Modeling Crash Frequency at Highway-Railroad Grade Crossings Using a Two-Stage Classification and Regression Tree Method

Shou-Ren HU [a], An-Chi HSIEH [a], Chi-Kang LEE [b]

[a] *Department of Transportation and Communication Management Science, National Cheng Kung University, Tainan City 70101, Taiwan; E-mail: shouren@mail.ncku.edu.tw*

[b] *Department of Marketing and Logistics Management, Southern Taiwan University of Technology, Tainan City 71005, Taiwan; E-mail: leeck@mail.stut.edu.tw*

Abstract: This research investigates the factors that have effects on the occurrence of traffic collisions at highway-railroad grade crossings (HRGXs) using a two-stage classification and regression approach. In particular, the collinearity problem generally confronted in linear regression models is avoided by a pre-stratified structure in the explanatory variables. Classification and regression tree (CART) is adopted to identify the key factors that are responsible for traffic collisions at HRGXs. Using the HRGX crash and grade crossing inventory data provided by Taiwan Railways Administration (TRA) during 2008~2010 in Taiwan, the empirical study results indicate that number of daily trains and highway width are positively associated with crash frequency, and the four classifiers identified in the variable classification stage are also found to be positively related to the number of traffic collisions at HRGXs.

Keywords: Highway-railroad Grade Crossing, Traffic Collision, Crash Frequency, Classification and Regression Tree

1. INTRODUCTION

According to the annual statistical report released by Taiwan Railways Administration (TRA), Ministry of Transportation and Communication (MOTC), till December 2011 there were still 555 Highway-Railroad Grade Crossings (HRGXs) in Taiwan. From 2008 to 2010, 105 crashes have occurred at 569 HRGXs in Taiwan and resulted in 26 fatalities and 25 injuries. These numbers are much higher than those in the international rail community (Laffey, 2010). Despite HRGX crashes are rare incidents, they might incur significant number of casualties and different degrees of property losses. In 2009, although only approximately 5% of the total incidents in the entire TRA system are HRGX related, the average casualty impact of the traffic collisions at HRGXs are three times higher than that of the total railway incidents

(TRA, 2010). Facing such a serious safety problem, how to identify the key factors causing traffic collisions and their potential impacts is a crucial issue for the rail authority in Taiwan.

According to the American Public Transit Association (APTA), an HRGX is that where the railroad crosses the road and users share the same right-of-way (ROW) and installs traffic control equipment on both sides to make sure that highway users and trains are passing safely. In this spatial area, because of different kinds of transportation modes passing through or driving around, there are two different traffic characteristics. On the railway side, trains with large mass can't arbitrarily change the direction and have slow reaction in emergencies. On the highway side, users including motor vehicle, pedestrian, and cyclist have shorter stopping distance and are able to easily control direction and quick reaction in case of emergencies. The purpose of traffic warning/control equipment is to promote a safer and more efficient operation of both rail and highway traffic at an HRGX. Both passive- and active-type traffic control devices are widely employed, including traffic sign, signal, markings, illumination, emergency button, train detection device, flashing light and warning bell, barrier-protected gate, and the addition of extra warning devices such as barrier or wayside horn, and law enforcement camera, etc.

In recent years, the TRA maintains a good crash database and HRGX inventory dataset. Hence a few local researches used statistical models to analyze these crash data. Hu *et al.* (2010; 2011; 2012) investigated crash severity, risk levels, and crash frequency using different categorical or count data regression models such as generalized logit model, zero-inflated Poisson model, and negative binomial model in which traffic exposure variables including number of daily trains and average daily roadway traffic are always found to be positively associated with the risk levels of an HRGX.

There are numerous researches modeling HRGX crash frequency and/or severity in the international rail community. Saccomanno *et al.* (2004) used a risk-based model to identify highway-rail grade crossing black spots in which two components are specifically modeled: crash frequency and consequence. The model was applied to the Canadian HRGX inventory and crash data for the period of 1994-2001. Poisson and negative binominal (NB) frequency prediction expressions were developed for crossings with three types of warning devices. Finally, an NB expression which has better fit to the crash frequency was developed for the crash consequence model. The spatial distribution of black spots is discussed with respect to the type of warning device, upgrades in warning device, geographic location, and historical crash frequency. In addition, some research used zero inflated models to investigate the traffic collisions at HRHXs because crashes are rare incidents and traditional statistical models might underestimate the probability (Lee *et al.*, 2005; Lord *et al.*, 2005; Lord *et al.*, 2007).

The application of count data statistical models, however, has been plagued by a number of methodological and practical issues, such as a lack of statistical significance of factor inputs, higher-order interaction between data, and the presence of collinearity among model inputs (Chang and Chang, 2005; Chang and Wang, 2006; Park and Saccomanno, 2005). To improve these deficiencies, Park and Saccomanno (2005) used tree-based data mining technique and statistical methods to estimate the main and interactive effects of introducing countermeasures at individual grade crossings in Canada. Yan *et al*. (2010) also applied a nonparametric statistical method, hierarchical tree-based regression (HTBR) model, to predict train-vehicle crash frequency for passive grade crossings controlled by cross bucks only and cross bucks combined with stop signs, respectively; and assess how the crash frequency changes after the stop-sign treatment is applied at the crossbuck-only-controlled crossings. Additionally, to analyze the severity levels of highway traffic crash, Chang and Chen (2005) and Chang and Wang (2006) considered that the classification and regression tree (CART) model is a good alternative to identify the risk factors that are associated with the occurrence of highway crashes.

As revealed in the above literature review, modeling traffic collisions using one-stage count data statistical models confronts with some theoretical problems. Thereby, in this research we use a two-stage hierarchical regression model framework to investigate the main and interaction effects of a set of explanatory variables, and explore the key factors that might contribute to traffic collisions at HRGXs. In summary, the objectives of this research are twofold: 1) to identify the risk factors associated with traffic collisions at HRGXs; and 2) evaluate the countermeasure effects of the identified significant variables. This research conducts the empirical study by using the HRGX crash dataset collected by the TRA during 2008~2010. The dataset also includes the basic properties and attributes of the investigated HRGXs, such as crossing types, highway geometric characteristics, daily trains, and highway vehicular traffic, etc.

The remainder of this paper is organized as follows. Section 2 describes the methodological aspect of the proposed two-stage hierarchical regression model framework. Section 3 depicts the crash dataset and crossing inventory data used in the empirical study. Section 4 provides the empirical study results and insight into the policy implications. Finally, findings and limitations of this research are summarized in Section 5, and future research directions are also suggested.

2. METHODOLOGY

In view that crash data are discrete and nonnegative integer in nature, we use a two-stage hierarchical regression model framework to explore the risk factors at HRGXs. In the first stage, a tree-based data mining method is adopted for crash data classification. In the second stage, a count-data statistical model is used to identify the causal relationship between the classified risk factors and crash frequency. Details of the methodologies are described below.

2.1 Tree-based Data-mining Methods

A hierarchical tree-based regression model which is a data exploration method classifies observations by recursively partitioning the predictor space. Due to its non-parametric nature and easy interpretation, it has received wide popularity in various fields (Chang and Chang, 2005; Chang and Wang, 2006). Compared with the traditional linear regression models or count-data statistical models, hierarchical tree-based regression models effectively deal with the complex relationships among the explanatory variables such as collinearity and interaction effects in a large data set (Conerly *et al.*, 2000).

Classification and regression tree (CART) which is a popular data mining technique is used in this research. The CART algorithm, originally proposed by Breiman *et al*. (1983), is a simple non-parametric method. The prediction rules are given in the form of binary decision trees and it is easy to understand, use, explain, and interpret. In this study, a commonly applied technique called recursive partitioning (RPART) is used to identify the interactions among a set of risk factors for traffic collisions at HRGXs.

RPART is characterized by the application of splitting rules in the data. It splits a sample into binary subsamples using a set of "yes-no" questions. The sample at a higher level is split into two left and right lower-level subsamples. Continue splitting sample according to node impurity until it cannot be split any further. At this time, the end node is called terminal node and represented by a rectangular. Besides, the other nodes are called non-terminal nodes (the decision nodes) and represented by a circle. Figure 1 shows an example of a tree structure.

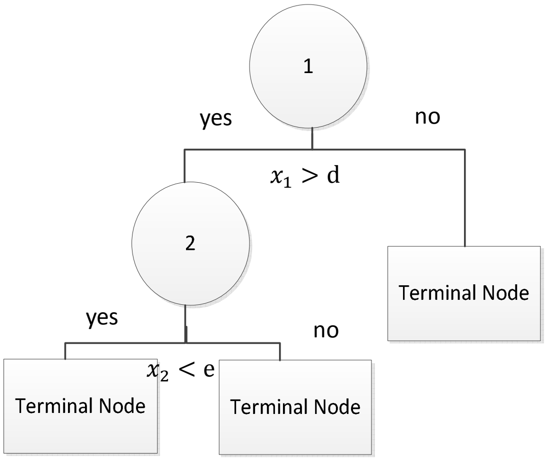

Figure 1. Tree structure using RPART

The RPART guiding principles used to construct the hierarchical regression trees contain the following four steps (Park and Saccomanno, 2005).

1. RPART starts with the root node; RPART performs all splits on each of the explanatory variables, applies a predefined node impurity measure to each split, and determines the reduction in impurity that is achieved.

2. RPART then selects the best split by applying a goodness-of-split criteria and partitioning the data set into left and right sub nodes.

3. Because RPART is recursive, it repeats Steps 1 and 2 for each of the nonterminal nodes, resulting in the largest possible tree. The change in impurity of node t of each split can be estimated using the following expression:

$$\vartriangle D(s,t) = D(t_C) - D(t_L) - D(t_R) \tag{1}$$

where,

$D(t_C) = \text{impurity (i.e., deviance) at current node } t,$
$D(t_L) = \text{impurity (i.e., deviance) at the left subnode } t_L,$ and
$D(t_R) = \text{impurity (i.e., deviance) at the right subnode } t_R.$

4. From the series of splits generated by a variable at a node, the rule is to choose the split that maximizes the reduction in the impurity at the current node. The best split is that associated with the highest $\vartriangle D(t)$ value.

In the likelihood-ratio (LR) criterion, each node's impurity is measured as the within-node deviance:

$$D(t) = \sum \left[y_i \, log\left(\frac{y_i}{\hat{\lambda} t_i}\right) - \left(y_i - \hat{\lambda} t_i\right) \right] \tag{2}$$

where,

$y_i = \text{observed event count for observation } i,$
$t_i = \text{baseline measure for observation } i \text{ (e.g., index of the time and space), and}$
$\hat{\lambda} = \frac{\sum y_i}{I} = \text{observed overall event rate.}$

The impurity measure has the property that $\mathrm{D}(t_C) \geq D(t_L) + D(t_R)$, meaning that the current impurity estimate is greater than or equal to the impurity estimates of the node created at the current split. Using this character, we can build the right-size tree for the causal analysis in the second stage. More details about tree-growing and tree-pruning algorithms for the RPART can be found in Breiman *et al.* (1983).

2.2 Count-data Statistical Models

After identifying the hierarchical relationship between crashes and the explanatory variables in the CART model, we will use count data statistical regression models to further investigate the key factors contributing to traffic collisions and/or casualties at HRGXs. For crash frequency analysis, many researches used Poisson regression, negative binominal (NB) regression, or Zero-inflated Poisson (ZIP) regression models. Since crash frequency is count data with discrete and nonnegative integer characteristics, hence Poisson regression model is more suitable than multiple linear regression models. However, the overdispersion problem frequently found in most crash data where the variance of data is greater than their mean; resulting that mean estimates given by Poisson regression models might be biased. On the other hand, negative binomial models provide a plausible general form that effectively captures data overdispersion phenomena, and it is applied to the present research.

The NB regression model is similar to a Poisson regression model except that an error term with Gamma distribution is associated with the mean of an event (here we define such an event is a traffic collision(s) at HRGXs). The functional form of a NB model is as following:

$$\lambda_i = e^{\beta X_i + \mathcal{E}_i} \tag{3}$$

It can be rewritten as:

$$\ln \lambda_i = \beta X_i + \mathcal{E}_i \tag{4}$$

where,

λ_i: expected number of collisions from Poisson regression; and

$\mathcal{E}_\mathrm{i} \sim Gamma(1, \alpha^2)$.

The conditional probability has the following form:

$$P(a_i|\mathcal{E}_i) = \frac{exp[-\lambda_i\, exp(\mathcal{E}_i)][\lambda_i\, exp(\mathcal{E}_i)]^{a_i}}{a_i!} \quad , \ a_i = 0,1,2, \dots; i = 1,2, \dots, n. \tag{5}$$

Eq. (5) formulates the probability of a_i times occurred at the i^{th} HRGX in a given period of time.

Using Eq. (5), we can obtain the following probability distribution by integral operation:

$$P(a_i) = \frac{\Gamma(\theta + a_i)}{\Gamma(\theta)a_i!} \mu_i^{\theta}(1 - \mu_i)^{a_i} \tag{6}$$

where $\mu = \frac{\theta}{\theta + \lambda_i}$, $\theta = \frac{1}{\alpha}$, and $\Gamma(\cdot)$ is the gamma function.

The expected value and variance are given below:

$$E[A_i] = \mu \tag{7}$$

$$Var[A_i] = \mu + \alpha\mu^2 \tag{8}$$

The maximum Likelihood estimation function of the NB regression model for λ_i is given as follows:

$$L(\lambda_i) = \prod_{i=1}^{n} \frac{\Gamma\left(\left(\frac{1}{\alpha}\right)+a_i\right)}{\Gamma\left(\frac{1}{\alpha}\right)a_i!} \left(\frac{\frac{1}{\alpha}}{\left(\frac{1}{\alpha}\right)+\lambda_i}\right)^{\frac{1}{\alpha}} \left(\frac{\lambda_i}{\left(\frac{1}{\alpha}\right)+\lambda_i}\right)^{a_i} \tag{9}$$

From Eqs. (7) and (8) one can see that this model allows the variance to exceed the mean, and $Var[A_i] = E[A_i][1 + \alpha E(A_i)] = E[A_i] + \alpha E[A_i]^2$. By this equation, the Poisson regression model can be regarded as a limiting model of the negative binomial regression model as α approaches 0 and the expected value is equal to its variance. In other words, once α is equal to 0, the NB distribution will be becoming the Poisson distribution. Hence, the Poisson regression model is a special case of the NB model, where α is the overdispersion parameter.

In consideration of the data overdispersion, Cameron and Trivedi (1990) and Greene (2000) proposed a statistic hypothesis test to decide a suitable model between Poisson and NB models. The hypotheses of this model selection test are listed below:

$$\begin{aligned} &H_0: \alpha = 0 \\ &H_1: \alpha \neq 0 \end{aligned} \tag{10}$$

We can test that if the overdispersion parameter α is zero or not by using the t-test. Finally, if we reject the null hypothesis (i.e. $\alpha \neq 0$), meaning that the expected value does not equal to variance of the data, and the decision rule is to choose the NB model; otherwise we choose the Poisson model. In both of the models, the likelihood ratio statistics ρ^2 is used as an indicator to examine the performance of the used model(s).

3. DATA DESCRIPTION

3.1 Data Collection

Up to 2011, there are 542 HRGXs at the Taiwanese traditional railway system (i.e. the TRA system). The crash data for the empirical study were collected by the TRA, MOTC during 2008~2010. Relevant crash history and crossing inventory data of a total of 795 HRGXs were collected. The dataset includes both crash and basic attributes of the investigated HRGXs.

In the traffic collision data, it includes crash frequency, number of injury and fatalities. Besides, the basic attributes collected are: railway operation (e.g., daily trains, train speed, and number of tracks), crossing facilities (e.g., crossing type, crossing angle, obstacle detection device, flash light and warning bell, and emergency button, etc.), and highway characteristics (e.g., highway type, grade, width, and daily vehicular traffic, etc.).

3.2 Description of the Empirical Data

In the 795-HRGXs dataset, by removing those being abolished, elevated, or closed to use of the HRGXs after 2010 and those with incomplete data registration, a total of 569 HRGXs were selected for the empirical study.

In the dataset, 105 crashes in total and 12 HRGXs confronting with one traffic collision during the three-year time period. These traffic collisions have resulted in 26 fatalities and 25 injuries. In addition, these crashes involved 32 automobiles, 24 pedestrians, and 22 motorcycles, which account for 74% of all the vehicle types. Table 1 provides the detailed description of the variables.

Table 1. Definition of the variables

Feature	Variable	Description	Type	Definition
Traffic accident and casualty	Crash	target variable	Categorical	0: no crash; 1: an crash
	Injury	number of injuries caused by a crash	Quantitative	number of injuries
	Fatality	number of fatalities caused by a crash	Quantitative	number of fatalities
Rail	Line	TRA's main operating lines around the Taiwan main island	Categorical	1: western line; 2: Taichung line; 3: Pingtung line; 4: Yilan line; 5: North-link line; 6: Taitung line; 7: South-link line; 8: Linkou line; 9: Neiwan line; 10: Jiji line; 11: Taichung harbor line; 12: Kaohsiung harbor line; 13: other lines
	Daily trains	number of trains at an HRGX	Quantitative	number of trains per day
	Train radio	radio used to provide message for protecting train from crash	Categorical	0: without; 1: with
Highway	Number of daily traffic	daily highway vehicular traffic	Quantitative	PCUs per day or AADT
	Road width	highway width of an HRGX	Quantitative	meter
	Highway level	highway function and administration	Categorical	1: provincial highway; 2: urban road; 3: prefectural road; 4: community road; 5:

				county road; 6: village road; 7: agriculture road; 8: special road; 9: port road
	Highway signal	number of coordinated highway intersections	Quantitative	number of highway signals coordinated
	Vehicle type	main vehicle or user type	Categorical	1: truck; 2: pickup; 3: automobile; 4: motorcycle; 5: bicycle; 6: pedestrian; 7: agricultural vehicle; 8: others
Crossing	Crossing type	type of HRGX	Categorical	1: type one; 2: type two; 3: type three; 4: manual; 5: special; 6: half-closure
	Crossing width	railway width of an HRGX	Quantitative	meter
	Crossing angle	angle of the railroad crossing the road	Quantitative	degree
	Surface type	HRGX's crossing surface type	Categorical	1: flat; 2: raised; 3: depression; 4: slant; 5: others
	6-meter arm	number of four-quadrant gates with 6-meter arm	Quantitative	positive integer
	8-meter arm	number of four-quadrant gates with 8-meter arm	Quantitative	positive integer
	Flash	flashing light	Categorical	0: without; 1: with
	Alarm	warning bell	Categorical	0: without; 1: with
	Surveillance	video surveillance camera	Categorical	0: without; 1: with
	Manual warning	manual warning device	Categorical	0: without; 1: with
	Auto warning	automatic warning device	Categorical	0: without; 1: with
	Small indicator	traditional train approaching direction indicator	Categorical	0: without; 1: with
	Large indicator	advanced LED train approaching direction indicator	Categorical	0: without; 1: with
	Detector	infrared obstacle	Categorical	0: without; 1: with

		detector		
	Emergency button	equipment used in emergency situations	Categorical	0: without; 1: with
	Gate	quadrant-type barrier at HRGXs to avoid breaking incidents	Categorical	0: without; 1: with
Others	Weather	climatic condition when a crash is occurred	Categorical	1: sunny; 2: cloudy; 3: rainy; 4: typhoon
	Responsibility	whether the TRA is responsible for the crash	Categorical	0: no; 1: yes
	Region	administrative region around the main island of Taiwan	Qualitative	1: north (Taipei, Keelung, Taoyuan, and Hsinchu); 2: west (Taichung, Nantou, Miaoli, Changhua, and Yunlin); 3: south (Chiayi, Tainan, Pingtung, and Kaohsiung); 4: east (Yilan, Hualien, and Taitung)

4. RESULTS AND DISCUSSION

4.1 Results of the CART Classification

Thirty variables were used in an attempt to identify the potential influence factors that might cause traffic collisions and/or casualties at HRGXs. This research has developed the CART model to classify the risk factors. Figure 2 shows the overall classification result. The tree finally produces a total of twenty three terminal nodes, and the tree can be classified according to the TRA's main operating lines, highway level, road width, number of daily traffic, and some crossing attributes. The result indicates that those variables are critical to explain the causes of crashes at HRGXs. It is easy to depict the hierarchical relationships among the explanatory variables. In the root node, the model starts splitting based on the variable of (train) **Line**. This indicates that TRA's main operating lines are suitable to classify traffic collisions at HRGXs, which can significantly distinguish the target variable by reducing the impurity of the nodes). CART classifies the lines of Pingtung, Taitung, South-link, Jiji, Taichung harbor, Kaohsiung harbor, and other lines into the left side of the tree (See Fig. 3). Because these lines possess similar operational characteristics; they serve relatively less number of daily trains on these lines. Therefore in node 2, we can conspicuously find that under these less busy lines, the probability of traffic collision only accounts for 3.4%. The next split variable is **Highway level**, shown in terminal node 1, including highway levels of provincial highway, urban road, lane, county road, village road, agriculture road, and port road. In terminal node 1, the probability of crash is almost zero. Looking at the left-hand side of the tree, we conclude that for those HRGXs located in Jiji line, Taichung harbor line, and the highway level is prefectural road or special road, a higher crash probability is observed.

In the right-hand side of the tree (see Figures 4 and 5), it is found that for the western, Taichung, Yilan, North-link, Linkou, and Neiwan lines, a higher crash probability is also found. It is because that more daily trains serving these lines; causing more traffic collisions in these service areas. Finally, under these different influence factors, the significant crash rates are found in terminal nodes 15, 18, 21, and 23. The result indicates that when a grade crossing located at an area with higher daily trains, highway width greater than 4.95 meters, crossing angle less than 97.5 degrees, and the highway level is urban road or some non-main artery, a higher crash probability is generally observed.

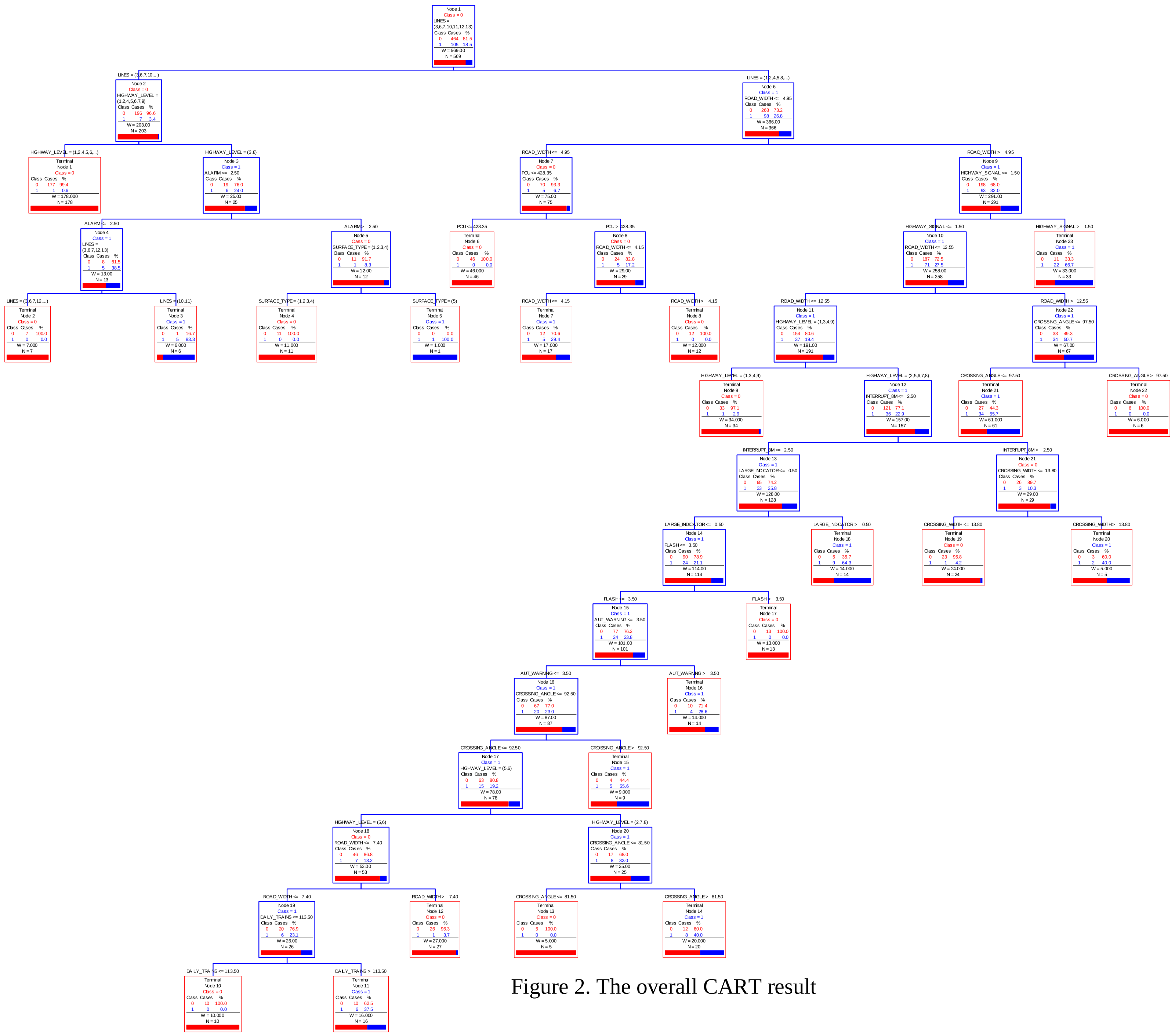

Figure 2. The overall CART result

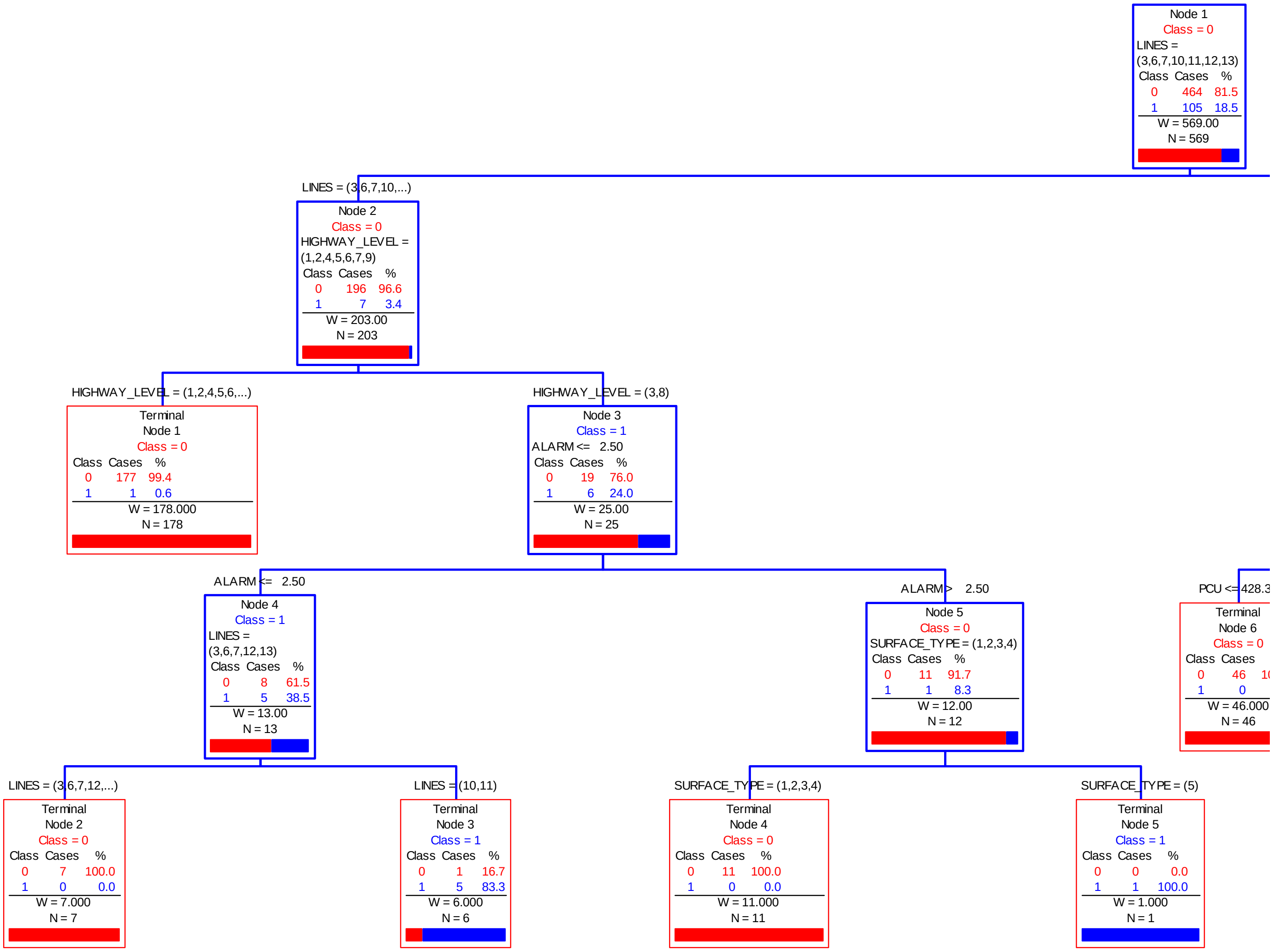

Figure 3. Left-hand side of the CART result

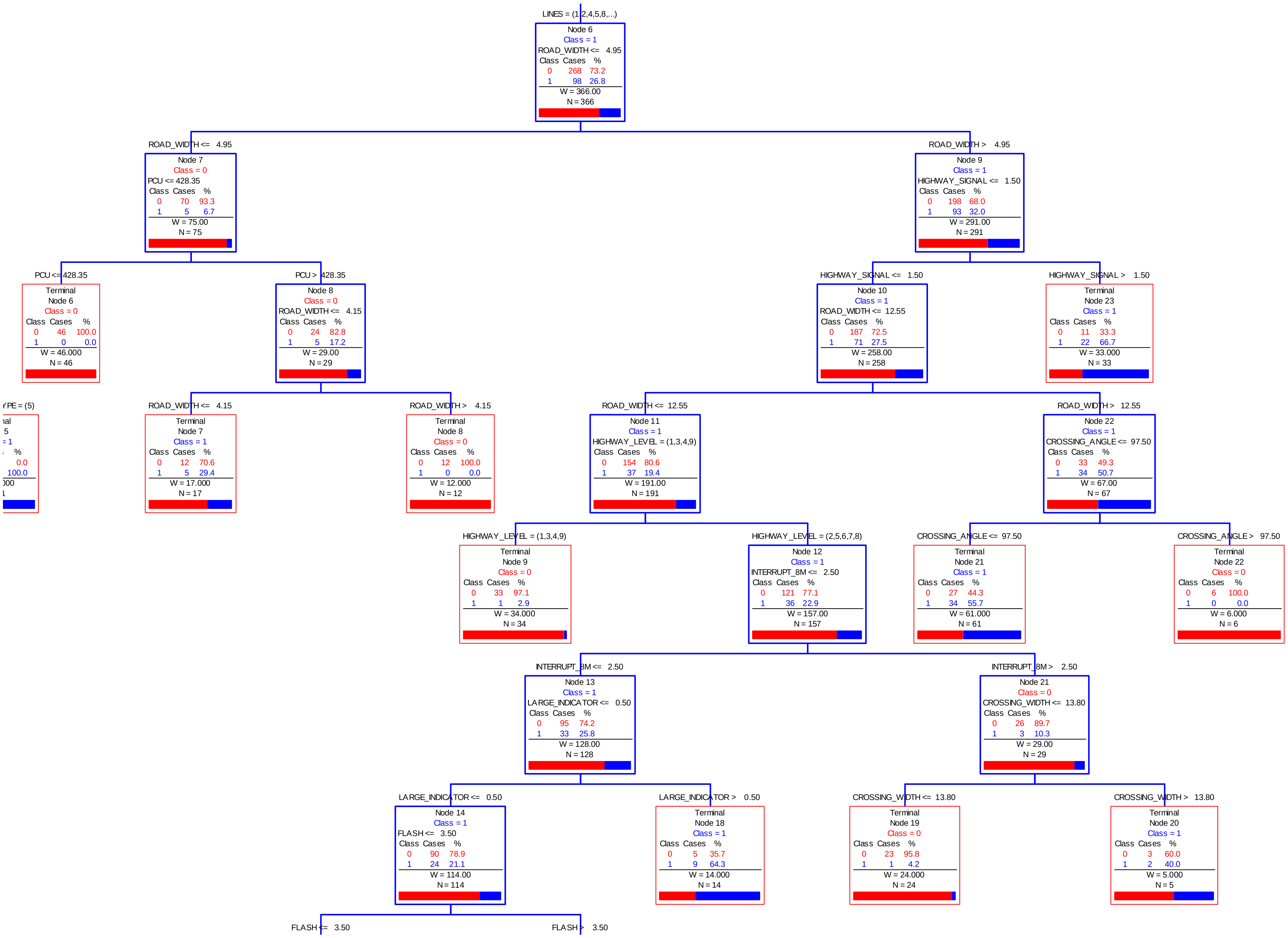

Figure 4. Right-hand side of the CART result (1/2)

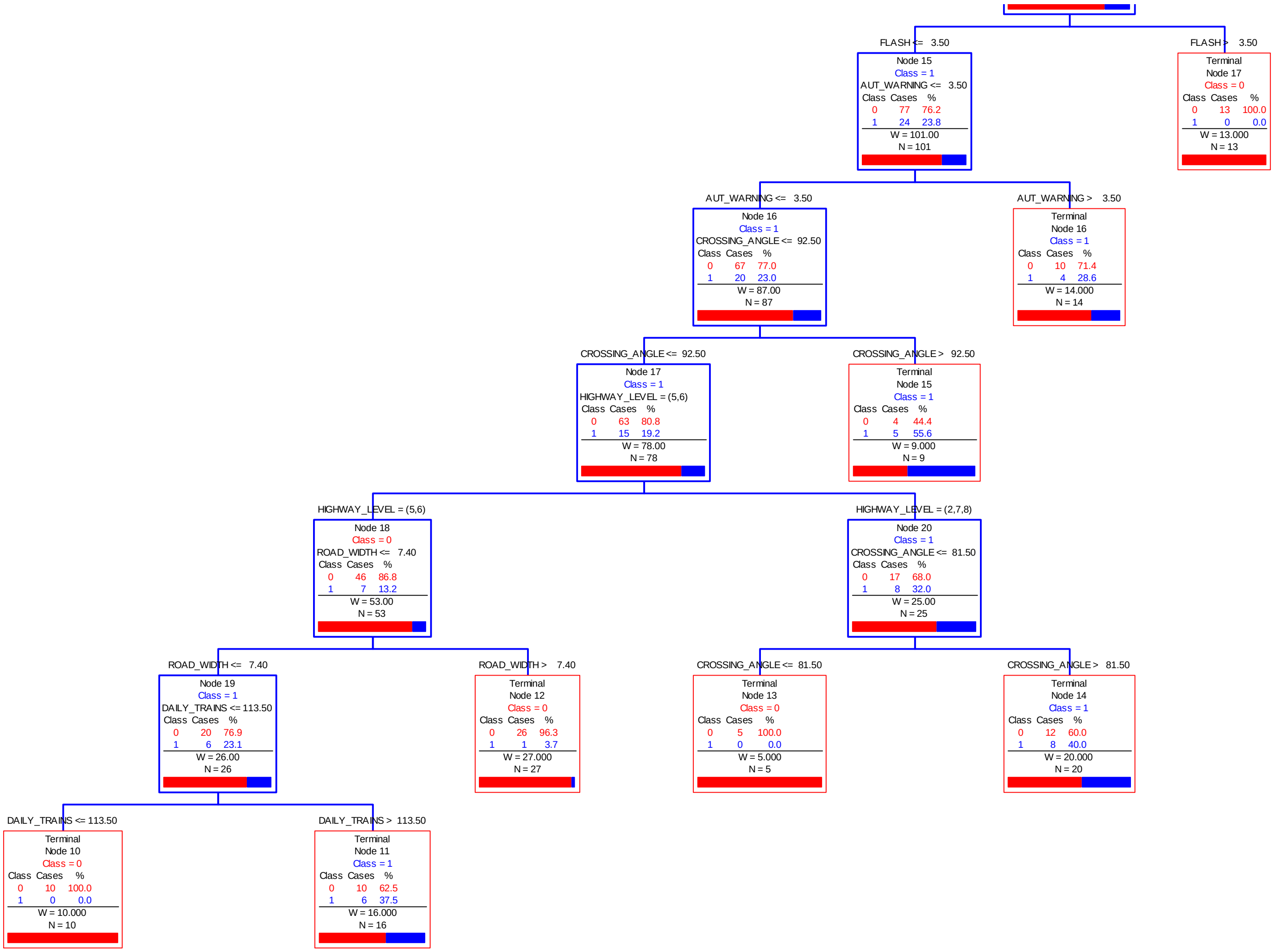

Figure 5. Right-hand side of the CART result (2/2)

4.2 Count-data Statistical Model for Crash Frequency

After identifying the key factors revealed in the first-stage of applying the CART algorithm, we can continue to conduct the count data statistical regression modeling to further explore the potential impact of the key factors on traffic collisions.

Before starting the NB regression analysis, we need to decide which variables to be incorporated into the crash frequency model. The CART algorithm also provides beneficial information to objectively conduct the variable selection process. As a result, we selected seven variables into the NB modeling, including **Crossing angle**, **Number of daily traffic**, **Daily trains**, **Line**, **Crossing width**, **Road width**, and **Highway level**; four classifiers (interaction) variables ***C1***, ***C2***, ***C3***, and ***C4*** are also incorporated into the model process. Details of variable settings and description are provided in Table 2.

The model estimation result shown in Table 3 indicates that the overdispersion parameter α is not significant ($t = 0.002 < 1.96$), meaning that the expected value of the data is equal to its variance, so we choose the Poisson regression model for later analysis. Under the 5% level of confidence and referring to the CRAT classification results, we find that most of the selected variables are significantly associated with the occurrence of crashes. The empirical study result shows that the geographical and/or environmental conditions (*C1~C4*) have more probability causing traffic collisions at HRGXs. For instance, *C1* represents terminal node 23 in the CART result where the HRGXs are located at western, Taichung, Yilan, North-link, Linkoue, and Neiwan lines. In addition, for those HRGXs with road width greater than 4.95 meters, highway signals are more than one. In Table 3, the coefficient of *C1* is about 1.161, meaning that under the *C1* environmental conditions, it significantly increases crash frequency (*t*-value = 3.403).

Beside, *line 2* also meanings that when an HRGX locate at the TRA’s main operation lines where relatively high traffic volume is observed, it could increase crash probability; the same finding is confirmed in the CART analysis. When crash frequency is increased, it would simultaneously increase the probability of injuries and/or fatalities. Finally, these coefficients of the selected variables are all reasonably positive, which conforms to our expectation. However, highway level and crossing angle are not found to be significant in this model.

Table 2. Variable settings and description for the NB regression model

Variable	Notation	Definition
Crossing width	*X1*	Meter (*m*)
Number of daily traffic	*X2*	Passenger car unit (PCU)
Daily trains	*X3*	Trains/day
Line	*Line 0*	1 for Pingtung line, Taitung line, South-Link line, Kaohsiung-Harbor line, and other lines; 0 for otherwise.
	Line 1	1 for Jiji line and Taichung harbor lines; 0 for otherwise.
	Line 2	1 for western, Taichung, Yilan, North-link, Linkou, and Neiwan lines; 0 for otherwise.
Crossing width	*X5*	Meter (*m*)
Road width	*X6*	Meter (*m*)
Highway level	*Highway 0*	1 for provincial highway, prefectural road, community road, and port road; 0 for otherwise.
	Highway 1	1 for urban road, county road, village road, agriculture road, and special road; 0 for otherwise.
Classifier 1	*C1*	1 for terminal node 23; 0 for otherwise.
Classifier 2	*C2*	1 for terminal node 21; 0 for otherwise.
Classifier 3	*C3*	1 for terminal node 18; 0 for otherwise.
Classifier 4	*C4*	1 for terminal node 15; 0 for otherwise.

In Table 3, we can further evaluate the model's capability of modeling the crash data. The likelihood ratio statistics ρ^2 index is given as:

$$\rho^2 = 1 - \frac{LL(\beta)}{LL(0)} \tag{11}$$

where $LL(\beta)$ is the log likelihood function with an intercept and predictors, $LL(0)$ is restricted log likelihood with only an intercept.

As shown in Table 3, the $\rho^2 = 0.186$ in the estimated model, indicating that the estimated model is statistically satisfactory.

Table 3. Results of the Poisson regression model

Variable	Coefficient	Standard Error	b/St. Er.
Constant	-3.998	0.469	-8.520
C1	1.161	0.341	3.403
C2	0.691	0.368	1.876
C3	1.295	0.473	2.739
C4	1.290	0.535	2.414
Line 2	1.686	0.515	3.275
X3	0.002	0.002	1.143
X6	0.029	0.017	1.737
Number of observations		502	
Log likelihood function		-204.615	
Info. Criterion: *AIC*		0.847	
Finite Sample: *AIC*		0.848	
Restricted log likelihood		-251.335	
McFadden Pseudo *R*-squared		0.186	
Chi squared		93.441	

5. CONCLUSIONS AND RECOMMENDATIONS

This research analyzes the newly collected HRGXs data in Taiwan using CART to classify the HRGXs crash and inventory data, and accordingly develop count data statistical regression models to explore the causal relationships between crash frequency and a set of influence factors aiming to find the key risk factors causing traffic collisions at HRGXs. Based on the empirical study results, we summarize the findings, address the limitations, and provide future research directions as below.

5.1 Findings and Conclusions

In the CART analysis result, the variables of line, daily trains, road width, highway level, crossing width, number of daily traffic, crossing angle, and some equipment variables are the important factors resulting in traffic collisions at HRGXs. Specifically, for those HRGXs located at the TRA's main operating lines, larger road width, and higher vehicle volume, it might result in a higher possibility of a crash. Secondly, by conducting the Poisson regression modeling process, we verify the same result of the selected influence factors. Furthermore, despite the newly collected HRGX data are relatively less in terms of number of traffic collisions in these three years, using the non-parametric method can help us identify the data collinearity and interaction problems. The empirical study results also show that the two-stage hierarchical regression model framework for the investigation of traffic collisions at HRGXs is satisfactory. Finally, the risk factors identified in this research are expected to provide the TRA with beneficial information to undertake effective safety improvement programs.

5.2 Limitations and Future Research

One of the research limitations is the relatively short time period of data collection (2008 through 2010). A lack of sufficient samples makes it difficult to reveal the key risk factors affecting the safety levels at HRGXs. For example, the current dataset can’t provide us with further crash risk analysis in terms of *crash severity* and/or *number of casualties in a given period of time* at HRGXs. For future research, a number of issues can be pursued. First, from the methodological perspective, some advanced statistical models for count data can be adopted and evaluated for data correlation research. By comparing the identified risk factors and prediction performance between different statistical models, it can provide valuable insights into the underlying relationships between a set of risk factors and traffic collisions at HRGXs. Second, in the data collection process, we did not evaluate data collection methods in this research. Future studies might investigate the effect of non-random samples on the analysis results of the adopted non-parameter models. Finally, in this study we only consider of geographical and/or environmental conditions at HRGXs; highway users’ behaviors at HRGXs is one of the main concerns that is worthy of further investigation.

ACKNOWLEDGEMENTS

The authors would like to thank TRA, which provided the historical crash data and grade-crossing inventory dataset in the empirical study. This research was partially supported by a grant (NSC 100-2221-E-006-219) from the National Science Council, Taiwan.

REFERENCES

Breiman, L., Friedman, J. H., Olshen, R. A., Stone, C. J. (1983) *Classification and Regression Trees*, Belmont, Wadsworth, California.

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

Chang, L. Y., Chen, W. C. (2005) Data mining of tree-based models to analyze freeway crash frequency. *Journal of Safety Research*, 36, 365-375.

Chang, L. Y., Wang, H. W. (2006) Analysis of traffic injury severity: An application of non-parametric classification tree techniques. *Accident Analysis and Prevention*, 38, 1019-1027.

Conerly, M., Gray, B., Kevin, B., Edward, M. (2000) Data mining and visualization of the Alabama crash database. *Technical Report*, University of Alabama, Transportation Center, Tuscaloosa.

Department of Statistics, MOTC. (2010) Analysis of traffic crash for the Taiwan Railways System. *Annual Report*, MOTC, Taiwan.

Greene, W. H. (2000) *Econometric Analysis*. Prentice Hall, New Jersey.

Haughton, D., Oulabi, S. (1993) Direct marking modeling with CART and CHAID. *Journal of Direct Marketing*, 7(3), 16-26.

Hu, S. R., Li, C. S., Lee, C. K. (2010) Investigation of Key Factors for Crash Severity at Railroad Grade Crossings by Using a Logit Model. *Safety Science*, Vol. 48, No. 2, pp. 186-194.

Hu, S. R., Li, C. S., Lee, C. K. (2011) Assessing Casualty Risk of Railroad-grade Crossing Crashes Using Zero-inflated Poisson Models. *Journal of Transportation Engineering*, 137(8), 527-536.

Hu, S. R., Li, C. S., Lee, C. K. (2012) Model crash frequency at highway–railroad grade crossings using negative binomial regression. *Journal of the Chinese Institute of Engineers*, 35(7), 841-852.

Laffey, S. (2010) International comparison of level crossing collisions. *TRB AHB60 Technical Report*, Transportation Research Board, National Research Council, Washington, D. C.

Lee, J., Nam, D., Park, D. (2005) Analyzing the relationship between grade crossing element and crash. *Journal of the Eastern Asia Society for Transportation Studies* (EASTS), 6, 3658-3668.

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.

Lord, D., Washington, S. P., Ivan, J. N. (2007) Further notes on the application of zero-inflated models in highway safety. *Accident Analysis and Prevention*, 39(1), 53-57.

Park, Y. J., Saccomanno, F. F. (2005) Collision frequency analysis using tree-based stratification. *Transportation Research Record: Journal of the Transportation Research Board*, 1918, 1-9.

Saccomanno, F. F., Fu, L., Miranda-Moreno, L. F. (2004) Risk-based model for identifying highway-rail grade crossing blackspots. *Transportation Research Record: Journal of the Transportation Research Board*, 1862, 127-135.

Yan, X., Richards, S., Su, X. (2010) Using hierarchical tree-based regression model to predict train-vehicle crashes at passive highway-rail grade crossings. *Accident Analysis and Prevention*, 42(1), 64-74.