p = 0.016$). The particular ZIP model considered in this study has the following form (Lambert, 1992):

$$\begin{aligned}
 P(Y_i = y_i) &= p_i + (1 - p_i)e^{-\lambda_i} \quad \text{if } y_i = 0 \\
 P(Y_i = y_i) &= (1 - p_i) \frac{e^{-y_i} \lambda_i^{y_i}}{y!} \quad \text{if } y_i = 1, 2, 3, \dots
 \end{aligned} \tag{3}$$

where p_i is the probability of being in the zero state and y is the number of events per period.

The coefficients for the final ZIP model are presented in Table 4. The model was obtained using the function zero-infl from the R-package (Zeileis et al., 2008). Eight independent variables out of 17 were found to be of significance for the model.

The variables regarding the warning device have been shown to be significant. The ZIP model chose two out of four dummy variables, namely road signs (VOSIG) $p = 2.76 \cdot 10^{-9}$ and full gates (VOBR) $p = 1.23 \cdot 10^{-5}$. The half gates variable (VROSP) here acts as a reference variable and the flashing light signal (VOSV) was excluded from the model probably because of lack of enough crossings with this type of device. This means that the probability number of accidents at crossings with road signs, which is the most common type of site protection, is higher than on reference (half gates) crossings. The full gates device has a negative coefficient, which shows us that they are superior to half gates regarding accident prevention. The prevalence of passive control devices may be attributed to a large number of sites with low traffic and train volumes for which the cost of upgrading to automated devices is not justifiable from a cost-benefit analysis in terms of the projected accident reductions at these sites (Ehrlich, 1989). One would suspect that the presence of gates would predictably result in fewer accidents, and the model estimation process did in fact result in a positive effect for the presence of half gate. The gates provide a physical blockade that serves as a deterrent to crossing, but is also cost prohibitive for implementation at all sites. Also, according to Wigglesworth & Uber (1991),

upgrading crossings with flashing light to boom barrier status reduce fatal accidents at crossings. However, many accidents are caused by vehicles running through a crossing to beat a train, occasionally around gates that are deployed (Caird et al., 2002; Cooper & Ragland, 2012). And indeed, automated control devices are not faultless. For example, automated signal control devices are susceptible to false alarms and excessive warning times, which may lead driver to rely on their own hazard judgment and ignore the signal, as well resorting to risky behavior by circumventing the lowered gates (Leibowitz, 1985; Meeker & Barr, 1989).

Table 4. ZIP accident prediction model result

Description	Independent variable	Estimated coefficients	Standard error	z-statistic	P _r (> z)
Model count					
Intercept	Constant	-1.668	0.287	-5.823	5.78e-09 ***
Road signs	VOSIG	0.984	0.166	5.945	2.76e-09 ***
Full gates	VOBR	-1.394	0.319	-4.373	1.23e-05 ***
Crossing width	SIRPPB	-0.530	0.155	-3.416	0.001 ***
Sqrt[AADT · daily trains]	EXPO	0.020	0.005	4.320	1.56e-05 ***
Maximal train speed	MBRZ	0.122	0.028	4.350	1.36e-05 ***
Number of tracks	BRKOLB	-0.370	0.180	-2.053	0.040 *
Crossing surface type	VRKOLA	-0.228	0.137	-1.662	0.096 .
Farm and non-categorized road	KPPNP	0.188	0.147	1.285	0.199
Log (theta)		-1.322	0.167	-7.896	2.87e-15 ***
Model zero					
Intercept	Constant	-1.160	1.626	-0.713	0.476
Road signs	VOSIG	2.012	1.293	1.556	0.120
Full gates	VOBR	-0.580	2.786	-0.208	0.835
Crossing width	SIRPPB	-22.424	1203.8	-0.019	0.985
Sqrt[AADT · daily trains]	EXPO	-0.058	0.060	-0.969	0.333
Maximal train speed	MBRZ	0.078	0.142	0.546	0.585
Number of tracks	BRKOLB	0.462	1.069	0.432	0.665
Crossing surface type	VRKOLA	-3.860	1.233	-3.132	0.002 **
Farm and non-categorized road	KPPNP	3.771	1.135	3.321	0.001 ***

Log-likelihood: -509.4 on 18 Df

AIC: 1054.79

Level of significance: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Two different traffic characteristics proved to be significant in affecting the railway crossing accident frequency. The exposure (EXPO) coefficient is positive ($p = 4.3 \cdot 10^{-7}$). The numbers of trains and road vehicles are often first variables that are considered in developing a model for accident prediction (Austin & Carson, 2002; Saccomanno et al., 2004; Miranda-Moreno et al., 2005). The higher the traffic volume, the more vehicles are exposed to risky situations with incoming trains, which

enlarges the accident probability. According to Austin & Carson (2002), higher numbers of trains and the AADT were found to increase a railway crossing accident frequency.

The maximum train speed at a given crossing (MBRZ) is also associated with a higher predicted accident frequency ($p = 1.36 \cdot 10^{-5}$). This is consistent with Austin & Carson (2002), the higher the defined maximum train speed, the higher the predicted accident frequency. As the train speed increases, the stopping distances of trains extend, and the time available for driver to spot the obstacle and stop the train decreases. On the other hand, difficulties that the drivers of vehicles have, concerning speed and distance of incoming train, are known (Leibowitz, 1985; Meeker et al., 1997; Meeker & Barr 1989; Kasalica et al., 2012), and as the train speed increases, the time the driver has got to react in order to change the wrong decision to cross decreases.

The crossing width, as well as the number of tracks, has been shown to have some significance. This finding is most likely related to the earlier; higher train and traffic volumes require a greater number of tracks and traffic lanes to operate. Road surface also have some significance, but this factor seems inconsequential compared to other railway, road or crossing characteristics likely to affect railway crossing safety.

It should be noted that some crossing characteristics (crossing angle or sight triangle), as well as road category do not have observable influence on the number of accidents.

From this model, we can conclude that the best way in which safety can be improved and the number of accidents can be reduced is upgrading the warning device system. The other variables are either out of our control (maximum train speed and exposure) or do not have significant influence on the number of accidents (road category, road geometry, road surface, etc.).

3.2. Accident severity model

The analysis of accident severity is performed using a multinomial logit model. Multinomial logit models have gained popularity for this type of data mainly because they can account for the dependent variable's ordinal nature.

Let $\pi_j(\mathbf{x}) = P(Y = j; \mathbf{x})$ be the probability of $Y = j, j = 0, 1, 2$. The multinomial logit model is given as follows (Hu et al., 2010):

$$\text{logit}[\pi_j(\mathbf{x})] = \log \frac{\pi_j(\mathbf{x})}{\pi_0(\mathbf{x})} = \alpha_j + \mathbf{x} \beta_j, j = 1, 2. \quad (4)$$

Here α_j is the intercept parameter, and $\beta_j = (\beta_{j\alpha_1}, \beta_{j2}, \dots, \beta_{j17})^T$ is 17- dimensional vector of regression parameters for j – the value of dependent variable. From Eq.(4), taking $\alpha_0 = 0$ and $\beta_0 = 0$, we obtain:

$$\pi_j(\mathbf{x}) = \frac{\exp(\alpha_j + \mathbf{x}\beta_j)}{\sum_{k=0}^2 \exp(\alpha_k + \mathbf{x}\beta_k)}, j = 0, 1, 2 \quad (5)$$

The analyses have been done using the R-function multinom (Venables & Ripley, 2002). Here, we also used the Akaike information criterion (AIC) stepwise procedure. The parameters were estimated using the maximum likelihood estimate (MLE) method. The results were presented using the function mlogit.display (Chongsuvivatwong, 2012). The coefficients for the final model accident severity are presented in Table 5.

Table 5. Multinomial logit model result for accident severity

Independent variable	Severity level (y = 1)		Confidence interval	Severity level (y = 2)		Confidence interval
	Coefficients/SE			Coefficients/SE		
Intercept	-5.76	0.694***	-	-5.55	0.826***	-
VOSIG(x ₁₄)	1.33	0.385***	3.78(1.78,8.04)	0.65	0.428	1.92(0.83,4.44)
VOBR(x ₁₆)	-1.28	0.667	0.28(0.08,1.03)	-1.93	1.051	0.15(0.02,1.14)
SIRPPB(x ₁₁)	1.26	0.297***	3.53(1.97,6.31)	1.20	0.378**	3.33(1.58,6.98)
MBRZ(x ₃)	0.22	0.068**	1.24(1.09,1.42)	0.12	0.083	1.12(0.95,1.32)
EXPO(x ₂)	0.08	0.015***	1.08(1.05,1.11)	0.05	0.017**	1.05(1.02,1.09)
BRKOLB (x ₄)	-0.87	0.427*	0.42(0.18,0.97)	-0.36	0.463	0.70(0.28,1.73)
KATPRM(x ₁)	-0.39	0.311	0.67(0.37,1.24)	0.74	0.430	2.09(0.90,4.87)

Residual Deviance: 668.53
AIC = 700.53

Level of significance: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

These estimated results of the effects of independent variables from this model by the use of the relative risk ratio (RRR) are described in the following section.

The relative risk ratio is the ratio of probabilities of a chosen state and the reference state. In Table 5, their RRR values and 95% confidence intervals (CI) are given. The ratios given are for Y = 1 (e.g., accident severity of 1) and Y = 2 (e.g., accident severity of 2), while Y = 0 (e.g., accident severity of 0). The variables regarding the warning device have been shown to be significant. The multinomial logit model chose two out of four dummy variables, namely road signs (VOSIG) and full gates (VOBR). The half gates variable (VROSP) here acts as a reference variable.

The value for RRR of road signs (VOSIG) for accident severity of Y = 1 is RRR = 3.78, 95% CI (1.78, 8.04), which means that upgrading road signs to reference half gates will result in a lower probability of accident severity of Y = 1. The value for RRR of road signs (VOSIG) for accident severity of Y = 2 is RRR = 1.92, 95% CI (0.83, 4.44), which means that upgrading road signs to reference half gates will also result in lower probability of accident severity of Y = 1. However, since the RRR is smaller in the second case, we can see that the upgrading from road signs to half gates will better prevent less severe accidents than more severe ones. From the RRR values for full gates (VOBR), we can see that upgrading from half gates to full gates will result in lowering probabilities for both accident severities of Y = 1 and Y = 2. However, this is much less significant change than the one when upgrading from road signs to half gates. This finding confirms the design-related issues of this investigation. This is in an agreement with Austin & Carson (2002): "... rather than focus on design-related improvements, one may want to consider improvements in the use of warning devices at railway crossings". However, railway crossings in Serbia are not adequately equipped with modern devices which are standardized, used and are part of the National Railway Level Crossing Safety Strategy in the developed countries. In such circumstances, the safety at railway crossings in Serbia depends mostly on human and physical factors according to Reports of safety and functionality of Serbian Railways (2002-2012). According to Cairney (2003), "the form of traffic control implemented at a railway crossing greatly affects the decision that has to be made by the driver of the road vehicle on the safety of the crossings".

Regarding crossing width variable, the RRR values indicate that wider crossings have higher probabilities for accident severities for both states, RRR =

3.53, 95% CI (1.97, 6.31) for $Y = 1$ and $RRR = 3.33$, 95% CI (1.58, 6.98) for $Y = 2$. It can be noted that this variable has about the same influence on less severe and more severe accidents.

The maximum train speed variable also has a significant influence on accident severity. The RRR value indicates that the ratio of probabilities for $Y = 1$ and $Y = 0$ will increase 1.24 times when the maximum train speed is increased by 10 km/h. The ratio of probabilities for $Y = 2$ and $Y = 0$ is increased 1.12 times when the maximum train speed is increased by 10 km/h. This means that increasing the speed of trains has not a direct impact on mortality.

The exposure to traffic also has a significant influence. For one unit increase of EXPO, the ratio of probabilities is increased $RRR = 1.08$ times ($Y = 1$) and $RRR = 1.05$ times ($Y = 2$). The exposure to traffic has been shown to be an important factor for accident prediction (Ogden, 2002; Austin & Carson, 2007; Hu et al., 2010). A higher number of trains and road vehicles, which is often found in urban areas, are shown in this paper to have an impact on high accident severity. It could be said that crossings with a higher exposure to traffic can provide higher probability of accidents and irregular behavior (Fitzpatrick et al., 1997). Also, according to Hu et al. (2010), one common characteristic found in the railway crossings with more severe accidents is that these railway crossings are usually located in urban areas where traffic exposure is relatively high, and compared to rural areas more traffic accidents are observed in these traffic busy areas.

Regarding number of tracks, the RRR indicates that the probabilities for accident severities are higher when there are multiple tracks as opposed to a single one. As for the last variable from the model, the railway category, it can be seen that the main railways have lower probabilities for accident severity of $Y = 1$, and higher probabilities for accident severity of $Y = 2$.

It can be noted that crossing parameters, such as road and railway geometry (crossing angle and sight triangle), as well as the road type, as it was the case with the ZIP model of accident frequencies, were not accepted by this model.

4. High-risk location analysis

One of the primary tasks in the development of the program for safety improvement in some parts of traffic infrastructure (e.g., road crossings or railway crossings), is to identify the locations that have a high accident risk. This process is also known as black spot identification (Saccomanno et al., 2003; Saccomanno & Lai, 2005). Identifying high-risk locations is the initial step of the process of improving safety (Persaud, 2001). This would then lead to further engineering treatment such as crossing closure or grade separation, improving the crossing geometry or upgrading warning devices to make the crossing safer.

One of the approaches to high-risk identification is based on regression models. This method uses some ranking criteria in order to sort the list of locations and identify the ones with the highest risk. Miranda-Moreno et al. (2009) proposed a Bayesian multinomial model in order to estimate the accident severity for each person involved in an accident. The total risk is defined as the product of accident frequency and its severity (Saccomanno et al., 2004; Miranda-Moreno et al., 2009).

The deaths and injuries are our main concern, so the total risk is the goal we want to achieve. Two criteria are considered for estimating the total risk. The first

criterion is the mean total risk for a crossing obtained as the product of mean accident frequency and mean accident severity, given as follows:

$$mTR_i = mFreq_i * mSev_i \tag{6}$$

where: mTR_i is mean total risk for crossing i ; $mFreq_i$ is mean accident frequency for crossing i obtained from the accident frequency model (Table 4) and $mSev_i$ is mean accident severity for crossing i obtained from the accident severity model (Table 5). The second criterion presented here is based on the empirical risk model. The estimate for total risk for a crossing i would be the mean empirical risk (mER_i). The reason for introducing this additional criterion is that in our data the variables accident frequency and accident severity are slightly correlated ($R = 0.36$).

Similar to what was done for modeling accident frequency; we tried different count data regression models for empirical risk modeling. Four types of models are considered: Poisson, NB, ZIP and ZINB. Those models were obtained using the stepwise AIC. Obtained models were then compared using Vuong's test (Table 6).

Table 6. The comparison of models with Vuong's statistic

First model	Second model	Value of V	p value	Better model
Poisson	NB	V = 4.60	$p = 1.70 \cdot 10^{-6}$	NB
NB	ZINB	V = 4.30	$p = 8.34 \cdot 10^{-6}$	ZINB
ZIP	ZINB	V = 1.98	$p = 0.029$	ZINB

The comparison with Vuong's test showed significant difference ($p = 0.029$) between zero inflated Poisson (ZIP) and zero inflated negative binomial (ZINB), which means that overdispersion is not only caused by excess of zeros, but there is another source of overdispersion. Suppose y is a discrete random variable consisting of the counts on n subjects, y_1, y_2, \dots, y_n . Observations that go into structural zeros ($y_i = 0$) have a degenerate distribution at zero with a probability of occurring is p . While the observations included in the NB counts ($y_i = 0, 1, 2, \dots$) follow a negative binomial distribution with probability of occurring is $(1 - p)$. Therefore, Y is ZINB distributed, which is defined by:

$$Y = \begin{cases} \text{structural zeros, with probability } p \\ \text{counting process, with probability } 1 - p \end{cases} \tag{7}$$

Based on the probability function of the zero-modified distribution, then probability mass function (pmf) for ZINB distribution is (Garay et al., 2015):

$$P_r(Y = y) = \begin{cases} p + (1 - p) \left(\frac{\emptyset}{\mu + \emptyset}\right)^\emptyset, & y = 0 \\ (1 - p) \frac{\Gamma(y + \emptyset)}{\Gamma(y + 1)\Gamma(\emptyset)} \left(\frac{\emptyset}{\mu + \emptyset}\right)^\emptyset \left(\frac{\mu}{\mu + \emptyset}\right)^y, & y = 1, 2, \dots \end{cases} \tag{8}$$

where $(\emptyset)^{-1}$, μ and $\Gamma(\cdot)$ represent a dispersion parameter, mean, and gamma function, respectively.

The estimated parameters of the ZINB model for empirical risk are given in Table 7.

Models for ranking railway crossings for safety improvement

Table 7. ZINB accident prediction model result for empirical risk

Description	Independent variable	Estimated coefficients	Standard error	z - statistic	$P_r(> z)$
Model count					
Intercept	Constant	-1.926	0.667	-2.886	0.004 **
Road signs	VOSIG	0.618	0.314	1.969	0.049 *
Full gates	VOBR	-2.121	0.520	-4.080	4.50e-05 ***
Crossing width	SIRPPB	-0.535	0.324	-1.650	0.099 .
Maximal train speed	MBRZ	0.240	0.066	3.666	2.46e-04***
Sqrt[AADT · daily trains]	EXPO	0.011	0.012	0.928	0.354
Crossing surface type	VRKOLA	-0.665	0.275	-2.416	0.016 *
Log (theta)		-1.322	0.167	-7.896	2.87e-15 ***
Model zero					
Intercept	Constant	9.613	3.863	2.489	0.013 *
Road signs	VOSIG	-5.095	2.013	-2.531	0.011 *
Full gates	VOBR	-5.181	3.118	-1.662	0.011 *
Crossing width	SIRPPB	-7.600	1.789	-4.248	2.15e-05 ***
Maximal train speed	MBRZ	0.829	0.489	1.695	0.090 .
Sqrt[AADT · daily trains]	EXPO	-0.999	0.321	-3.115	0.002 **
Crossing surface type	VRKOLA	2.868	1.531	-1.873	0.061 .
Theta = 0.2666					
Log-likelihood: -509.4 on 18 Df					

The variable that has the highest impact on empirical risk is maximum train speed ($p = 0.000246$). Other variables of importance are road signs ($p = 0.048910$), exposure to traffic ($p = 0.353520$), crossing width ($p = 0.098971$) and road surface type ($p = 0.06110$). Therefore, the probability of the railway crossing being at risk of a fatality varied with these risk factors.

Two lists of railway crossings are compared using two methods, namely percentage deviation and the Spearman correlation coefficient. These two criteria were used in order to create a list of high-risk locations for crossings on the Serbian railway network. Based on each criterion, two lists were made. A simple way to compare the two lists is the percentage deviation. For this purpose, a certain number of top locations from both lists were selected. The percentage deviation is defined in the following way (Miranda-Moreno & Fu, 2006):

$$\% \text{ deviation} = 100 \times (1 - b/m) \quad (9)$$

where b is the number of common locations on the two lists, and m is the number of selected top locations. The percentage deviation is calculated for various thresholds (No. of crossings). The value of deviation is between 40% and 60%. It can be noted that this deviation is greater when there are shorter lists of top locations, and gradually goes down when the length of lists is increased.

The Spearman correlation coefficient is a non-parametric technique to measure the linear correlation between two variables (Miranda-Moreno et al., 2005). Here, the Spearman coefficient is calculated to measure the correlation between these two risk models. In other words, it measures the degree of matching between the two lists. It is calculated in the following way (Miranda-Moreno et al., 2005):

$$r = 1 - \frac{6 \cdot \sum_{i=1}^n d_i^2}{m \cdot (m^2 - 1)} \quad (10)$$

where: r is the Spearman coefficient, d_i is the difference in ranks between the two models for the same crossing i and, m is the number of selected top locations. The correlation is $R = 0.50$.

In Table 8, there is a list of 20 crossings that were identified as high-risk locations that were common for both estimation criteria. For each crossing it is shown how many top locations appear in both lists, as well as the value of mean total risk and mean empirical risk. In this list, we can notice that 60% of crossings are with road signs, and 40% are with half gates. The high-risk location list in Table 8 shows that most crossings, 90%, are located on main railways in urban areas. One of the possible explanations is that urban areas experience greater volume of traffic, which can cause a higher accident rate. On the other hand, the mean maximum train speed at these 20 crossings is 100 km/h, which is significantly higher than the mean maximum speed for the whole sample (70 km/h). This confirms that the maximum train speed has a more pronounced effect on the number of injuries and deaths.

Table 8. List of high-risk locations based on two criteria

First 20 high-risk locations	Crossing No.	Km position	Competent railway station	Warning devices	Mean total risk	Mean empirical risk
10	87	20+993	Batajnica	PB	4.116	2.879
10	28	7+070	Rakovica	DS	2.220	2.495
10	94	34+694	Stara Pazova	PB	1.123	2.502
20	22	253+700	Belotince	DS	1.053	2.108
20	121	74+241	Pirot	DS	1.032	1.990
20	276	252+523	Niš	DS	0.945	2.057
20	298	82+030	Sr. Mitrovica	PB	0.668	4.238
30	27	335+818	Suva Morava	DS	0.901	1.670
30	90	116+080	Šid	PB	0.831	1.573
30	164	57+306	Odžaci	DS	0.651	1.857
30	521	76+983	Ruma	DS	0.637	2.670
30	92	74+019	Voganj	PB	0.581	2.180
30	300	99+549	Sr. Mitrovica	PB	0.581	2.180
30	36	94+920	Velika Plana	DS	0.558	2.110
40	244	119+207	Vrbas	DS	0.854	1.354
40	72	31+037	Kr. Trnovče	DS	0.733	1.310
40	257	78+247	Palanka	PB	0.615	1.294
40	96	97+785	Sr. Mitrovica	DS	0.562	1.298
40	24	79+362	Palanka	PB	0.518	1.254
40	79	26+019	Loz. Saraoci	DS	0.691	1.294

Note: PB = half gates; DS = road signs; a value for a period of five years

It should be noted that many of the predicted high-risk locations were not upgraded during the analysis period, suggesting that possible high-risk crossings as predicted by the model were not considered for safety intervention.

To assess the railway crossing for safety intervention, the Serbian Railways method is based on engineering judgment supplemented by simple statistical analysis of the historical accident data. We examined the original data (first 20 locations according to the highest accident frequency and accident severity). Then, we compared the lists of predicted and historical high-risk locations. Four crossings have been found common in both lists for accident frequency, and two crossings for accident severity. In this paper, it is asserted that high-risk locations cannot be established solely on the basis of historical accident experience. This is supported by (Saccomanno et al., 2004), a longer view of accident risk is needed to reflect the expected risk over a given period of time. Such estimates can be obtained only with accurate and reliable accident frequency and severity prediction models.

5. Conclusion

Ideally, the final outcome of this work would be reducing the number or the severity of accidents at railway crossings in Serbia. In order to achieve this goal, it was important to develop a model that can be applicable to the limited data set at our disposal and to estimate the influence of various parameters contained within the data. We have considered three risk models and two criteria for the identification of high-risk locations.

Data points used in this study, i.e. accident reports and railway crossings' characteristics were extracted from two official data-bases containing actual events and site descriptions. Prior to the modeling, we had analyzed available data in order to ensure that the data sample used was truly representative.

The first considered risk model was the Accident Frequency Model. In this model, we have found that the ZIP regression model produces the best fit for the data used. The second risk model we considered was also well-known in literature – the Accident Severity Model and the Multinomial Logit regression analysis. Both mentioned models provide useful information about the risks involved. However, our needs regarding the risk assessment required some additional quantitative parameters. A novel third model – Empirical Risk Model was introduced in order to satisfy these requirements.

Two criteria for the identification and risk-ranking of the railway crossings in Serbia were presented. One criterion was calculated as a product of mean accident frequency and mean accident severity. The other criterion was obtained using the Empirical Risk Model and we suggest the name mean empirical risk for the name of this quantity.

Because crash frequency and severity jointly determine the casualty risk level at a railway crossing, one can alternatively predict the casualty risk level by using a bivariate count data model. To incorporate the accident severity and some of the key factors such as vehicle occupancy into a total risk model (Miranda-Moreno et al., 2009), the use of posterior distributions through the Bayesian approach (Miranda-Moreno & Fu, 2006; Persaud et al., 1999) has been widely recommended for identification of high-risk locations. Furthermore, an in-depth investigation on vehicle drivers' behavior at individual railway crossings, which is currently being conducted by the authors, might answer the contradictory model estimation results found in this research. These are topics that are worthy of future research.

Appendix

Analysis of the site selection criteria

The sample of 745 railway crossings out of total 2,138 is located throughout the whole network of Serbian Railways. The railway line of major international importance through Serbia (according to European agreement on main international railway lines - AGC) is reference line E70 (Croatia)–Šid–Beograd–Niš–Dimitrovgrad–(Bulgaria). On this reference line, 215 out of total of 329 railway crossings were considered. For other main lines, the numbers of included railway crossings are the following: on reference line E85 (Hungary)–Subotica–Beograd–Preševo–(Macedonia) 39 out of 68; on intermediate line E79 Beograd–Vrbnica–(Montenegro) 11 out of 30; on E66 intermediate line Beograd–Pančevo–Vršac–(Romania) 25 out of 75; on supplementary line Lapovo–Kraljevo–Kosovo Polje 42 out of 78. On other supplementary railway lines, 413 out of remaining 1,558 railway crossings were considered. Therefore, railway crossings of national and local importance, which are located both in urban and rural areas, were included. The sample of 745 railway crossings is composed of 231 (31%) railway crossings on which accidents occurred in a period 2007–2011 according to the Accident database of Serbian railway crossings (2007-2011), and also of 514 (69%) on which accidents did not occur, for which we had adequate data according to the Serbian railway crossing inventory database (2007-2011). With all the data that was at our disposal, we believe that the truncation of railway crossings with incomplete parameters was fairly random, although we cannot exclude some higher order correlations. We believe that our sample is suitable for the purpose of identifying the high-risk crossings in Serbia.

References

- Austin, R. D., & Carson, J. L. (2002). An alternative accident prediction model for highway-rail interfaces. *Accident Analysis and Prevention*, 34(1), 31- 42.
- Caird, J. K., Creaser, J. I., Edwards, C. J., & Dewar, R. E. (2002). A human factors analysis of highway-railway grade crossing accidents in Canada. Report TP 13938E, Canada Transport Development Centre, Transport Canada.
- Cairney, P. (2003). Prospects of improving the conspicuity of trains at passive railway crossings. Report CR 217, Australian Transport Safety Bureau, Canberra.
- Cameron, A. C., & Trivedi, P. K. (1998). *Regression analysis of count data*. Cambridge University Press, Cambridge, UK.
- Chongsuivatwong, V. (2012). *epicalc: Epidemiological calculator*. R package version 2.14.1.6. <http://CRAN.R-project.org/package=epicalc>.
- Cooper, D., & Ragland, D. (2012). Applying safety treatments to rail-highway at grade crossings. Safe TREC. University of California, Berkeley.
- Ehrlich, D. (1989). Driver behavior and decision making. Paper presented at the National Conference on Rail-Highway Safety, San Diego.
- European agreement on main international railway lines (AGC). ECE/TRANS/SC.2/2006/4. Economic and Social Council. United Nations.

- Fitzpatrick, K. M., Carlson, P. J., Bean, J. A., & Bartoskewitz, R. E. (1997). Traffic violation at gated highway-railroad grade crossings. Report TX-98/2987-1, Texas transportation institute, Texas department of transportation.
- Garay, A. M., Lachos, V. H., Bolfarine, H. (2015), Bayesian estimation and case influence diagnostics for the zero-inflated negative binomial regression model. *Journal of Applied Statistics*, 42(6), 1148-1165.
- Hu, Shou-Ren., Li, Chin-Shang., & Lee, Chi-Kang. (2010). Investigation of key factors for accident severity at railroad grade crossings by using a logit model. *Safety Science*, 48(2), 186-194.
- Hu, Shou-Ren., Li, Chin-Shang., & Lee, Chi-Kang. (2011). Assessing Casualty Risk of Railroad-Grade Crossing Crashes Using Zero-Inflated Poisson Models. *Journal of Transportation Engineering*, 137(8), 527-536.
- Kasalica, S., Vukadinović, R., & Lučanin, V. (2012). Study of drivers behaviour at a passive railway crossing. *Promet - Traffic and Transportation*, 24 (3), 193-201.
- Lambert, D. (1992). Zero-inflated Poisson regression with an application to defects in manufacturing. *Technometrics*, 34 (1), 1-14.
- Leibowitz, H. W. (1985). Grade crossing accidents and human factors engineering. *American Scientists*, 95, 558-562.
- Lord, D., Washington, S. P., & Ivan, J. N. (2005). Poisson, Poisson-gamma and zero-inflated regression models of motor vehicle crashes: balancing statistical fit and theory. *Accident Analysis and Prevention*, 37(1), 35-46.
- Meeker, F. L., & Barr, R. A. (1989). An observational study of driver behaviour at a protected railroad grade crossing as trains approach. *Accident Analysis and Prevention*, 21(3), 255-262.
- Meeker, F. L., Fox, D., & Weber, C. (1997). A comparison of driver behavior at railroad grade crossings with two different protection systems. *Accident Analysis and Prevention*, 29(1), 11-16.
- Miaou, S. P. (1994). The relationship between truck accidents and geometric design of road section: Poisson versus negative binomial regressions. *Accident Analysis and Prevention*, 26(4), 471-482.
- Miranda-Moreno, L. F., & Fu, L. (2006). A comparative study of alternative model structures and criteria for ranking locations for safety improvements. *Networks and Spatial Economics*, 6: 97-110.
- Miranda-Moreno, L. F., Fu, L., Saccomanno, F. F., & Labbe, A. (2005). Alternative risk models for ranking locations for safety improvement. *Transportation Research Record*, 1908, 1-8.
- Miranda-Moreno, L. F., Fu, L., Ukkusuri, S., & Lord, D. (2009). How to incorporate accident severity and vehicle occupancy into the hotspot identification process? Paper No. 09-2824 presented at 88th Annual Meeting of the Transportation Research Board, Washington, DC.
- Official Gazette of the Republic of Serbia no. 34/2010; in Serbian. Regulation on personal compensation.
- Ogden, B. D. (2007). *Railroad-highway grade crossing handbook: Revised second edition*. FHWA-SA-07-010, Federal Highway Administration, U.S. DOT, Washington, DC.

Persaud, N.B. (2001). Statistical method in highway safety analysis: a synthesis of highway practice. NCHRP synthesis 295, Transportation Research Board, Washington, DC.

Persaud, N.B., Lyon, C., & Nguyen, T. (1999). Empirical Bayes procedure for ranking sites for safety investigation by potential for safety improvement. Transportation Research Record, 1665. 7-12.

R Development Core Team. (2012). R: A language and environment for statistical computing. Vienna, Austria: R Foundation for Statistical Computing. Retrieved from <http://www.R-project.org/>.

Rovšek, V., Batista, M., & Bogunović, B. (2014). Identifying the key risk factors of traffic accident injury severity on Slovenian roads using a non-parametric classification tree. Transport, 1-10.

Sacomanno, F. F., & Lai, X. (2005). A model for evaluating countermeasures at highway-railway grade crossings. Transportation Research Record, 1918. 18-25.

Sacomanno, F. F., Fu, L., & Miranda-Moreno, L. F. (2004). Risk-based model for identifying highway-rail grade crossings blackspots. Transportation Research Record, 1862. 127-135.

Sacomanno, F. F., Fu, L., Ren, C., & Miranda-Moreno, L. F. (2003). Identifying highway-rail grade crossing black spots. Report T8200-011518/001/MTB, Department of civil engineering University of Waterloo, Ontario, Canada.

Serbian Rail Administration. Accident database of Serbian railway crossings (2007-2011), Serbian Railways; in Serbian.

Serbian Rail Administration. Statistics on accidents on Serbian railways 2011; in Serbian.

Serbian Rail Administration. The Serbian railway crossing inventory database (2007-2011), Serbian Railways; in Serbian.

Serbian Rail Expert Service. Reports of Safety and Functionality of Serbian Railways. Annual reports 2002-2012; in Serbian.

Venables, W. N., & Ripley, B. D. (2002). Modern applied statistics with S. Fourth Edition. Springer, New York. ISBN 0-387-95457-0.

Washington, S. P., Haque, M. M., Oh, J., & Lee, D. (2014) Applying quantile regression for modeling equivalent property damage only crashes to identify accident blackspots. Accident Analysis and Prevention, 66(1), 136-146.

Washington, S. P., Karlaftis, M. G., & Mannering, F. L. (2003). Statistical and econometric methods for transportation data analysis. Chapman and Hall/CRC, Washington, DC.

Wigglesworth, E. C., & Uber, C. B. (1991). An evaluation of the railway level crossing boom barrier program in Victoria Australia. Journal of Safety Research, 22(3), 133-140.

Zeileis, A., Kleiber, C., & Jackman, S. (2008). Regression models for count data in R. Journal of Statistical Software, 27(8), 1-25.

© 2020 by the authors. Submitted for possible open access publication under the terms and conditions of the Creative Commons Attribution (CC BY) license (<http://creativecommons.org/licenses/by/4.0/>).

