Developing accident prediction model for railway level crossings

Ci Liang, Mohamed Ghazel, Olivier Cazier, El Miloudi El Koursi

This paper can be referred to "Ci LIANG, Mohamed GHAZEL, Olivier CAZIER, El-Miloudi EL KOURSI, 2017, Developing accident prediction model for railway level crossings, Safety science, 101, Elsevier, pp.48-59, DOI: 10.1016/j.ssci.2017.08.013"

Developing accident prediction model for railway level crossings

Ci Liang [a,b,c,*], Mohamed Ghazel [b,a,c], Olivier Cazier [d,a], El-Miloudi El-Koursi [b,a,c]

[a] *FCS Railenium, Valenciennes, France*
[b] *IFSTTAR-COSYS / ESTAS, Lille-Villeneuve d'Ascq, France*
[c] *University Lille 1, Lille-Villeneuve d'Ascq, France*
[d] *SNCF Réseau, Paris, France*

Abstract

Railway level crossing (LX) safety continues to be one of the most critical issues for railways, despite an ever-increasing focus on improving design and application practices. Accidents at European LXs account for about one-third of the entire railway accidents and result in more than 300 deaths every year in Europe. Due to the non-deterministic causes, the complex operation background and the lack of thorough statistical analysis based on accident/incident data, the risk assessment of LXs remains a challenging task. In the present paper, some LX accident prediction models are developed. Such models allow for highlighting the influence of the main impacting parameters, i.e., the average daily road traffic, the average daily railway traffic, the annual road accidents, the vertical road profile, the horizontal road alignment, the road width, the crossing length, the railway speed limit and the geographic region. The Ordinary Least-Squares (OLS) and Nonlinear Least-Squares (NLS) methods are employed to estimate the respective coefficients of variables in the prediction models, based on the LX accident/incident data. The validation and comparison process is performed through statistical means to examine how well the estimation of the models fits the reality. The outcomes of validation and comparison attest that the improved accident prediction model has statistic-based approbatory quality. Moreover, the improved accident prediction model combined with the NB distribution shows relatively high predictive accuracy of the probability of accident occurrence.

Keywords: Level crossing safety, Train-car collision, Accident prediction modeling, Statistical analysis;

1. Context and related works

Accidents at railway level crossings (LXs) often give rise to serious material and human damage and hamper railway safety reputation, although the majority of accidents are caused by vehicle driver violations. LX safety is one of the most critical issues for railways which needs to be tackled urgently (Ghazel, 2009; Mekki et al., 2012; Liu et al., 2016). In 2012, there were more than 118,000 LXs in the 28 countries of the European Union (E.U.) which correspond to an average of 5 LXs per 10 line-km (ERA, 2014). Accidents at European LXs account for about one-third of the entire railway accidents. They result in more than 300 deaths every year in Europe (Liu et al., 2016). In some European countries, accidents at LXs account for up to 50% of railway accidents (Ghazel and El-Koursi, 2014; Evans, 2011b). In the entire E.U. zone, the overall number of deaths per fatal accident in railways from 1990 to 2009 is 4.10, with no apparent long-term change over time (Evans, 2011a). In France, the railway network shows more than 18,000 LXs for 30,000 km of railway lines, which are crossed daily by 16 million vehicles on average, and around 13,000 LXs show heavy road and railway traffic (SNCF Réseau, 2011). Despite numerous measures already taken to improve the LX safety, SNCF Réseau (the French national railway infrastructure manager) counted

*Corresponding author. IFSTTAR, Lille-Villeneuve d'Ascq, 20 Rue Élisée Reclus, BP 70317 F-59666 Villeneuve d'Ascq Cedex.
Email: ci.liang@railenium.eu. Tel.: +33(0)320438310.

100 collisions at LXs which led to 25 deaths in 2014. This number was half the total number of collisions per year at LXs a decade ago, but still too large (SNCF Réseau, 2015). In order to significantly reduce the accidents and their related consequences at LXs, it is crucial to establish a high quality accident prediction model and carry out a thorough analysis to understand the potential reasons for accidents occurring at LXs. Indeed, this paves the way for making appropriate safety diagnoses at LXs.

Many existing works dealing with LX safety are devoted to developing qualitative approaches, in order to understand the potential reasons causing accidents at LXs, such as surveys (Wigglesworth, 2001), interviews (Read et al., 2016), focus group methods (Stefanova et al., 2015) or driving simulators (Larue et al., 2015), rather than collecting real field data. In recent years, a systems analysis framework (Leveson, 2011; Read et al., 2016; Wilson, 2014) and a psychological schema theory (Salmon et al., 2013; Stanton and Walker, 2011) have been used to analyze the contributory factors underlying the accidents occurring at LXs. A study presented by Salmon et al. (2013) described a collision between a loaded semi-trailer truck and a train, which occurred in North Victoria, Australia, when the truck crossed the LX while the LX is occupied by railways without lights flashing. According to the investigation of the Office of the Chief Investigator (OCI), the truck driver in this study was not aware of the train and the activated state of the level crossing until it was too late to stop the truck. A study conducted by Davey et al. (2008) discussed the intentional violation of vehicle drivers crossing LXs, particularly focusing on vehicle driver's complacency due to the high level of familiarity. Tey et al. (2011) conducted an experiment to measure vehicle drivers' responses to LXs equipped with stop signs (passive), flashing lights and half barriers with flashing lights (active) respectively. In this study, the vehicle drivers' responses result from both the field survey and a driving simulator. Although these available qualitative approaches are beneficial to understand factors causing LX accidents, they do not allow for predicting the number or the probability of accident occurrence, or quantifying the contribution degree of the various impacting factors. Thereby, quantitative safety analysis approaches are crucial to thoroughly understand the impacting factors and enable the identification of practical design and improvement recommendations to prevent accidents at LXs.

One can notice that a number of quantitative studies on statistical models to predict LX accident frequency open a significant vista on understanding the risk related to LX accidents. In 1941, L. E. Peabody and T. B. Dimmick of the U.S. Bureau of Public Roads developed one of the earliest railway-highway crossing accident prediction models to estimate the number of accidents at railway-highway crossings in 5 years, named Peabody-Dimmick Formula (Ogden, 2007). This formula was developed based on the accident data of rural railway-highway crossings in 29 states in the U.S. and was utilized through the 1950s. As shown in Eq. (1), the parameters considered in this formula are the average daily road traffic V, the average daily railway traffic T, and the protection coefficient indicative of warning devices adopted P. K is an additional parameter.

$$A_5 = \frac{1.28 \times (V^{0.170} \times T^{0.151})}{P^{0.171}} + K \tag{1}$$

However, advances in both warning device technologies and LX design features quickly led to an unavailability of the predefined formula form and coefficients that reflected the conditions pertaining to LX accidents in 1941.

The next evolutionary step in LX accident prediction was the New Hampshire Index (Oh et al., 2006) which is given as follows:

$$HI = V \times T \times P_f \tag{2}$$

where HI represents the hazard index; V is the average daily road traffic; T is the average daily railway traffic and P_f is the protection factor indicative of the warning devices adopted.

The New Hampshire model is a relative formula which can be used to rank the importance of crossing upgrades. Due to its simplicity, it has been widely used across the U.S. However, it is limited in that it does not predict the expected number of collisions, but only gives some indications about the priorities in terms of LX safety.

The accident prediction formula developed by the U.S. Department of Transportation (USDOT) in the early 1980s sought to overcome the limitations of earlier models (Chadwick et al., 2014). This comprehensive formula comprises three primary equations:

$$a = K \times EI \times MT \times DT \times HP \times MS \times HT \times HL \tag{3}$$

$$B = \frac{T_0}{T_0 + T} \times a + \frac{T}{T_0 + T} \times (\frac{N}{T}),\ T_0 = \frac{1}{0.05 + a} \tag{4}$$

$$A = \begin{cases} 0.7159 \times B, \text{ for passive devices;} \\ 0.5292 \times B, \text{ for flashing lights;} \\ 0.4921 \times B, \text{ for gates;} \end{cases} \quad (5)$$

where a is the initial collision prediction (collisions per year at a given LX); K is the formula constant; EI is the exposure index (a variant of traffic moment) based on the product of highway and railway traffic; MT is the index for the number of main tracks; DT is the index for daily through trains during daylight; HP is the index for highway paved; MS is the index for maximum train speed; HT is the index for highway type; HL is the index for highway lanes. B is the adjusted accident frequency; T_0 is the weighting factor and N is the number of accidents observed in T years at a given LX. Finally, A is the normalized accident frequency.

The USDOT formula is the most commonly used model in the U.S. today. A specified table of USDOT provides each of the indexes for LXs equipped with passive controls, flashing lights and gates (Austin and Carson, 2002). Although the formula is comprehensive, its current definition makes it difficult to identify or prioritize design or improvement activities that will most effectively address LX safety-related problems, since it does not provide the magnitude of the characteristics' contribution to the LX safety.

The Australian Level Crossing Assessment Model (ALCAM) is a location specific and parameterized risk model which provides a method for assessing risks to LX users, train passengers and train staff (Woods et al., 2008). The ALCAM model is given as follows:

$$\text{ALCAM Risk Score} = \text{Infrastructure Factor} \times \text{Exposure Factor} \times \text{Consequence Factor} \quad (6)$$

where the Infrastructure Factor is the output of a complex scoring algorithm that assesses how the physical properties at each LX site will affect human behavior; the Exposure Factor is a function of the LX control type, vehicle (or pedestrian) volumes and train volumes (i.e., the Peabody-Dimmick Formula is used as the Exposure Factor function) to address the combined exposure of trains and road vehicles (or pedestrians) pertaining to various LX control types; the Consequence Factor is the expected consequence of a collision which includes deaths and injuries involving both railway and roadway. The Infrastructure Factor adjusts the accident probability per year to reflect the actual LX site conditions. Multiplying the Infrastructure Factor by the Exposure Factor will give the actual annual likelihood of an accident occurring at a particular LX (National ALCAM Committee, 2012). The Consequence Factor is expressed in terms of an expected number of equivalent fatalities per year. An equivalent fatality is a combination of all types of harm using the ratio: 1 fatality = 10 serious injuries = 200 minor injuries. The ALCAM has been applied across all Australian states and in New Zealand since 2003, and overseen by a committee of representatives from the various jurisdictions of these countries to ensure its consistency in terms of development and application. However, the ALCAM does not cover all kinds of LX accidents, since its main focus is deliberate and accidental collisions involving user errors but excluding vandalism and suicide. It should be noticed that some LX physical properties considered in ALCAM show a high correlation between each other, which implies the existence of a kind of redundancy between the model inputs, and consequently a bias in terms of the outputs.

In recent studies, authors tended to adopt the Poisson regression model, the NB regression model or variants of the Poisson regression model (e.g., zero-inflated Poisson (ZIP) and zero-inflated negative binomial (ZINB)) combined with the estimated $\hat{\lambda} = e^{\sum_{j=1}^{m} \beta_j x_j + \sigma}$ (x_j is the independent variable considered and β_j is the estimated coefficient of x_j) (Cameron and Trivedi, 1986; Lawless, 1987; Cameron and Trivedi, 1990; Miaou, 1994; Austin and Carson, 2002; Chang, 2005; Lu and Tolliver, 2016) to deal with accident statistics. However, this form of estimated $\hat{\lambda}$ is not appropriate in our case. According to the constraints between the LX accident frequency and impacting variables, presented in section 3.2, some variables (e.g., the average daily railway traffic, the average daily road traffic and the road traffic accidents) should not be used in an exponential form, due to the logical assumption that the case where these variables are equal to 0, would directly lead to 0 accident occurrence. Therefore, these aforementioned approaches combined with traditional $\hat{\lambda}$ will introduce high bias when predicting the LX accident frequency and the probability of accident occurrence.

These aforementioned investigations indicate a strong need for an appropriate accident prediction model that is comprehensive in its consideration of contributing factors to LX safety. More importantly, such a model should have good statistical quality and relatively high predictive accuracy. Therefore, in the present study, a new accident prediction model and an improved model based on the new model are developed to predict the accident frequency at

(a) SAL4 (b) SAL2

(c) SAL0 (d) Crossbuck LX

Fig. 1. Four types of LXs in France.

LXs. The current paper begins with a description of the research context and a review of previous literature on LX safety analysis. The coefficients associated with the parameters of our models are estimated based on the French LX accident/incident data provided by SNCF Réseau. A thorough statistical analysis for examining the model quality and a comparison between predictive accuracies of the two models combined respectively with the Poisson distribution and the negative binomial (NB) distribution are then performed. Moreover, the contributions of various parameters considered to LX accident occurrence are discussed thoroughly. This paper concludes with a summary of the present study and directions for future research.

2. Study subject

There are four LX types in France (SNCF, 2015), as shown in Fig. 1:
a) SAL4: Automated LXs with four half barriers and flashing lights;
b) SAL2: Automated LXs with two half barriers and flashing lights;
c) SAL0: Automated LXs with flashing lights but without barriers;
d) Crossbuck LXs, without automatic signaling.

As shown in Table 1, SAL2 (more than 10,000) is the most widely used type of LX in France. Moreover, more than 4,000 accidents at SAL2 LXs contributed most to the total number of accidents at LXs from 1974 to 2014.

Table 1. Accidents at different types of LXs in France from 1974 to 2014.

Type of LX	Number	# Accident
SAL4	> 600	> 600
SAL2	> 10,000	> 4,000
SAL0	> 60	> 50
Crossbuck LX	> 3,000	> 700

LX accidents are caused by the following transport modes: 1) motorized vehicle (MV), 2) pedestrian and bicycle

(PB). As illustrated in Fig. 2, the motorized vehicle is the main transport mode causing LX accidents in the 21 geographical regions in France. Moreover, as the LX accident frequency caused by motorized vehicles increases, the entire LX accident frequency increases accordingly. On the contrary, pedestrians and cyclists contribute very little to the overall risk[1] related to LX accidents (Liang et al., 2017).

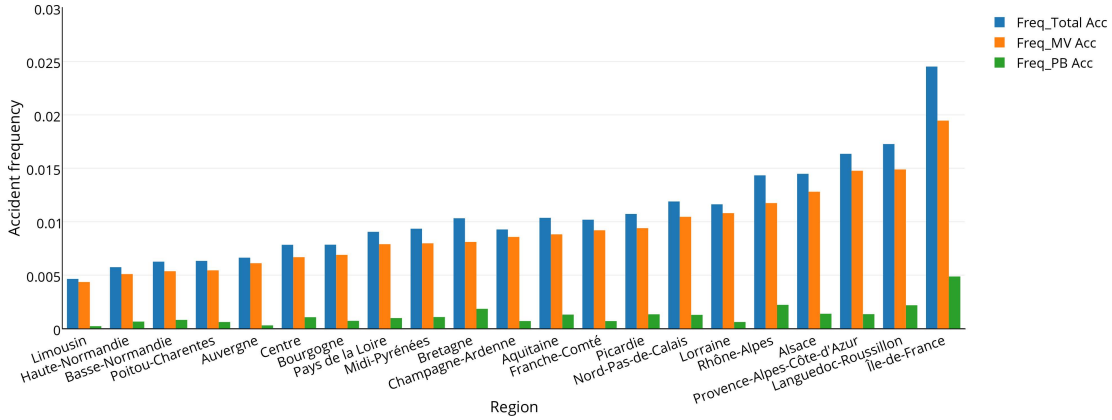

Fig. 2. Comprehensive accident frequency in different regions.

Considering the train/motorized vehicle (train-MV) collisions, SAL2 LXs also have the most part of LX accidents according to the statistics shown in Fig. 3. For all these reasons, we will limit the scope of our analysis to train-MV accidents occurring at SAL2 LXs; in fact, from the aforementioned observation, these accidents can be considered as the most representative for LX accidents in general.

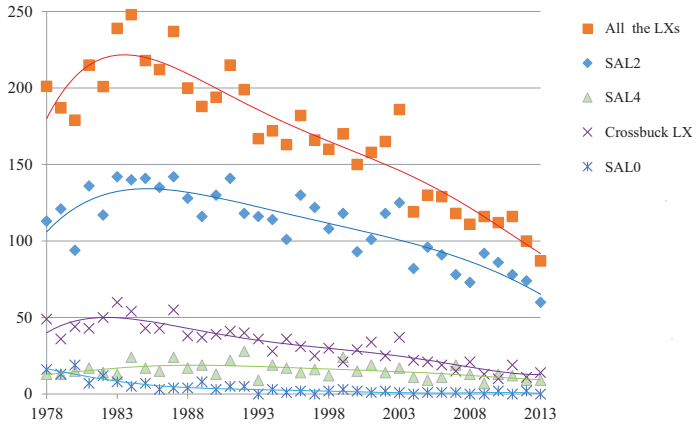

Fig. 3. The number of collisions (train-MV) at different types of LX in France from 1978 to 2013.

3. Methodology

3.1. Data sources and coding

The data to support our investigation come from a dedicated accident/incident database provided by SNCF Réseau. SNCF Réseau investigated and recorded various attributes of LX accidents/incidents, railway and roadway traffic

[1] In this paper, most of the time the term "risk" stands for the occurrence likelihood.

characteristics, surrounding characteristics of LXs. Therefore, the accident/incident data and involved LX information that cover SAL2 LXs in 21 geographical administrative regions in mainland France from 1990 to 2013 are obtained.

In the present study, an adequate sample is selected to include the data in the decade from 2004 to 2013, which provides reliable and sufficient information about both LX accidents and railway, roadway and LX characteristics. Namely, the selected LX inventory presents the LX identification number, the LX location, the LX accident timestamp, the railway traffic volume, the road traffic volume, the LX dimension, the profile and alignment of the entered road and so on. There are 8,332 public SAL2 LXs involved in our investigation. The total number of SAL2 LXs in France is about 10,000. However, in our investigation we considered only those about which we had enough information (8,332). Using the LX identification number and the LX accident timestamp in the accident/incident database, the annual accident frequency at a given SAL2 is obtained. Then, a new database containing 10 years of data is created, using again the LX identification number as a common data element, which includes annual LX accident frequency, railway, roadway and LX characteristics at a given SAL2 and annual roadway accident statistics. Impacting parameters pertaining to LX accidents considered in our investigation should be thought to be: (1) important in determining accident frequency, (2) more permanent in nature (e.g., sight obstruction noted as a problematic factor due to involved alterable construction topography, vegetation and other environmental elements) and (3) not accident-dependent (Austin and Carson, 2002). This combined database formed the basis of our investigation. The parameters considered in this investigation are shown in Table 2. As shown in Table 2, some minor data transformations in the combined database were necessary. Variables that have multiple non-numeric choices (e.g., profile, alignment) are encoded as singular indicator variables. Numerical variables, such as the average daily road traffic, the average daily railway traffic, the railway speed limit, the LX width and the crossing length are used as they are without transformation. The region risk factor is determined by the general accident frequency per SAL2 in the region. The road accident factor is determined by the ratio of the annual number of road accidents in a given year to the average number of road accidents per year over the period of 10 years considered. The statistical characterization of the variables considered is given in Table 3.

It is worth noticing that by using the Spearman correlation checking (Borkowf, 2002), we found that some other parameters tested were not significant (e.g., the road-rail track angle at a given LX) or highly correlated with the parameters considered in our analysis (e.g., the number of lanes at a given LX is highly correlated with the LX width).

3.2. Preliminary accident prediction model

Based on some preliminary analyses, it is worth noticing that five constraints need to be considered so as to develop the model for predicting annual accident frequency at a given SAL2:

- The predicted accident frequency should always be non-negative;
- It should be 0 if the average daily railway traffic is 0;
- It should be 0 if the average daily road traffic is 0;
- It should be 0 if the annual road traffic accidents are 0;
- The model should be time-dependent, i.e., it should reflect the variation of accident frequency as time advances.

For the preliminary accident prediction model, we considered only three parameters in Table 2, which are the average daily railway traffic, the average daily road traffic and the annual road accidents. The preliminary model is developed as follows:

$$\lambda_{10P} = K \times F_{RAcc} \times V^a \times T^b \tag{7}$$

where λ_{10P} represents the annual accident frequency at a given SAL2 during the period of 10 years considered; K is the constant coefficient; F_{RAcc} is the road accident factor; V is the average daily road traffic and T is the average daily railway traffic. Here, F_{RAcc} is a time-dependent variable which can reflect the variation of annual road accidents as time advances.

The conventional formula of the traffic moment is given by: Traffic moment = Road traffic frequency × Railway traffic frequency (Liang et al., 2017). However, based on some previous analyses, we adopt a variant called "corrected moment", or CM for short. $CM = V^a \times T^b$, where $b = 1 - a$ and the best value of a in terms of fitting is computed to be $a = 0.354$ based on the previous statistical analysis performed by SNCF Réseau (SNCF Réseau, 2010). Therefore,

Table 2. Parameters considered and data coding.

Parameter	Explanation	Data coding
Railway traffic characteristics		
Average daily railway traffic	The average number of trains crossing the LX daily	Numerical, used directly;
Railway speed limit	The maximum permission speed of train within the LX section	Numerical, used directly;
Roadway traffic characteristics		
Average daily road traffic	The average number of road vehicles crossing the LX daily	Numerical, used directly;
Annual road accidents	The number of road accidents in a given year	Road accident factor: *Annual road accidents in a given year / Average road accidents per year over the period observed*;
LX characteristics		
Alignment	Horizontal road alignment shape: "straight", "curve" or "S"	Alignment indicator; 0, 1 and 2 represent "straight", "curve" and "S", respectively;
Profile	Vertical road profile shape: "normal" or "hump or cavity"	Profile indicator; 0 and 1 represent "normal" and "hump or cavity", respectively;
LX width	The entered road width	Numerical, used directly;
Crossing length	The length of LX that road vehicles need to cross	Numerical, used directly;
Region risk	The region of the LX considered	Region risk factor, highlighting the general LX-accident-prone region: *The number of SAL2 accidents over the observation period in the region considered / The number of SAL2 LXs in the region considered*;

Table 3. Statistical characterization of variables considered.

Variable	Mean	Variance	StdDev	Min	Max
Annual LX accident	0.0057	0.0060	0.0776	0	2
Average daily railway traffic	26.0636	914.5413	30.2413	0.5000	330
Railway speed limit	92.4599	1.7963e+03	42.3829	5	160
Average daily road traffic	826.8022	3.1718e+06	1.7810e+03	0.5700	2.5570e+04
Corrected moment	51.4744	3.7377e+03	61.1367	1.2781	938.5449
Road accident factor	1.0001	0.0189	0.1378	0.8058	1.1988
Alignment	0.2587	0.3209	0.5665	0	2
Profile	0.1488	0.1266	0.3559	0	1
Length	9.6766	14.9545	3.8671	3	59
Width	5.4504	1.8414	1.3569	2	24
Region risk factor	0.3487	0.0142	0.1194	0.1739	0.7747

we consider $(V^{0.354} \times T^{0.646})$ as an integrated parameter that reflects the combined exposure frequency of both railway and road traffic. One can notice that Eq. (7) can be rewritten as $\lambda_{10P} = K \times RM$, where $RM = F_{RAcc} \times V^{0.354} \times T^{0.646}$. Thus, this model can be regarded as a linear model with respect to the composite parameter RM. The Ordinary Least-Squares (OLS) method is employed to estimate coefficient K. As shown in Fig. 4, K is estimated as 1.319e-04 ($t - statistic = 33.72 > 1.96$ corresponding to a 95% confidence level).

Fig. 4 indicates that this preliminary model shows that, for high values of corrected moment, there is a significant deviation between observed accident frequencies and predicted accident frequencies at SAL2 LXs. Therefore, further statistical analysis is carried out to evaluate the quality of the transformed linear model. In this case, we make group

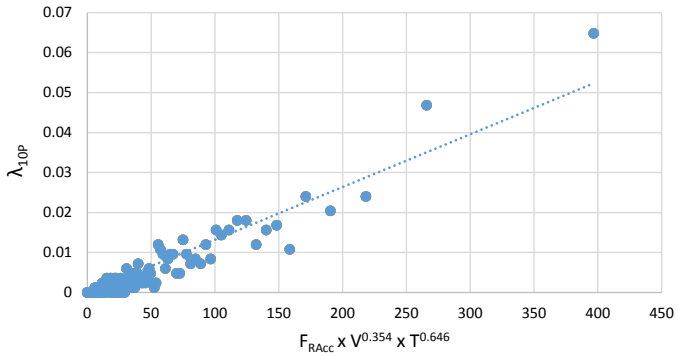

Fig. 4. The preliminary accident frequency prediction model λ_{10P} vs. $(F_{RAcc} \times V^{0.354} \times T^{0.646})$.

classification, which means that the data set is divided into 100 groups with the same number of samples in each group. Then, the mean value of λ_{10P} and RM of each group are computed respectively to generate a linear relationship between the group-mean λ_{10P} and the group-mean RM. Hence, we can adopt "Residuals vs. Fitted" graph, "Normal Q-Q" graph, "Scale-Location" graph and "Residual vs. Leverage" graph (Anscombe, 1973; Chatterjee and Hadi, 2015) to check the linearity, the normal distribution of residuals, the homoscedasticity (Jarque and Bera, 1980) and the abnormal values of the model, respectively. These four graphs pertaining to our group classification analysis are shown in Fig. 5. For an idealized linear model: 1) the red line in Fig. 5a should be a horizontal line at 0 and residuals should randomly distribute around this line; 2) the standardized residuals shown in Fig. 5b should fall in the 45° direct line, which can attest the normal distribution of residuals; 3) the red line shown in Fig. 5c should be a horizontal line at a certain value and square roots of standardized residuals should randomly distribute around this line; 4) Fig. 5d can identify abnormal values and significant values which have an important impact on the model fitting, through Cook's distance. One can notice that residuals in the groups with big ID (i.e., 95, 99 and 100) are significant. These residuals correspond to SAL2 LXs with high corrected moment in our data set. Therefore, the statistical test results attest the significant deviation between the observed accident frequencies and the predicted accident frequencies at SAL2 LXs having high corrected moment. Through checking the accident/incident data, these SAL2 LXs with high corrected moment are correspondingly accident-prone LXs in general. We conjecture that this preliminary accident prediction model is not appropriate to predict the annual accident frequency at SAL2 LXs with high corrected moment. A thorough analysis needs to be performed to develop an improved model which can predict the annual accident frequency at a given SAL2 more accurately. Meanwhile, the model should consider more impacting variables.

3.3. Improved accident prediction model

Based on the previous analysis in section 3.2, we consequently developed an improved version of the prediction model, as shown below:

$$\lambda_{10Y} = K \times F_{RAcc} \times (V^{0.354} \times T^{0.646}) \times e^{(C_{Profile} \times I_{Profile} + C_{Align} \times I_{Align} + C_{Wid} \times Wid + C_{Leng} \times Leng + C_{RSL} \times RSL + C_{Reg} \times F_{Reg})} \quad (8)$$

where λ_{10Y} represents the annual accident frequency at a given SAL2 for a period of 10 years; K is the constant coefficient; F_{RAcc} is the road accident factor; V is the average daily road traffic; T is the average daily railway traffic and $V^{0.354} \times T^{0.646}$ is regarded as the corrected moment; $I_{Profile}$ and $C_{Profile}$ are respectively the profile indicator and its corresponding coefficient; I_{Align} and C_{Align} are respectively the alignment indicator and its corresponding coefficient; Wid and C_{Wid} are respectively the LX width and its corresponding coefficient; $Leng$ and C_{Leng} are respectively the crossing length and its corresponding coefficient; RSL and C_{RSL} are respectively the railway speed limit and its corresponding coefficient; F_{Reg} and C_{Reg} are respectively the region factor and its corresponding coefficient (see Table 2).

Note that appropriate higher orders and interaction terms of covariates can be included in Eq. (8) without difficulty, due to the use of exponential form (Miaou, 1994). The Nonlinear Least-Squares (NLS) method and Gauss-Newton

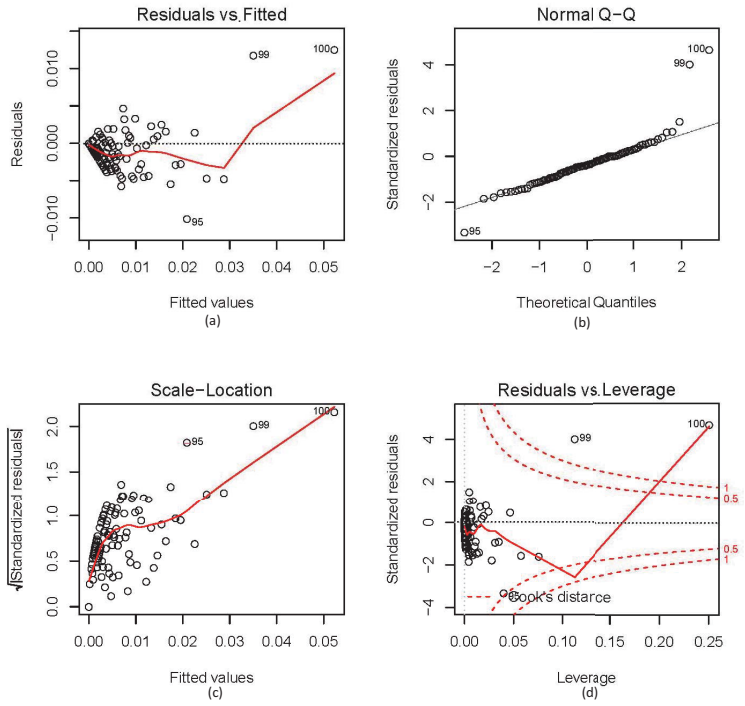

Fig. 5. Statistical evaluation of the model quality: (a) Residuals vs. Fitted, (b) Normal Q-Q, (c) Scale-Location, (d) Residual vs. Leverage.

algorithm (Madsen et al., 2004) are adopted to estimate the coefficients of variables. Estimated coefficients are provided in Table 4. A $|t - statistic| > 1.96$ is introduced to identify significant parameters corresponding to a 95% confidence level. The average daily railway traffic, the railway speed limit, the average daily road traffic, the annual road accidents, the road alignment, the LX width, the crossing length and the region LX-accident-prone factor have significant and positive influence on SAL2 accident frequency. However, the test shows that the road profile does not have a significant impact. The $t - statistic$ for road profile is not significant which indicates that the coefficient of road profile $C_{Profile}$ should be 0 and the impact of road profile could be neglected (The interpretation of this fact is discussed in section 5.1.3). Moreover, the coefficients of the considered variables with the exponential form can reflect their respective contribution degrees to the SAL2 accident frequency. According to these contribution degrees (ranked in brackets), the region LX-accident-prone factor is the most important contributor among these variables.

Table 4. Estimated coefficients of the improved accident prediction model.

Parameter	Coefficient	Estimated value	Standard error	$t - statistic$	Significant
	K	2.703e-05	5.078e-06	5.322	×
$I_{Profile}$	$C_{Profile}$	3.626e-02	5.706e-02	0.635	
I_{Align}	C_{Align}	3.427e-01 (2)	2.942e-02	11.648	×
Wid	C_{Wid}	9.847e-02 (3)	1.494e-02	6.589	×
$Leng$	C_{Leng}	2.084e-02 (4)	4.284e-03	4.865	×
RSL	C_{RSL}	3.089e-03 (5)	7.586e-04	4.072	×
F_{Reg}	C_{Reg}	4.962e-01 (1)	1.722e-01	2.882	×

4. Model validation and distribution identification

In this section, we will validate the quality of the two prediction models and identify an appropriate statistical distribution combined with the prediction model of accident frequency, in such a way as to make a more accurate estimation of the probability of accidents occurring at a given SAL2 in a given year.

4.1. Statistical test evaluation

We applied the two prediction models to estimate the annual accident frequency based on the 10-year accident data of SAL2 LXs. The Monte-Carlo test for randomly sampling annual accident frequencies which meet the condition that the predicted annual accident frequency at a given SAL2 is equal to or more than the observed annual accident frequency at the SAL2 considered is performed (Considering a safety strict principle, the predicted annual accident frequency should not be lower than the observed annual accident frequency). Then, the percentages of randomly sampled annual accident frequencies that meet this condition are computed to compare with the actual percentages of specified entire sampled annual accident frequencies (e.g., as for the entire 80,000 annual accident frequencies sampled out of 83,320, the actual entire percentage is computed as 80,000/83,320; while k annual accident frequencies within the 80,000 frequencies sampled meet the above condition, thus, the percentage of randomly sampled annual accident frequencies meeting this condition is computed as k/83,320). Table 5 shows the Monte-Carlo test results. One can notice that, for the specified entire random sampling size 80,000, 40,000, 10,000, 5,000 and 500, the percentages of randomly sampled annual accident frequencies meeting the aforementioned condition computed using λ_{10Y} are all closer to the actual percentages of specified entire sampled annual accident frequencies, compared with the percentages of randomly sampled annual accident frequencies computed using λ_{10P}. Moreover, the similarity between the percentages of randomly sampled annual accident frequencies meeting the aforementioned condition, which are computed using λ_{10Y}, and the actual specified entire percentages is relatively high.

Although the Monte-Carlo test results indicate that the λ_{10Y} model seems more appropriate, the tested percentages of annual accident frequencies sampled according to the aforementioned condition closer to the actual percentages are not able to thoroughly attest to the fact that the quality of λ_{10Y} model is definitely better, since the predicted accident frequency may be much higher than the accident frequency observed. Therefore, further statistical tests are required to comprehensively evaluate the model quality.

Table 5. Monte-Carlo test results.

# Samples	Actual percentage of annual accident frequencies sampled	λ_{10Y}-model estimated percentage of annual accident frequencies sampled	λ_{10P}-model estimated percentage of annual accident frequencies sampled
80,000	0.96015	0.95482	0.94191
40,000	0.48008	0.47747	0.45463
10,000	0.12002	0.11946	0.11416
5,000	0.06001	0.05959	0.05665
500	0.00600	0.00598	0.00576

Akaike's information criterion (AIC) (Bozdogan, 1987), the Bayesian information criterion (BIC) (Weakliem, 1999), the Pearson chi-square statistic test (PCS) (Dahiya and Gurland, 1972) and the degree of freedom (DF) are computed to evaluate the goodness of fit (GOF) of the model. They can be expressed as follows:

$$\mathrm{AIC} = n + n \times ln(2\pi) + n \times ln(\mathrm{RSS}/n) + 2(l+1) \tag{9}$$

$$\mathrm{BIC} = n + n \times ln(2\pi) + n \times ln(\mathrm{RSS}/n) + (l+1)ln(n) \tag{10}$$

$$\mathrm{PCS} = \sum_{i=1}^{n} \frac{(O_i - \lambda_i)^2}{\lambda_i} \tag{11}$$

$$\mathrm{DF} = n - (l + 1) \tag{12}$$

where n is the sample size; RSS is the sum of the squares of residuals between the annual accident frequencies observed and the annual accident frequencies estimated; l is the number of independent exponential parameters; O_i is the annual accident frequency observed and λ_i is the annual accident frequency expected.

The AIC and BIC are two statistical measures to test the relative quality of models for a given set of data. Smaller AIC and BIC values indicate a better model fitting. The PCS test is used to determine whether there is a significant difference between the values expected and the values observed. The PCS is roughly equal to the model DF if the model fits the data perfectly without any dispersion. In other words, the closer the PCS is to the DF, the better the model fits the data (Lu and Tolliver, 2016). These statistical test results are shown in Table 6 with the goodness ranked in brackets. Some findings can be noticed: 1) considering AIC and BIC, the λ_{10Y} model gives better results, since the AIC and BIC values corresponding to the λ_{10Y} model are much smaller than those for the λ_{10P} model; 2) as for PCS, the λ_{10Y} model is also the preferred one, since the PCS of the λ_{10Y} model is closer to DF (DFs of the λ_{10Y} and the λ_{10P} are considerably approximative).

4.2. Statistical distribution identification

In this section, further analysis is performed to identify a more appropriate statistic distribution combined with the accident frequency prediction model. The Poisson distribution shown as Eq. (13) is a natural first choice for modeling such accident occurrence. Chang (2005) indicates that accident frequency is likely to be over-dispersed (cf. Eq. (14)) and suggests using the negative binomial (NB) distribution as an alternative.

$$Poi(X = k) = \frac{\lambda^k e^{-\lambda}}{k!},\ k = 0, 1, 2, \ldots \tag{13}$$

where $Poi(X = k)$ is the probability of k accidents occurring, $k \in IN$ and λ is the expectation value of the number of accidents. In our study, λ is expressed by Eq. (7) or Eq. (8).

$$VAR(X)\begin{cases} = E(X) \\ > E(X), \text{ over-dispersed} \\ < E(X), \text{ under-dispersed} \end{cases} \tag{14}$$

The NB model as a special case of Poisson-Gamma mixture model is a variant of the Poisson model designed to deal with over-dispersed data (Lord and Mannering, 2010; Buddhavarapu et al., 2016). The over-dispersion could come from several possible sources, e.g., omitted variables, uncertainty in exposure data, covariates or non-homogeneous LX environment (Miaou, 1994). The NB model considered in this study has the following form:

$$P_{NB}(X = k) = \frac{\Gamma\left(k + \frac{1}{\alpha}\right)}{\Gamma(k+1)\,\Gamma\left(\frac{1}{\alpha}\right)} \left(\frac{1}{1+\alpha\lambda}\right)^{1/\alpha} \left(\frac{\alpha\lambda}{1+\alpha\lambda}\right)^{k},\ k = 0, 1, 2, \ldots \tag{15}$$

where $P_{NB}(X = k)$ is the probability of k accidents occurring, $k \in IN$; λ is the expectation value of the number of accidents and α is the dispersion parameter.

The relationship between the mean value and the variance in the NB model is given as follows:

$$VAR(X) = E(X) + \alpha E(X)^2 \tag{16}$$

If $\alpha > 0$, there is an over-dispersion; if $\alpha < 0$, there is an under-dispersion and, in the case where $\alpha = 0$, the NB model reduces to the Poisson model.

However, the NB model is limited to handling under-dispersed data ($\alpha < 0$) (Lord and Mannering, 2010). If the dispersion parameter α is set as a negative value to try to handle under-dispersion issue, it would no longer be an NB model and would lead to unreliable estimation, especially when the sample mean is low and the sample size is small

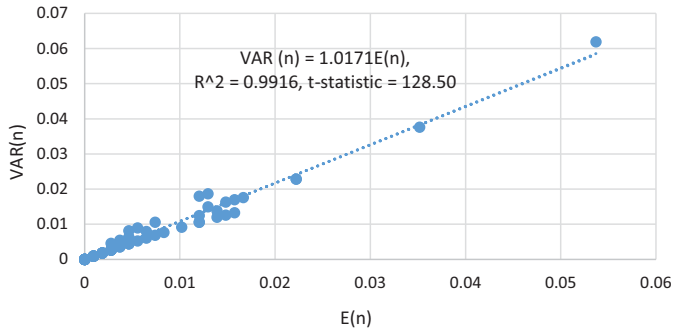

Fig. 6. Constraint between the group variance and the group mean of annual accidents for all SAL2 LXs.

(Lu and Tolliver, 2016). Oh et al. (2006) proposed the Gamma model to handle under-dispersed samples. The Gamma model is given as follows:

$$P_G(X = k) = Gamma(\beta k, \lambda) - Gamma(\beta(k+1), \lambda) \tag{17}$$

where $P_G(X = k)$ is the probability of k accidents occurring, $k \in IN$; λ is the expectation value of the number of accidents and β is the dispersion parameter. If $\beta < 1$, there is an over-dispersion; if $\beta > 1$, there is an under-dispersion and if $\beta = 1$, the Gamma model reduces to the Poisson model. However, the Gamma model shown as Eq. (18) is limited to the time-dependent observation assumption and zero observations (Lord and Mannering, 2010), since general $\Gamma(x)$ restricts discrete responses to positive values.

$$Gamma(\beta k, \lambda) = \begin{cases} 1, \text{if } k = 0 \\ \dfrac{1}{\Gamma(\beta k) \int_0^{\lambda} u^{\beta k - 1} e^{-u} du}, \text{if } k > 0 \end{cases} \tag{18}$$

Therefore, the restriction between variance and mean value is significant to identify an appropriate statistical distribution. Firstly, we adopted group classification to make preliminary variance analysis, which is that the annual accidents at a given SAL2 during the 10 years were divided into 100 groups with the same number of samples in each group. Then, the mean value and variance of accidents in each group were computed respectively to analyze the relationship between the group variance and the group mean value. The variance analysis is shown in Fig. 6. It seems that there is a slight over-dispersion of the data set, since the variance $VAR(n)$ is a bit bigger than the mean $E(n)$ ($VAR(n) = 1.0171E(n)$).

Since the mean value and the variance are very close to each other, we performed meticulous analyses to assess both the Poisson and the NB models with regard to all SAL2 LXs in our accident database so as to identify which model is more effective. When applying the NB distribution, we adopt the Maximum Likelihood Estimation (MLE) method to estimate the dispersion parameter α of the data set (Dai et al., 2013). Using R language to perform the MLE method, α is estimated at 1.9594.

The log-likelihood statistic (LL) is adopted to assess the GOF of the accident frequency prediction model combined with a statistical distribution and identify the more appropriate distribution for estimating the probabilities of accident frequency observed. The larger the LL, the more preferred the model (Lu and Tolliver, 2016). The mathematical description of the LL is given as:

$$\text{LL} = \sum_{i=1}^{n} ln(\hat{P}_i) \tag{19}$$

where n is the sample size and $\hat{P}_i$ is the estimated probability of accident frequency observed. $\hat{P}_i$ is computed respectively according to the accident frequency prediction model combined with the Poisson or the NB distribution.

LL test results are shown in Table 6. One can notice that for λ_{10Y} model combined with either the Poisson or NB distribution, its GOFs are significantly better than λ_{10P} model's GOFs according to LL results. Furthermore, the GOF of λ_{10Y} combined with the NB distribution is better than when combined with the Poisson distribution.

Table 6. Model GOF comparison.

Parameter	λ_{10Y} **Poisson**	λ_{10Y} **NB**	λ_{10P} **Poisson**	λ_{10P} **NB**
Railway traffic characteristics				
Average daily railway traffic	×	×	×	×
Railway speed limit	×	×		
Roadway traffic characteristics				
Average daily road traffic	×	×	×	×
Annual road accidents	×	×	×	×
LX characteristics				
Alignment	×	×		
Profile	×	×		
LX width	×	×		
Crossing length	×	×		
Region	×	×		
AIC	-190,744 (1)	-190,744 (1)	-190,591 (2)	-190,591 (2)
BIC	-190,670 (1)	-190,670 (1)	-190,573 (2)	-190,573 (2)
PCS	65,796 (1)	65,796 (1)	53,108 (2)	53,108 (2)
DF	83,313	83,313	83,319	83,319
LL	-2,599 (2)	-2,596 (1)	-2,631 (4)	-2,629 (3)
Goodness score (the lower, the better)	5	4	10	9

Based on the predicted probability of the accident frequency observed, further Cumulative Distribution Function (CDF) analysis with regard to the Poisson and the NB distributions is performed to evaluate the quality of the accident frequency prediction model combined with these two statistical distributions. As shown in Fig. 7, the relationship between the CDF and the corresponding probability of a given event is depicted. $\hat{P}(\bullet)$ denotes the predicted probability of a given event obtained through the Poisson or NB distribution; O_i is the observed accident frequency and λ_i is the estimated accident frequency. The blue curve "CDF NB λ_{10P}, $O_i > \lambda_i$" represents the CDF of event " $O_i > \lambda_i$" obtained through the NB distribution combined with the λ_{10P}; the red curve "CDF NB λ_{10P}, $O_i <= \lambda_i$" represents the CDF of event " $O_i <= \lambda_i$" obtained through the NB distribution combined with the λ_{10P}; the green curve "CDF POI λ_{10P}, $O_i > \lambda_i$" represents the CDF of event " $O_i > \lambda_i$" obtained through the Poisson distribution combined with the λ_{10P}; the violet curve "CDF POI λ_{10P}, $O_i <= \lambda_i$" represents the CDF of event " $O_i <= \lambda_i$" obtained through the Poisson distribution combined with the λ_{10P}. The interpretation of the remaining curves involving the λ_{10Y} can be similarly obtained. Given that some curves are almost covered by some others in Fig. 7, the extracted results of CDF analysis shown in Table 7 become clearer for discussion.

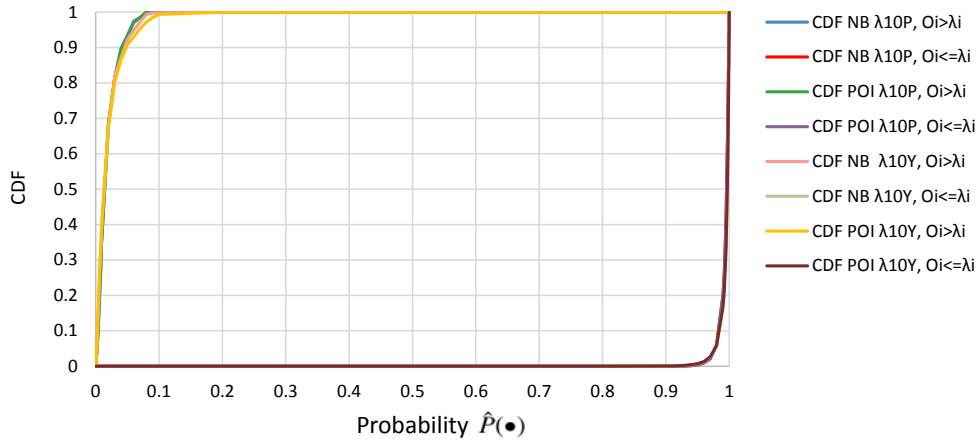

Fig. 7. CDF of the Poisson and the NB distributions combined with the λ_{10P} and λ_{10Y} models according to the estimated probability.

Table 7. The extracted results of CDF analysis.

Model CDF	$\hat{P}(O_i > \lambda_i) > 0.005$ (CDF in percent)	$\hat{P}(O_i > \lambda_i) > 0.05$ (CDF in percent)	$\hat{P}(O_i <= \lambda_i) > 0.95$ (CDF in percent)	$\hat{P}(O_i <= \lambda_i) > 0.995$ (CDF in percent)
CDF NB λ_{10P}	85.29 (3)	6.61 (1)	99.62 (1)	57.19 (3)
CDF NB λ_{10Y}	79.10 (2)	7.68 (3)	99.36 (3)	66.07 (1)
CDF POI λ_{10P}	85.29 (3)	6.82 (2)	99.61 (2)	57.15 (4)
CDF POI λ_{10Y}	78.89 (1)	9.17 (4)	99.27 (4)	65.94 (2)

Table 7 indicates that:

1) CDF NB λ_{10P}, $O_i > \lambda_i$:
   In 85.29% of cases, $\hat{P}(O_i > \lambda_i)$ is more than 0.005; in 6.61% of cases, $\hat{P}(O_i > \lambda_i)$ is more than 0.05;
2) CDF POI λ_{10P}, $O_i > \lambda_i$:
   In 85.29% of cases, $\hat{P}(O_i > \lambda_i)$ is more than 0.005; in 6.82% of cases, $\hat{P}(O_i > \lambda_i)$ is more than 0.05;
3) CDF NB λ_{10Y}, $O_i > \lambda_i$:
   In 79.10% of cases, $\hat{P}(O_i > \lambda_i)$ is more than 0.005; in 7.68% of cases, $\hat{P}(O_i > \lambda_i)$ is more than 0.05;
4) CDF POI λ_{10Y}, $O_i > \lambda_i$:
   In 78.89% of cases, $\hat{P}(O_i > \lambda_i)$ is more than 0.005; in 9.17% of cases, $\hat{P}(O_i > \lambda_i)$ is more than 0.05;
5) CDF NB λ_{10P}, $O_i <= \lambda_i$:
   In 99.62% of cases, $\hat{P}(O_i <= \lambda_i)$ is more than 0.95; in 57.19% of cases, $\hat{P}(O_i <= \lambda_i)$ is more than 0.995;
6) CDF POI λ_{10P}, $O_i <= \lambda_i$:
   In 99.61% of cases, $\hat{P}(O_i <= \lambda_i)$ is more than 0.95; in 57.15% of cases, $\hat{P}(O_i <= \lambda_i)$ is more than 0.995;
7) CDF NB λ_{10Y}, $O_i <= \lambda_i$:
   In 99.36% of cases, $\hat{P}(O_i <= \lambda_i)$ is more than 0.95; in 66.07% of cases, $\hat{P}(O_i <= \lambda_i)$ is more than 0.995;
8) CDF POI λ_{10Y}, $O_i <= \lambda_i$:
   In 99.27% of cases, $\hat{P}(O_i <= \lambda_i)$ is more than 0.95; in 65.94% of cases, $\hat{P}(O_i <= \lambda_i)$ is more than 0.995;

According to the CDF analysis results shown in Table 7, in the cases of "$\hat{P}(O_i > \lambda_i) > 0.005$" and "$\hat{P}(O_i <= \lambda_i) > 0.995$", for the λ_{10Y} model combined with either the Poisson or the NB distribution, its GOFs are significantly better than λ_{10P} model's GOFs. In the cases of "$\hat{P}(O_i > \lambda_i) > 0.05$" and "$\hat{P}(O_i <= \lambda_i) > 0.95$", the criteria of the two models combined with the Poisson and the NB distributions have no obvious distinction, in particular, for the criterion "$\hat{P}(O_i <= \lambda_i) > 0.95$". Furthermore, λ_{10Y} combined with the NB distribution shows a slightly better quality than when combined with the Poisson distribution.

At a later stage in this paper, we carry out a comparison between the predicted probabilities of annual accident frequency observed at a given SAL2 according to the λ_{10Y} and the λ_{10P} combined with the Poisson and the NB distribution. In this context, group classification analysis is adopted to compute the mean predicted probability and the mean observed accident frequency of each group with the same sample size. The relationship between the mean predicted probability and the mean observed accident frequency is shown in Fig. 8. The blue and red scatters respectively represent the λ_{10Y} model combined with the Poisson and NB distributions. The interpretation of the remaining scatters involving the λ_{10P} can be similarly obtained. One can notice that the λ_{10Y} model combined with the Poisson and NB distributions respectively predict higher probabilities of observed accident frequencies than the λ_{10P} model combined with them. This difference is particularly noticeable in the case where the two prediction models are combined with the NB distribution (i.e., the red and green scatters). Moreover, one can also notice that the probability of accident occurrence estimated by the Poisson distribution combined with the λ_{10Y} model is higher than that estimated by the NB distribution combined with the λ_{10Y} model.

However, it should be recalled that the higher probability predicted does not necessarily indicate a higher predictive accuracy, since the probability of accident occurrence in reality would not be high. Therefore, further analysis to assess the predictive accuracy of the Poisson and the NB distributions should be carried out. As shown in Table 8, f_k denotes the percentage of samples of observed annual accident frequency with k accidents involved in a given year (f_k = the number of samples of observed annual accident frequency involving k accidents occurring in a given year / the total number of samples n). The estimated relative annual accident frequency reflected by estimated probabilities

Fig. 8. Average predicted Probability of annual accidents observed at a given SAL2 using λ_{10P} and λ_{10Y} combined with the Poisson and the NB distribution.

Table 8. The predictive accuracy comparison between the Poisson distribution and the NB distribution.

# Annual accidents considered (k)	Observed annual accident frequency (f_k in percent)	NB-λ_{10Y} estimated relative annual accident frequency ($\hat{f}_k$ in percent)	POI-λ_{10Y} estimated relative annual accident frequency ($\hat{f}_k$ in percent)	NB-λ_{10P} estimated relative annual accident frequency ($\hat{f}_k$ in percent)	POI-λ_{10P} estimated relative annual accident frequency ($\hat{f}_k$ in percent)
0	99.4371	99.3915 (1)	99.3903 (2)	99.3279 (3)	99.3255 (4)
1	0.5485	0.5999 (1)	0.6033 (2)	0.6647 (3)	0.6673 (4)
2	0.0144	0.0077 (1)	0.0062 (3)	0.0069 (2)	0.0055 (4)
> 2	0	0.0002 (2)	0.0001 (1)	0.0001 (1)	0.0001 (1)
Goodness score (the lower, the better)		5	8	9	13

on average is computed as: $\hat{f}_k = \sum_{i=1}^{n} \hat{P}(X_i = k)/n$, where $\hat{P}(X_i = k)$ is the estimated probability of k accidents occurring at a given SAL2 in a given year. According to the goodness of predictive accuracy ranked in brackets, the NB distribution shows a higher predictive accuracy with regard to various annual numbers of accidents occurring at a given SAL2 during the 10-year period, particularly, when combining with the λ_{10Y}. In the cases of 0, 1 and 2 accidents occurring at a given SAL2 in a given year, the predictive accuracy of the NB distribution combined with the λ_{10Y} takes the first place in all the cases, which means that the probabilities of accident occurrence predicted by the NB distribution combined with the λ_{10Y} are closest to the actual frequencies of accident occurrence. In the case of more than 2 accidents occurring at a given SAL2 in a given year, the predictive accuracy of the NB distribution combined with the λ_{10Y} takes the second place, with the deviation of only 0.0002% compared with f_k, the actual percentage of observed annual accident frequency samples. In fact, there are no SAL2 LXs showing more than 2 accidents in the same year during the 10-year period considered.

In fact, the prediction performance of ZIP and ZINB regression methods were also examined but resulted in no higher goodness-of-fit values and a quite small number of significant parameters (only 4 and 3 significant parameters corresponding to ZIP and ZINB, respectively) compared with the Poisson and NB models and, hence, were not reported in the comparison.

5. Discussion

5.1. Accident frequency prediction model evaluation

Based on the aforementioned analyses using the Monte-Carlo, AIC, BIC and PCS tests, one can notice that λ_{10Y} model for accident frequency prediction has better GOF and considers impacting variables more comprehensively (cf. Table 6, 9 characteristics considered in λ_{10Y} vs. 3 characteristics considered in λ_{10P}). In this section, we will scrutinize the impact of variables considered in λ_{10Y} model on SAL2 accident frequency.

5.1.1. Railway traffic characteristics

Two different characteristics regarding railway traffic are proved to be significant in terms of affecting SAL2 accident frequency, namely the average daily railway traffic and the railway speed limit. They positively influence the annual accident frequency at a given SAL2. The average daily railway traffic with a power of 0.646 has a more decisive impact on the accident frequency than the average daily road traffic with a power of 0.354, since the higher the railway traffic frequency appearing at SAL2 LXs, the much higher the SAL2 accident risk. Furthermore, the higher the railway speed limit, the higher the SAL2 accident risk. A tentative to explain this finding is that a high railway speed limit corresponds to a high actual train speed, therefore, the risk for train-MV collisions is higher.

5.1.2. Roadway traffic characteristics

There are two roadway traffic characteristics which have a significant impact on SAL2 accident frequency, namely the average daily road traffic and the annual road accidents. The average daily road traffic with a power of 0.354 has a positive impact on the annual accident frequency at a given SAL2. If there is no road traffic at a given SAL2, there would be no accidents caused by motorized vehicles at this SAL2. Moreover, the higher the combined exposure of railway and roadway traffic, the higher the likelihood of an accident occurring.

One of the most important characteristics to estimate the annual accident frequency at a given SAL2 is the annual road accidents. The impact of road accidents was likely to be ignored in the previous studies pertaining to LX safety analysis. The present study has clearly shown that the accidents at LXs are above all road accidents and they highly depend on the road safety level. Moreover, the road accident factor in λ_{10Y} model is time-dependent, since road accidents have an annual variation (vary every year). Correspondingly, the risk related to LX accidents has an annual variation as well.

5.1.3. LX characteristics

Four LX characteristics, namely the road alignment, the LX width, the crossing length and the LX-accident-prone factor of the region, have a significantly positive impact on the annual accident frequency at a given SAL2. It is worth recalling that the vertical road profile has no significant impact ($|t - statistic|$ of profile < 1.96). A tentative is discussed with SNCF experts to explain this fact as follows: on the one hand, the "hump or cavity" profile would cause an increasing risk of accidents involving long/heavy vehicles (trucks, buses, etc.), with relatively low population; on the other hand, for most of ordinary cars, such a profile obliges ordinary cars to cross the LX with low speeds, thus helps reduce the risk of LX accident occurrence. Besides, for LXs lacking the road profile information, they are treated as normal situations (no hump or cavity) according to SNCF experts' advice. Namely, 0 is used as the profile indicator for them when performing data coding. These conjectural reasons need to be further investigated in future works. In addition, although the above reasons may be suitable for French LX situation, they would depend on case-by-case scenarios and the policy implications in other countries.

For the other significant LX characteristics, a higher region LX-accident-prone factor, a poorer road alignment trace, a larger LX width and a larger crossing length correspond to a higher risk of LX accidents. According to the contribution degrees of variables ranked in Table 4, the contribution of the region LX-accident-prone factor to the risk of LX accidents takes the first place followed by the contribution of the road alignment, the LX width and the crossing length successively.

5.2. Statistic distribution evaluation

According to LL test results shown in Table 6, the NB distribution combined with the λ_{10Y} shows a better quality for predicting accident frequency with over-dispersion. Considering the CDF analysis shown in Fig. 7 and Table 7, the NB distribution combined with the λ_{10Y} also shows higher percentages of cases meeting the two conditions that the probability of the event "the accident frequency estimated no less than the actual accident frequency observed" is higher than 0.95 and 0.995, than those of the Poisson distribution combined with the λ_{10Y}. On the other hand, the NB distribution combined with the λ_{10Y} shows a lower percentage of cases meeting the condition that the probability of the event "the accident frequency estimated less than the actual accident frequency observed" is higher than 0.05, than those of the Poisson distribution combined with the λ_{10Y}.

Moreover, in terms of estimated probabilities of actual accidents occurring at a given SAL2 (cf. Table 8), the NB distribution combined with the λ_{10Y} shows a higher predictive accuracy with regard to various annual numbers of accidents occurring at a given SAL2 during the 10 years.

It is worth noticing that, although the NB distribution is more effective than the Poisson distribution when dealing with over-dispersed accident count, it requires more extensive computations to estimate the model parameters as well as the dispersion parameter and, to generate inferential statistics, compared with the Poisson model.

6. Conclusions

In the present study, we have developed an accident frequency prediction model based on LX accident statistics and various impacting factors. This model allows for predicting accident occurrence with a considerably high accuracy and has a more appropriate form compared with the existing models pertaining to LX accident prediction. Although the developed prediction model is tailored to SAL2 LX accidents in France, the general formula form of the model and the methodology adopted to set up the model and validate its quality can be applied to different contexts. The scientific selection process in our study ensures that the main impacting variables are considered and redundant variables are excluded. In fact, impacting variables pertaining to LX risk should be thought to be important in determining accident frequency, more permanent in nature and not accident-dependent.

The region LX-accident-prone factor which can indicate the impact of regional LX safety level on LX accident frequency is originally utilized in the improved model. CM, a more effective factor, is proposed in this study to replace the conventional traffic moment, single average daily railway traffic or single average daily road traffic when explaining the likelihood of LX collisions. The significant impact of road accidents, almost ignored in the past studies, is well considered in our study.

A validation of the model quality is performed by means of a set of comprehensive statistical approaches, namely the Monte-Carlo, AIC, BIC and PCS tests, which all indicate that the improved accident prediction model involving various influential factors is trustworthy and sound. Moreover, the LL test, CDF analysis and predictive accuracy analysis are conjunctively employed to identify a more appropriate statistical distribution for predicting the probability of LX accident occurrence. The results obtained attest that the improved model combined with NB distribution has relatively high predictive accuracy of the probability of accident occurrence. To the best of our knowledge, such a thorough validation process is rarely achieved in similar existing works.

To sum up, the above contributions of the present study offer an in-depth perspective on potential parameters causing LX accidents. Moreover, these contributions pave the way for identifying practical design measures and improvement recommendations to prevent accidents at LXs. In future works, we will establish Bayesian risk models to quantify the causal relationships between the impacting parameters and the risk related to LX accidents, and assess their respective impact on the LX safety level. In addition, the effectiveness of various technical solutions will be investigated based on some experiments that we have carried out at several LXs.

Acknowledgements

This work has been conducted in the framework of "MORIPAN project: MOdèle de RIsque pour les PAssages à Niveau" within the Railenium technological research institute, in partnership with the National Society of French

Railway Networks (SNCF Réseau) and the French Institute of Science and Technology for Transport, Development and Networks (IFSTTAR).

References

Anscombe, F.J., 1973. Graphs in statistical analysis. The American Statistician, 27 (1), 17-21.

Austin, R.D., Carson, J.L., 2002. An alternative accident prediction model for highway-rail interfaces. Accident Analysis & Prevention 34 (1), 31-42.

Borkowf, C.B., 2002. Computing the nonnull asymptotic variance and the asymptotic relative efficiency of Spearman's rank correlation. Computational statistics & data analysis 39 (3), 271-286.

Bozdogan, H., 1987. Model selection and Akaike's information criterion (AIC): The general theory and its analytical extensions. Psychometrika 52 (3), 345-370.

Buddhavarapu, P., Scott, J.G., Prozzi, J.A., 2016. Modeling unobserved heterogeneity using finite mixture random parameters for spatially correlated discrete count data. Transportation Research Part B: Methodological 91, 492-510.

Cameron, A.C., Trivedi, P.K., 1986. Econometric models based on count data: Comparisons and applications of some estimators and tests. Journal of applied econometrics 1 (1), 29-53.

Cameron, A.C., Trivedi, P.K., 1990. Regression-based tests for overdispersion in the Poisson model. Journal of econometrics 46 (3), 347-364.

Chadwick, S.G., Zhou, N., Saat, M.R., 2014. Highway-rail grade crossing safety challenges for shared operations of high-speed passenger and heavy freight rail in the US. Safety Science 68, 128-137.

Chang, L.Y., 2005. Analysis of freeway accident frequencies: Negative binomial regression versus artificial neural network. Safety science 43 (8), 541-557.

Chatterjee, S., Hadi, A.S., 2015. Regression analysis by example. John Wiley & Sons.

Dahiya, R.C., Gurland, J., 1972. Pearson chi-squared test of fit with random intervals. Biometrika 59 (1), 147-153.

Dai, H., Bao, Y., Bao, M., 2013. Maximum likelihood estimate for the dispersion parameter of the negative binomial distribution. Statistics & Probability Letters 83 (1), 21-27.

Davey, J., Wallace, A., Stenson, N., Freeman, J., 2008. The experiences and perceptions of heavy vehicle drivers and train drivers of dangers at railway level crossings. Accident Analysis & Prevention 40 (3), 1217-1222.

European Railway Agency (ERA), 2014. Railway safety performance in the European Union. 9 (2) Agency Regulation, 881/2004/EC.

Evans, A.W., 2011. Fatal accidents at railway level crossings in Great Britain 1946-2009. Accident Analysis & Prevention 43 (5), 1837-1845.

Evans, A.W., 2011. Fatal train accidents on Europe's railways: 1980-2009. Accident Analysis & Prevention 43 (1), 391-401.

Ghazel, M., 2009. Using stochastic Petri nets for level-crossing collision risk assessment. IEEE Trans. on Intelligent Transportation Systems 10 (4), 668-677.

Ghazel, M., El-Koursi, E.-M., 2014. Two-half-barrier level crossings versus four-half-barrier level crossings: A comparative risk analysis study. IEEE Trans. on Intelligent Transportation Systems 15 (3), 1123-1133.

Jarque, C.M., Bera, A.K., 1980. Efficient tests for normality, homoscedasticity and serial independence of regression residuals. Economics letters, 6 (3), 255-259.

Larue, G.S., Rakotonirainy, A., Haworth, N.L., Darvell, M., 2015. Assessing driver acceptance of Intelligent Transport Systems in the context of railway level crossings. Transportation Research Part F: Traffic Psychology and Behaviour 30, 1-13.

Lawless, J.F., 1987. Negative binomial and mixed Poisson regression. Canadian Journal of Statistics 15 (3), 209-225.

Leveson, N.G., 2011. Engineering a Safer World: Systems Thinking Applied to Safety. MIT Press, Cambridge.

Liang, C., Ghazel, M., Cazier, O., El-Koursi, E.-M., 2017. Risk analysis on level crossings using a causal Bayesian network based approach. Transportation Research Procedia 25, 2172-2186.

Liu, B., Ghazel, M., Toguyeni, A., 2016. Model-Based Diagnosis of Multi-Track Level Crossing Plants. IEEE Trans. on Intelligent Transportation Systems 17 (2), 546-556.

Lord, D., Mannering, F., 2010. The statistical analysis of crash-frequency data: A review and assessment of methodological alternatives. Transportation Research Part A: Policy and Practice 44 (5), 291-305.

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

Madsen, K., Nielsen, H.B., Tingleff, O., 2004. Methods for non-linear least squares problems (2nd Edition). Informatics and Mathematical Modelling, Technical University of Denmark.

Mekki, A., Ghazel, M., Toguyeni, A., 2012. Validation of a new functional design of automatic protection systems at level crossings with model-checking techniques. IEEE Trans. on Intelligent Transportation Systems 13 (2), 714-723.

Miaou, S.P., 1994. The relationship between truck accidents and geometric design of road sections: Poisson versus negative binomial regressions. Accident Analysis & Prevention 26 (4), 471-482.

National ALCAM Committee, 2012. ALCAM in Detail-An Introduction to the new ALCAM models, Australia.

Ogden, B.D., 2007. Railroad-Highway Grade Crossing Handbook, revised second ed. FHWA-SA-07-010, Springfield, Virginia 22161.

Oh, J., Washington, S.P., Nam, D., 2006. Accident prediction model for railway-highway interfaces. Accident Analysis & Prevention 38 (2), 346-356.

Read, G.J., Salmon, P.M., Lenné, M.G., Stanton, N.A., 2016. Walking the line: Understanding pedestrian behaviour and risk at rail level crossings with cognitive work analysis. Applied ergonomics 53, 209-227.

Salmon, P.M., Read, G.J., Stanton, N.A., Lenné, M.G., 2013. The crash at Kerang: Investigating systemic and psychological factors leading to unintentional non-compliance at rail level crossings. Accident Analysis & Prevention 50, 1278-1288.

SNCF, 2015. Research on the material of level crossing in 2014, France.

SNCF Réseau, 2011. World Conference of Road Safety at Level Crossings (Journée Mondiale de Sécurité Routière aux Passages à Niveau), France. From http://www.planetoscope.com/automobile/1271-nombre-de-collisions-aux-passages-a-niveau-en-france.html

SNCF Réseau, 2015. 8th National Conference of Road Safety at Level Crossings (8ème Journée Nationale de Sécurité Routière aux Passages à Niveau), France. From http://www.sncf-reseau.fr/fr/dossier-de-presse-8eme-journee-nationale-de-securite-routiere-aux-passages-a-niveau

SNCF Réseau, 2010. Statistical analysis of accidents at LXs, France.

Stanton, N.A., Walker, G.H., 2011. Exploring the psychological factors involved in the Ladbroke Grove rail accident. Accident Analysis & Prevention 43 (3), 1117-1127.

Stefanova, T., Burkhardt, J.-M., Filtness, A., Wullems, C., Rakotonirainy, A., Delhomme, P., 2015. Systems-based approach to investigate unsafe pedestrian behaviour at level crossings. Accident Analysis & Prevention 81 (0), 167-186.

Tey, L.S., Ferreira, L., Wallace, A., 2011. Measuring driver responses at railway level crossings. Accident Analysis & Prevention, 43 (6), 2134-2141.

Weakliem, D.L., 1999. A critique of the Bayesian information criterion for model selection. Sociological Methods & Research 27 (3), 359-397.

Wigglesworth, E.C., 2001. A human factors commentary on innovations at railroad-highway grade crossings in Australia. Journal of Safety Research 32 (3), 309-321.

Wilson, J.R., 2014. Fundamentals of systems ergonomics/human factors. Applied Ergonomics 45, 5-13.

Woods, M.D., Slovak, R., Schnieder, E. et al., 2008. Safer European Level Crossing Appraisal and Technology (SELCAT)-D3 Report on Risk Modeling Techniques for level crossing risk and system safety evaluation. Rail Safety and Standards Board (RSSB), 66-67.