applied sciences

MDPI

Article

Analysis of the Railway Accident-Related Damages in South Korea

Man Sik Park [1], Jin Ki Eom [2,*], Jungsoon Choi [3,*] and Tae-Young Heo [4,*]

[1] Department of Statistics, Sungshin Women's University, 2 Bomun-ro 34da-gil, Seongbuk-gu, Seoul 02844, Korea; mansikpark@sungshin.ac.kr
[2] Railway Policy Research Team, Korea Railroad Research Institute, Uiwang, Gyeonggi 16105, Korea
[3] Department of Mathematics, College of Natural Sciences, Hanyang University, Seoul 04763, Korea
[4] Department of Information & Statistics, Chungbuk National University, 1 Chungdae-ro, Seowon-Gu, Cheongju Chungbuk 28644, Korea
* Correspondence: jkom00@krri.re.kr (J.K.E.); jungsoonchoi@hanyang.ac.kr (J.C.); theo@chungbuk.ac.kr (T.-Y.H.)

Received: 13 November 2020; Accepted: 4 December 2020; Published: 8 December 2020

Abstract: Railway accidents are critical issues characterized by a large number of injuries and fatalities per accident due to massive public transport systems. This study proposes a new approach for evaluating the damages resulting from railway accidents using the two-part models (TPMs) such as the zero-inflated Poisson regression model (ZIP model) and the zero-inflated negative-binomial regression model (ZINB model) for the non-negative count measurements and the zero-inflated gamma regression model (ZIG model) and the zero-inflated log-normal regression model (ZILN model) for the semi-continuous measurements. The models are employed for the evaluation of the railway accidents on Korea Railroad, considering the accident damages, such as the train delay time, the number of trains delayed and the cost of considering the accident count responses, for the period 2008 to 2016. From the results obtained, we found that the human-related factors, the high-speed railway system or the Korea Train Express (KTX) and the number of casualties, are the main cost-escalating factors. The number of trains delayed and the amount of delay time tend to increase both the probability of incurring costs and the amount of cost. For better evaluation, the railway accident data should contain accurate information with less recurrence of zeros.

Keywords: railway; accidents; damages; evaluation; two-part models

1. Introduction

The evaluation factors of the railway accidents are represented by the severity of injuries, the fatalities, the damage of rolling stock and the associated infrastructure, and the environmental cost. Therefore, the evaluation of railway safety is mainly related to the most common factors that have an impact on accidents and their significance to the severity of the injury caused [1]. The cost of trains getting delayed, which is required for handling accidents, are commonly included in the transport network cost [2,3]. In order to evaluate and quantify the costs of railway accidents, conventionally, the accident data is analyzed by using the aggregated cost of a long-term period, since railway accidents are not as frequent as road accidents.

The studies for evaluating railroad accident-related damages have not been conducted as much as road accident studies. This is because road accidents happen more often and plenty of data is available to the public. Railway accidents rarely occur and access to the data is generally restricted by the railway operators. Due to the limited number of studies on railway accidents, some studies are related to the railway accident prediction models [4–10]. These studies analyzed accident factors and constructed a reliable accident prediction model for the prevention of the railway accidents.

The factors for predicting accidents consist of human, rolling stock, facilities, and operational factors [4–6]. In particular, the analysis of accidents occurring at railroad-crossing related to facilities research was dominant [7–9]. These studies have explored factors which have an impact on accidents occurring between vehicles and trains, such as the number of train tracks, the number of highway lanes, train and traffic volumes, train and vehicle speeds, site and surface characteristics, road/rail-side appurtenances and so forth. Based on the accident factors, these studies introduced a logistic regression model for the occurrence/non-occurrence accident data and a Poisson regression model for the number of accident data.

Unlike accident prediction models, accident evaluation models have to consider the non-negative and zero-inflated nature of accident damage data as dependent measures like the train delay time, the number of trains delayed and the cost of considering the accident count responses. Even if an accident occurs, there are quite a lot of cases where casualties, train delays and accident costs do not occur. These cases are recorded as 'zero' in accident damage data. Therefore, statistical models for accident evaluation should be applied differently from existing accident prediction models.

In order to introduce appropriate statistical models for railway accident evaluation, we review the literature related to the statistical models implemented on railway accident data and the models dealing with non-negative and the zero-inflated nature of data. Then, we propose reliable statistical models and implement them to the train delay time, the number of trains delayed and the cost by using railway accident data observed in Korea. The appropriate accident evaluation models can accurately assess the damages caused by railroad accidents and accordingly, the railroad operator can reasonably establish a plan for necessary actions to reduce the accident cost in the future.

2. State of Art Statistical Models

With respect to the statistical accident modeling efforts, some models were developed earlier for predicting accidents, based on a multiple linear regression model [10]. Since then, the statistical models applied to railway accident prediction have become more sophisticated by introducing categorical data analysis along with accident severity data in the form of logistics. Hu et al. [5] developed a generalized logit model to determine the categorical characteristics of accident severity on railroad grade crossing. Different levels of injuries in accidents are modeled by either ordered logit or probit models [11–13]. By comparing various logit model structures such as the ordered probit, the multinomial logit and the random parameter logit model, Zhao and Khattak [14] showed that the random parameter logit model was the most suitable to evaluate the severities of injuries in railway level crossing accidents.

Due to the random, discrete and non-negative nature of accident data, the models such as the Poisson regression model and the negative-binomial regression models [4] were widely used, instead of the linear regression models. However, the two heterogeneous distributions of measurements (for example, zero or positive) of accident data have not properly been explained by any classical models. In order to handle the excessive frequencies of zero in railway accident data, with no injuries and fatalities, the two-part models (TPMs) emerged as a solution, which consists of a degenerate distribution at zero and a non-zero distribution otherwise. Until now, a few studies suggested a zero-inflated negative-binomial and the zero-inflated Poisson model [15–17].

According to the non-zero distribution of a random component (or dependent variable), TPMs are of two types; 1) zero-inflated regression models for non-negative count data and 2) zero-inflated regression models for semi-continuous data; for example, zero-inflated Poisson model (ZIP) and zero-inflated negative-binomial model (ZINB) belong to the former type and zero-inflated gamma regression model (ZIG) and zero-inflated log-normal model (ZILN) belong to the latter type.

There has been a lot of research on the two-part models. Lambert [18] applied ZIP to predict the number of defects in manufacturing. Ridout et al. [19] reviewed contemporary statistical models for count data with excessive zeros. Joe and Zhu [20] compared generalized Poisson models with ZINB. Mwalili et al. [21] contributed to significant correction for misclassification in caries research by using ZINB. Neelon et al. [22] proposed a Bayesian model for zero-inflated count data with an

analysis of the psychiatric outpatient service use. Neelon et al. [23] summarized TPMs for non-negative count measurements and semi-continuous data. Kern and Wasser [24] considered ZIG to analyze health care costs including a large proportion of $0 data. Nobre et al. [25] analyzed time spent on leisure time physical activity using ZIG. Risio et al. [26] applied ZILN for the Prosopis caldenia pod production data at tree level in the Argentinean semiarid pampas. Tong et al. [27] suggest a zero-adjusted gamma model for mortgage loan loss given default contained extensive numbers of zeroes. Neelon et al. [28] employed ZIP model with spatial effects to examine emergency department visits. Ghosh et al. [29] proposed the Bayesian modeling approach for fitting zero-inflated regression model. Bayesian approaches for modeling semi-continuous data are proposed by References [30–32].

This study proposes a new approach for evaluating the damages resulting from railway accidents using the TPMs such as the ZIP model and the ZINB model for the non-negative count measurements and the ZIG model and the ZILN model for the semi-continuous measurements. These models consist of a degenerate distribution at zero and a non-zero distribution using railway accident data. For real application, we extracted all the recorded accidents data for the period 2008 to 2016 obtained from the Korea Transportation Safety Authority (KOTSA) with respect to the train types (urban train, general train and high-speed train), the organization types (metro and national railway), the accident factors (non-human related and human-related) and the accident types (Traffic, Safety, Misc., rolling stock). Then, we employed the statistical models to identify the independent variables that are highly correlated with the accident types and the magnitude of accidents reflecting train delay and causalities.

3. Two-Part Model

The two-part model framework provides an appropriate structure for modeling two types of data—the non-negative count data and the semi-continuous data [19,21]. Let Y_i be the non-negative (or semi-continuous) outcome for subject i and $\boldsymbol{x}_i = \left(X_{i1}, X_{i2}, \cdots X_{ip}\right)^T$ be a vector of covariates for subject i; let $\boldsymbol{\beta} = \left(\beta_1, \beta_2, \cdots \beta_p\right)^T$ be the parameter vector used in modeling the probability of positive responses and let $\boldsymbol{\theta} = \left(\theta_1, \theta_2, \cdots \theta_q\right)^T$ represent the mean and the dispersion parameters of the conditional distribution of the positive responses. The two parts of the model can contain different sets of covariates but we have assumed that they are the same and use $\{\boldsymbol{x}_i,\ i = 1, 2, \cdots n\}$ for both parts of the model; hence, $p = q$. Thus, the probability of a positive response can be denoted as $p_i = P(Y_i > 0 \big| \boldsymbol{x}_i, \boldsymbol{\beta})$ and the conditional distribution of the positive responses can be represented as $g_{\boldsymbol{\theta}}\left(y_i \big| y_i > 0, \boldsymbol{x}_i\right)$. Therefore, the distribution of the response, $g_{\boldsymbol{\theta}}(y_i \big| \boldsymbol{x}_i)$ is expressed as

$$g_{\boldsymbol{\theta}}(y_i \big| \boldsymbol{x}_i) = g_{\boldsymbol{\theta}}(y_i \Big| y_i = 0, \boldsymbol{x}_i) \times \mathbf{I}(Y_i = 0) + g_{\boldsymbol{\theta}}\left(y_i \big| y_i > 0, \boldsymbol{x}_i\right) \times (1 - \mathbf{I}(Y_i = 0)),$$

where the indicator function, $\mathbf{I}(Y_i = 0)$ is defined such that if $Y_i = 0$ then, $\mathbf{I}(Y_i = 0) = 1$, else, $\mathbf{I}(Y_i = 0) = 0$. Thus, this framework results in the following mixed probability function and likelihood: for $i = 1, 2, \cdots, n$,

$$f(y_i \big| \boldsymbol{x}_i, \boldsymbol{\beta}, \boldsymbol{\theta}) = \left[(1 - p_i) + p_i \cdot g_{\boldsymbol{\theta}}(y_i \big| y_i = 0, \boldsymbol{x}_i)\right]^{\mathbf{I}(y_i = 0)} \times \left[p_i \cdot g_{\boldsymbol{\theta}}\left(y_i \big| y_i > 0, \boldsymbol{x}_i\right)\right]^{1 - \mathbf{I}(y_i = 0)}.$$

$$L(\boldsymbol{\beta},\ \boldsymbol{\theta} \big| \boldsymbol{y},\ \boldsymbol{X}) = \left[\prod\nolimits_{y_i = 0} (1 - p_i) + p_i \cdot g_{\boldsymbol{\theta}}(y_i \big| y_i = 0, \boldsymbol{x}_i)\right]\left[\prod\nolimits_{y_i > 0} p_i \cdot g_{\boldsymbol{\theta}}\left(y_i \big| y_i > 0, \boldsymbol{x}_i\right)\right]. \quad (1)$$

Here, $\boldsymbol{y} = (Y_1, Y_2, \cdots, Y_n)^T$ and $\boldsymbol{X}$ is a covariate matrix with a size of $n \times p$. In case of non-negative count responses, $g_{\boldsymbol{\theta}}(y_i \big| y_i = 0, \boldsymbol{x}_i)$ has its own positive probability whereas the distribution of semi-continuous measurements is $g_{\boldsymbol{\theta}}\left(y_i \big| y_i > 0, \boldsymbol{x}_i\right) = g_{\boldsymbol{\theta}}(y_i \big| \boldsymbol{x}_i)$. Equation (1) can be factored into two parts: one related

to the $\boldsymbol{\beta}$ parameter vector involved in estimating p_i's and the other with only the parameters involved in estimating the $\boldsymbol{\theta}$ parameter vector. The first part (logit part) for estimating $\boldsymbol{\beta}$ is expressed as follows:

$$\text{logit}(p_i) = \log\left(\frac{p_i}{1-p_i}\right) = \boldsymbol{x}_i{}^T\boldsymbol{\beta}.$$

In this study, we consider two different TPMs in terms of the measurement types of the dependent variable. In Section 3.1, the TPM for non-negative count measurements is introduced, especially ZIP and ZINB models. In Section 3.2, the TPM for non-negatively continuous measurements is introduced, including ZIG and ZILN models.

3.1. TPMs for Non-Negative Count Measurements

In general, Poisson regression analysis is the primary model which is used to find out the causal relationship between the independent variables and the dependent variable. The dependent variable must distribute as Poisson distribution. One of the assumptions of Poisson regression is that the mean and the variance are equal but most of the data will have a larger variance or overdispersion. Poisson distribution is a representative model for count responses, for example, the count of defective products in a manufacturing process or the number of visits to the hospital. We frequently encounter excessive zeros, which is more than that can be handled under a Poisson distribution [18,19]. For example, the number of trains delayed due to railway accidents in this study had 24% zero measurements. To counter this problem, we construct a model with alternative discrete distributions. The ZIP (zero-inflated Poisson) distribution can be used to model the count responses having excessive zeros. Lambert [18] first introduced the ZIP model in terms of mixed distribution, where one distribution is the point mass at zero with a probability weight of $1-p_i$ and the other is a Poisson distribution with the mean rate, λ_i and weight, p_i. The expectation maximization algorithm of Dempster et al. [28] is generally used to find the maximum likelihood estimation of the ZIP model.

In this section, we briefly explain the ZIP distribution along with its property. Let Y denote the count response variable that follows a mixture of the two distributions, the perfect one with zero state degenerating at 0 and the other with Poisson distribution denoted as $Poi(\lambda_i)$; with probabilities $(1-p_i)$ and p_i, respectively, where $0 \le p_i \le 1$. The probability distribution of Y_i's can be expressed as

$$f(y_i|\boldsymbol{x}_i, \boldsymbol{\beta}, \boldsymbol{\theta}) = \begin{cases} (1-p_i) + p_i \cdot e^{-\lambda_i}, & y_i = 0, \\ p_i \cdot \frac{\lambda_i{}^{y_i} e^{-\lambda_i}}{y_i!}, & y_i = 1, 2, \cdots. \end{cases} \quad (2)$$

As can be seen from Equation (2), ZIP distribution has an inflated probability of zeros by the amount $(1-p_i)\left(1-e^{-\lambda_i}\right)$ compared to $e^{-\lambda_i}$. Using the moment generating function of the ZIP, which is of the form $E\left(e^{tY_i}\right) = (1-p_i) + p_i \cdot e^{\lambda_i(e^t-1)}$, we can easily find the mean and variance of Y_i, $E(Y_i) = \lambda_i p_i$ and $V(Y_i) = \lambda_i p_i[1 + (1-p_i)\lambda_i]$. The higher value of the variance compared with the mean denotes an overdispersion in count data.

Based on the second part (Poisson part) of the Equation (2), the systematic component and its link function is represented as follows:

$$\log(\lambda_i) = \boldsymbol{x}_i{}^T\boldsymbol{\theta}.$$

Given a dataset $\{(y_i; \boldsymbol{x}_i)\}$ of size n, we can write the log-likelihood function, $l(\cdot)$ of the ZIP model as

$$l(\boldsymbol{\beta}, \boldsymbol{\theta}|\boldsymbol{y}, \boldsymbol{X}) = \sum\nolimits_{y_i=0} log\left[(1-p_i) + p_i \cdot e^{-\lambda_i}\right] + \sum\nolimits_{y_i>0} log\left(p_i \cdot \frac{\lambda_i{}^{y_i} e^{-\lambda_i}}{y_i!}\right).$$

We have explored analysis methods for the non-negative count responses containing excessive zeros. This type of data can be modeled through a two-step process which involves modeling the probability of a non-zero outcome and modeling the mean of the non-zero outcomes.

The ZINB distribution is also used for count responses in modeling the overdispersion with excessive zeros. The ZINB model is almost the same as the ZIP except for the probability distribution of Y_i's, which is expressed as

$$f(y_i|\boldsymbol{x}_i, \boldsymbol{\beta}, \boldsymbol{\theta}) = \begin{cases} (1-p_i) + p_i \cdot \left(\frac{\delta}{\lambda_i+\delta}\right)^{\delta}, & y_i = 0, \\ p_i \cdot \frac{\Gamma(y_i+\delta)}{\Gamma(y_i+1)\Gamma(\delta)} \left(\frac{\delta}{\lambda_i+\delta}\right)^{\delta} \left(\frac{\lambda_i}{\lambda_i+\delta}\right)^{y_i}, & y_i = 1, 2, \cdots, \end{cases}$$

and the log-likelihood function,

$$l(\boldsymbol{\beta},\ \boldsymbol{\theta}|\boldsymbol{y},\ \boldsymbol{X}) = \sum_{y_i=0} log\left[(1-p_i) + p_i \cdot \left(\frac{\delta}{\lambda_i+\delta}\right)^{\delta}\right] + \sum_{y_i>0} log\left(p_i \cdot \frac{\Gamma(y_i+\delta)}{\Gamma(y_i+1)\Gamma(\delta)} \left(\frac{\delta}{\lambda_i+\delta}\right)^{\delta} \left(\frac{\lambda_i}{\lambda_i+\delta}\right)^{y_i}\right),$$

where $E(Y_i) = \lambda_i p_i$ and $V(Y_i) = \lambda_i p_i[1 + (1-p_i)\lambda_i + \lambda_i/\delta]$. If $\delta \to \infty$ and $p_i \to 0$, then the ZINB reduces to the Poisson regression model.

3.2. TPMs for Semi-Continuous Measurements

When a variable is non-negatively continuous and has excessive zeros, we regard it as semi-continuous. This type of data is frequently observed in economics, climatology, microbiology, medical applications and so on. In this study, we introduce the TPMs for semi-continuous measurements, which are the ZIG and ZILN models.

The ZIG model uses a gamma regression with a log link function to model the non-zero values. Semi-continuous data can be modeled in two parts: one part (logit part) consisting of the probability of a non-zero value and the other part (gamma part) consisting of the distribution of the continuous non-zero values.

The ZIG likelihood follows the format of the Equation (1), where, $\text{logit}(p_i) = \boldsymbol{x_i}^T\boldsymbol{\beta}$ and $g_{\boldsymbol{\theta}}(y_i|\boldsymbol{x}_i, \mu_i, \nu)$, such that

$$g_{\boldsymbol{\theta}}(y_i|\boldsymbol{x}_i, \mu_i,\ \nu) = \frac{1}{\Gamma(\nu)} \left(\frac{\nu}{\mu_i}\right)^{\nu} \mathrm{y}_i^{\nu-1} exp\left(-\frac{\nu y_i}{\mu_i}\right),$$

where y_i modeled as $log(\mu_i) = \boldsymbol{x_i}^T\boldsymbol{\theta}$ and ν is the dispersion parameter [17]. This leads to the following likelihood function:

$$L(\boldsymbol{\beta},\ \boldsymbol{\theta}|\boldsymbol{y},\ \boldsymbol{X},\ \boldsymbol{\mu},\ \nu) = \left[\prod_{y_i=0}(1-p_i)\right]\left[\prod_{y_i>0} p_i \cdot g_{\boldsymbol{\theta}}(y_i|\boldsymbol{x}_i, \mu_i, \nu)\right]. \quad (3)$$

As can be seen in the Equation (3), the ZIG likelihood is factorable into one part with $\boldsymbol{\beta}$ and the other part with $\boldsymbol{\theta}$ and ν. Maximizing each of the parts separately will lead to the maximization of the overall likelihood. This can be performed via the Newton-Raphson algorithm for each part.

ZILN regression also follows the Equation (1). The difference between the two models is the type of the continuous distribution, $g_{\boldsymbol{\theta}}(\cdot)$. For the ZILN regression, $g_{\boldsymbol{\theta}}\left(y_i|\boldsymbol{x}_i, \mu_i, \sigma^2\right)$ is defined as

$$g_{\boldsymbol{\theta}}\left(y_i|\boldsymbol{x}_i, \mu_i, \sigma^2\right) = \frac{1}{y_i\sqrt{2\pi\sigma^2}} exp\left(-\frac{(\log y_i - \mu_i)^2}{2\sigma^2}\right).$$

Here, σ^2 is the population variance of the natural log of the data. So, its likelihood is expressed by

$$L\left(\boldsymbol{\beta},\ \boldsymbol{\theta}|\boldsymbol{y},\ \boldsymbol{X},\ \boldsymbol{\mu},\ \sigma^2\right) = \left[\prod_{y_i=0}(1-p_i)\right]\left[\prod_{y_i>0} p_i \cdot g_{\boldsymbol{\theta}}\left(y_i|\boldsymbol{x}_i, \mu_i, \sigma^2\right)\right].$$

4. Real Application

Table 1 summarizes Korea's railroad statistics for the year 2018, from the Ministry of Land, Infrastructure and Transport (MOLIT) in Korea. There are three types of railways in operation:

The Korea Train Express (KTX) and the general railroads, operated by KORAIL; the Korean National Railroad; and 15 routes in 9 cities operated by 11 Metro private sectors. The high-speed rail has a total length of 643 km on three routes, with a speed of 300 km/h and the general railway has 107 lines in total, with a speed of 60–100 km/h, of which 52 lines are operated for passenger transportation. With respect to the annual passenger transport statistics, Metros had a total of 3618 million passengers and the high-speed railway carried 297.6 million passengers, while the general railway carried, merely, a total of 92.1 million passengers due to its low frequency of passenger services.

Table 1. Description of railway statistics in Korea (2018).

Type	Descriptions	Operated Speed (km/h)	Length (km)	No. Lines	Passenger (Million/Year)
KTX (Korea Train Express)	High-speed railway (operated by KORAIL)	300	643	3	297.6
General Railway	Regional railway (operated by KORAIL)	60~100	3492	107 (52 *)	92.1
Urban railway	Metro (11 operators)	20~30	704	15	3618.3

Note: (*) represents the number of railway lines for passenger transport service.

We analyzed a railway accident dataset in South Korea for the period from 2008 to 2016, which was obtained from the KOTSA. There were 5051 railway accidents during the mentioned period. Figure 1 presents the significantly decreasing temporal pattern of the number of railway accidents from 2008 to 2016. The number of accidents were 704 and 360 in 2008 and 2016, respectively, representing almost 50% decrease in accidents in 9 years. In the Figure 1, the number of accidents over the years were compared by the railroad types (KTX, Urban and General) and it was found that general railway-related accidents dramatically decreased from 2008 to 2016, compared with other train types.

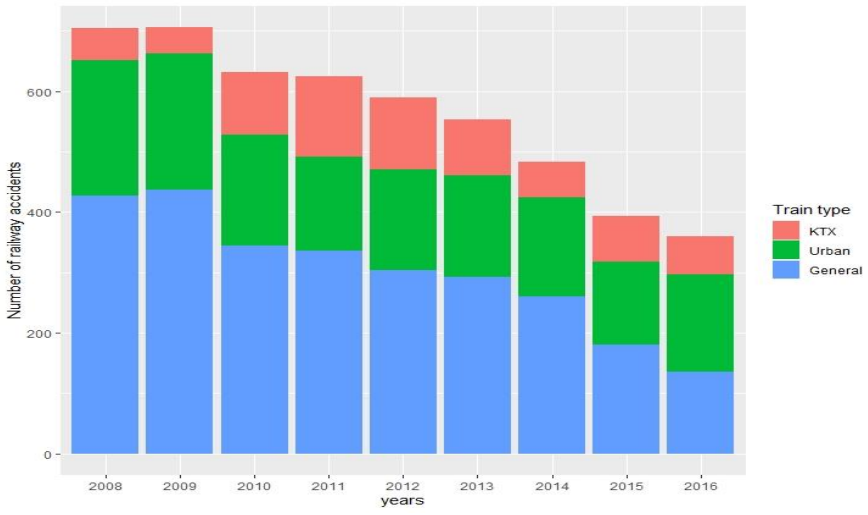

Figure 1. Time series plot of the number of railway accidents (2008–2016).

According to MOLIT, the reasons for the reduction of railway accidents are in three folds: (1) continuous railway facility improvements such as double tracks, electrification, new rolling stocks, modernization of maintenance equipment and so forth, (2) investments in the expansion of safety facilities such as three-dimensional crossing, installation of platform screen-doors (PSDs) and safety fences along the roads, (3) reinforcement of education related to accident prevention and training for railroad employees. These three actions have strengthened the overall railroad safety management in the form of systematic establishment and maintenance of the railroad safety system and the implementation of periodic facility safety inspections.

Figure 2 shows the map of the number of railway accidents in 17 provinces of South Korea. There were 892 accidents in Seoul, the capital of South Korea within the study period and 857 accidents in Gyeonggi province, the area surrounding the capital. Thus, the total number of railway accidents in Seoul and Gyeonggi province was 1749, which is about 34.6% of all accidents. The frequencies of all KTX and the majority of urban railway lines have increased dramatically in this area in order to meet the rising travel demand. As a result, train accidents are concentrated, mainly in the metropolitan area. Therefore, it can be concluded that the number of accidents escalate with rise in train frequency.

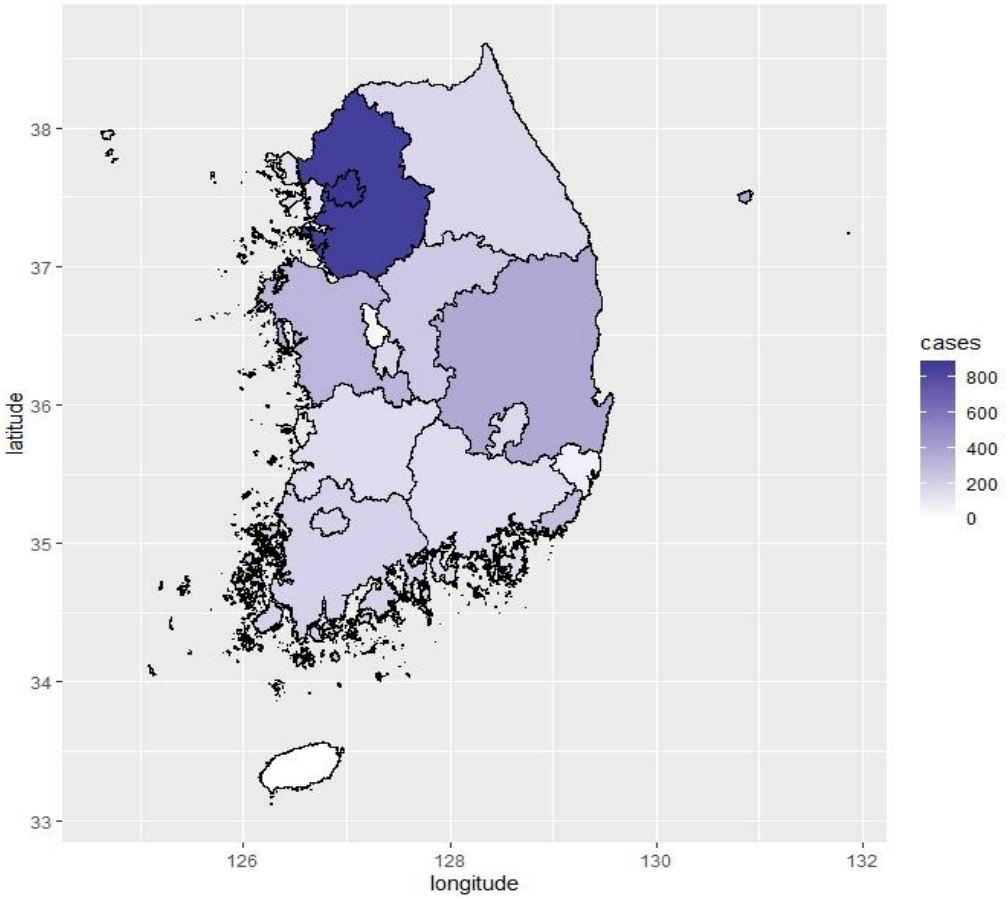

Figure 2. Map of the number of railway accidents.

In this study, we considered the following three damage variables from the railway accidents: number of delayed trains, delay time and amount of costs incurred in handling the accidents. The covariates related with the accidents are year, organization, train type, accident factor, accident type, the number of fatalities (casualties) and the number of derailed trains. These variables were given from the KOTSA dataset. For example, the accident factor was already defined as human and non-human factors. The human factors are the accidents caused by passenger or driver. The non-human factors are defined as the accidents are caused by all factors other than human such as weather, traffic, nal failure, rolling-stock malfunctioned and so forth. Table 2 presents the list of variables used in the analysis.

Table 2. Description of the Variables in Korea Transportation Safety Authority (KOTSA) Railway Accidents Dataset.

Type	Information	Description
Dependent Variables	No. of Delayed Trains	The number of trains delayed due to railway accidents
	Delay Time	Train delay time due to railway accidents (minute)
	Amount of Costs	Amount of costs due to railway accidents (million won)
Independent Variables	Year	Year (2008–2016)
	Organization	KORAIL (88.8%), Metro (11.2%)
	Railroad Type	KTX (14.7%), Urban (31.4%), General (53.8%), missing (0.1%)
	Accident Factor	Human-related (51.6%), Non-human-related (48.4%)
	Accident Type	Traffic (30.8%), Safety (13.8%), Misc. (8.71%), rolling stock (46.7%)
	No. of Derailed Trains	The number of derailed trains due to railway accidents
	No. of Fatalities	The number of fatalities due to railway accidents
	No. of Casualties	The number of casualties due to railway accidents

From the descriptive statistics of the railway accident data, it is found that about 88.8% of railway accidents occur in the railways operated by KORAIL, which dominate the railroad service network. As the accident factor, the human-related accidents and non-human-related accidents were 2606 (51.6%) and 2445 (48.4%), respectively. The railroad type, which was categorized into high-speed railway, urban railway and general railway were 740 (14.7%), 1587 (31.4%) and 2718 (53.8%), respectively. The accident type is divided into four groups: traffic accident, safety-related accident, rolling-stock related accident and miscellaneous accident. The number of traffic accidents were 1557 (30.8%), which can be classified into collisions between, vehicle(s) and trains at a railway-crossing, passenger(s) and trains and road worker(s) and trains. The number of safety related accidents were 697 (13.8%), which occurred at railway facilities around the station, the platform and the train. The number of rolling-stock related accidents were 2357 (46.7%) that happened due to fire in the train, train collision and derailment. The number of miscellaneous accidents were 440 (8.7%) which are not included in the previous three categories. We considered the number of fatalities, casualties and derailed trains for modeling and evaluating the impact of railway accidents.

For the real application of the railway accident data, we used the SAS software (Version 9.4, SAS Institute Inc., Cary, NC, USA.) for modeling TPMs and the R software (Version 3.6.3, R Foundation for Statistical Computing, Vienna, Austria) for visualizing the data.

The continuous dependent variables in this study are delay time (Figure 3) and the amount of costs incurred (Figure 4) due to railway accidents. As can be seen in Figures 3 and 4, 22.5% (1139) and 79.8% (4029) of the sample size were recorded as no time delay and no cost, respectively. The histograms in the figures show that the positive measurements for each variable are extremely right-skewed and, thus we considered the gamma distribution and the log-normal distribution as good fits. The five-number summary (minimum, 25th percentile, median, 75th percentile, maximum) for delay time is (0, 10, 22, 37 and 1432). For the positive measurements (20.2%) of the cost amount, the five-number summary is expressed as (0.001, 0.434, 1.724, 6.6 and 13,933). No probability distribution of positive real-value

random variables includes zero as one of the possible realizations. Therefore, TPMs for semi-continuous data are essential to build both logit part for zero measurements and the gamma (or log-normal) part for non-zero (or positive) measurements.

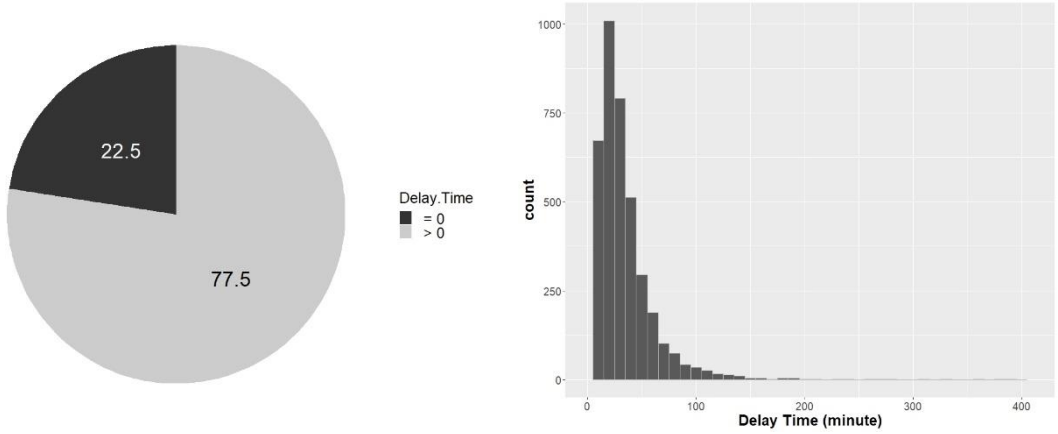

Figure 3. Distribution of Delay Time (the measurements above 400 min (0.6%) are not displayed in the histogram).

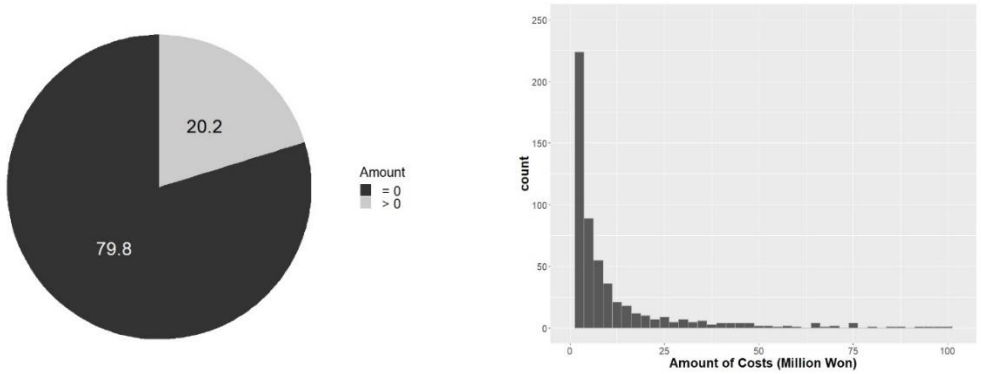

Figure 4. Distribution of Amount of Costs (the measurements above 100 million won (0.6%) are not displayed in the histogram).

Table 3 shows the frequencies of each of the categorized non-negative count variables such as the number of fatalities (casualties), the number of derailed trains and the number of trains delayed due to the railway accidents. All the variables in Table 3 have a lot of zero measurements (24.0–99.2%). Hence, zero-inflated regression model for the variables is applied for the number of delayed trains.

Table 3. Contingency Table for Categorized Discrete Variables.

Discrete Variables	0	1	2	3	4	5	6–10	11–20	21–50	50–
No. of Fatalities	4070 −80.6	965 −19.1	15 −0.3	-	-	1 −0.02				
No. of Casualties	2847 −56.4	2126 −42.1	44 −0.87	12 −0.24	7 −0.14	2 −0.04	6 −0.12	3 −0.06	3 −0.06	1 −0.02
No. of Derailed Trains	5008 −99.2	19 −0.38	12 −0.24	2 −0.04	-	1 −0.02	8 −0.16	-	1 −0.02	
No. of Delayed Trains	1211 −24	1950 −38.6	546 −10.8	338 −6.69	210 −4.16	168 −3.33	322 −6.37	177 −3.5	110 −2.18	19 −0.38

The values are expressed as number of cases (percentage).

Table 4 illustrates the regression analyses of the number of trains delayed due to the railway accidents with some of the independent variables shown in Table 2. In general, positive estimates increase the probability of having more trains delayed (or the number of trains delayed) but negative estimates decrease them. As can be seen in Table 4, the estimation results are quite similar irrespective of the models considered. For example, more trains are estimated to be delayed during the period of our interest due to the railway accidents as the corresponding estimates are positive (0.063 in ZIP; 0.066 in ZINB) and KORAIL has a smaller number of trains delayed than the other railway organizations in that its associated estimates (−0.628 in ZIP; −0.891 in ZINB) are negative. Human-related accidents resulted in less trains delayed compared to the non-human related accidents because of the negative estimate (−0.165 in ZIP; −0.706 in ZINB). We can also see that the number of derailed trains has a positive impact (0.063 in ZIP; 0.442 in ZINB) on the number of delayed trains. The ZIP model shows that the KTX (0.005) does not cause more delayed trains under the significance level of 0.05 but the urban trains (0.646) result in more delayed time than the general trains. We can also see that number of fatalities increases the number of delayed trains. The interpretation also works for the ZINB model. The logit part in the ZIP model has statistically significant effect of the accident factor (human-related factor vs. non-human-related factor). However, none of the independent variables are of significance in the logit part of ZINB model. In terms of Akaike Information Criterion Corrected (AICC) [24], the ZINB model outperforms the ZIP model.

As mentioned earlier, we employed two different types of the TPM; (1) ZIP model and ZINB model for discrete variables such as number of delayed trains (Table 4) due to railway accidents and (2) ZIG model and ZILN model for semi-continuous variables. Tables 5 and 6 illustrate the statistical results from the analyses of the TPMs for the two semi-continuous dependent variables: delayed time and the amount of costs due to the railway accidents.

Table 5 displays the statistical results of the regression analysis on the amount of delayed time. For every model, both the logit part and the positively continuous part (Gamma part or Log-normal part) have the statistically significant coefficient estimates. Especially, the coefficient estimates in logit parts are quite similar with each other. The possibility that a railway accident causes any time delay increases when the accident is involved in KORAIL (0.580) compared to Metro. Human-related accidents (−2.777) and number of fatalities (1.810) have meaningful impact on time delay resulting from a railway accident. In case of the other part, the independent variables such as year and number of casualties are only significant in the Gamma part though number of trains derailed due to railway accidents are not significant in either of the two models. KORAIL (−1.141 in ZIG; −0.182 in ZILN) tends to be have lesser delayed time and a human-related accident (−0.110 in ZILN) has a lower probability of having a positive delayed time than a non-human-related accident. Train types such as KTX (−0.582 in ZIG; −0.597 in ZILN) and urban trains (−0.341 in ZIG; −0.405 in ZILN) result in less delayed time than general trains. Noteworthy in this case is that higher the number of casualties (0.099) due to the railway accidents the longer the hours of train delay, especially in the ZIG model. From the comparison between the two TPMs, the ZILN model slightly outperforms the ZIG model. This is because the log-likelihood value of the former is larger and its AICC is smaller.

Table 4. Two-Part Regression Analyses of Number of Trains delayed due to Railway Accidents.

Independent Variables	Zero-Inflated Poisson Regression						Zero-Inflated Negative-Binomial Regression					
	Logit Part			Poisson Part			Logit Part			Negative Binomial Part		
	Est.	**S.E.**	***p*-Value**	**Est.**	**S.E.**	***p*-Value**	**Est.**	**S.E.**	***p*-Value**	**Est.**	**S.E.**	***p*-Value**
Year (2008–2016)	0.004	0.016	0.827	0.063	0.003	<0.001	0.012	0.047	0.786	0.066	0.008	<0.001
Organization (vs. Metro)												
KORAIL	0.004	0.016	0.827	−0.628	0.023	<0.001	−0.002	0.091	0.998	−0.891	0.073	<0.001
Accident Factor (vs. Non-human-related)												
Human-related	−0.955	0.107	<0.001	−0.165	0.022	<0.001	0.512	0.278	0.065	−0.706	0.045	<0.001
Train Type (vs. General)												
KTX				−0.037	0.027	0.177				0.005	0.056	0.936
Urban				0.702	0.020	<0.001				0.646	0.045	<0.001
No. of Fatalities				0.024	0.024	0.328				0.460	0.054	<0.001
No. of Trains Derailed				0.063	0.009	<0.001				0.442	0.095	<0.001
Log-likelihood					−18,647.2						−10,706.0	
AICC					37,318.4						21,438.0	

Est., Estimate; S.E., standard error; AICC, Akaike Information Criterion Corrected (Hurvich and Tsai [33]).

Table 5. Two-Part Regression Analyses of Delay Time (Minute) due to Railway Accidents.

Independent Variables	Zero-Inflated Gamma Regression						Zero-Inflated Log-Normal Regression					
	Logit Part			Gamma Part			Logit Part			Log-Normal Part		
	Est.	**S.E.**	***p*-Value**	**Est.**	**S.E.**	***p*-Value**	**Est.**	**S.E.**	***p*-Value**	**Est.**	**S.E.**	***p*-Value**
Year (2008–2016)				−0.017	0.005	<0.001				−0.001	0.004	0.826
Organization (vs. Metro)												
KORAIL	0.580	0.114	<0.001	−1.141	0.048	<0.001	0.580	0.114	<0.001	−0.182	0.026	<0.001
Accident Factor (vs. Non-human-related)												
Human-related	−2.777	0.099	<0.001	−0.048	0.031	0.122	−2.777	0.099	<0.001	−0.110	0.026	<0.001
Train Type (vs. General)												
KTX				−0.582	0.033	<0.001				−0.597	0.031	<0.001
Urban				−0.341	0.031	<0.001				−0.405	0.028	<0.001
No. of Casualties				0.099	0.021	<0.001				0.019	0.015	0.208
No. of Fatalities	1.810	0.103	<0.001				1.810	0.102	<0.001			
No. of Trains Derailed				0.061	0.038	0.102				0.042	0.022	0.055
Log-likelihood					−19,770.7						−19,144.6	
AICC					39,567.5						38,315.2	

Est., Estimate; S.E., standard error; AICC, Akaike Information Criterion Corrected (Hurvich and Tsai [33]).

Table 6. Two-Part Regression Analyses of Amount of Costs due to Railway Accidents.

Independent Variables	Zero-Inflated Gamma Regression						Zero-Inflated Log-Normal Regression					
	Logit Part			Gamma Part			Logit Part			Log-normal Part		
	Est.	**S.E.**	***p*-Value**	**Est.**	**S.E.**	***p*-Value**	**Est.**	**S.E.**	***p*-Value**	**Est.**	**S.E.**	***p*-Value**
Year (2008–2016)	−0.228	0.019	<0.001	−0.018	0.022	0.414	−0.228	0.019	<0.001	0.004	0.026	0.886
Organization (vs. Metro)												
KORAIL	−2.842	0.200	<0.001	0.255	0.172	0.137	−2.841	0.200	<0.001	0.131	0.301	0.664
Accident Factor (vs. Non-human-related)												
Human-related				0.485	0.122	<0.001				0.459	0.153	0.003
Train Type (vs. General)												
KTX	−0.222	0.225	0.326	0.790	0.133	<0.001	−0.218	0.225	0.333	1.593	0.150	<0.001
Urban	−1.551	0.205	<0.001	0.534	0.169	0.001	−1.549	0.205	<0.101	−0.303	0.296	0.307
Accident Type (vs. rolling-stock)												
Traffic	−0.871	0.164	<0.001				−0.871	0.164	<0.001			
Safety	−1.103	0.677	0.103				−1.098	0.676	0.104			
Miscellaneous	−0.330	0.156	0.034				−0.331	0.156	0.034			
No. of Delayed Trains	0.022	0.011	0.044	0.094	0.009	<0.001	0.022	0.011	0.043	0.053	0.006	<0.001
Interaction between No. of Trains Delayed and Train Types												
KTX	−0.003	0.031	0.918				−0.004	0.031	0.902			
Urban	0.042	0.014	0.003				0.042	0.014	0.003			
Delay Time	0.011	0.002	<0.001	0.007	0.001	<0.001	0.011	0.002	<0.001	0.001	0.001	0.145
Interaction between Delay Time and Train Types												
KTX	0.048	0.009	<0.001				0.048	0.009	<0.001			
Urban	−0.012	0.002	<0.001				−0.012	0.002	<0.001			
No. of Casualties	0.069	0.118	0.557	0.149	0.066	0.024	0.069	0.118	0.557	0.197	0.075	0.009
Log-likelihood					−4156.3						−3999.3	
AICC					8362.9						8048.9	

Est., Estimate; S.E., standard error; AICC, Akaike Information Criterion Corrected (Hurvich and Tsai [33]).

Table 6 illustrates the statistical results obtained from the TPM analysis on the amount of costs incurred due to the railway accidents during the 9 years' period of study. The severity of a railway accident is generally assessed by the amount of costs, which was our primary end-point variable. Therefore, we regarded some of the previously used dependent variables such as number of delayed trains and amount of delayed time as potential independent variables, so as to assess their effects on the amount of costs incurred during railway accidents. We also considered some interaction effects between the two factors in the model, for example, the number of trains delayed and train types and the amount of delayed time and train types. As mentioned earlier, we employed two different TPMs, the ZIG model and ZILN model to take care of the dependent variables having excessive zeros and extremely right-skewed distribution. Table 6 shows that the regression results of the two models are quite similar and the two parts of each model are statistically meaningful. It was found that with the passage of time, railway accidents incurring no costs are more likely to occur in that the corresponding estimates are mostly negative (−0.228 and −0.018 in ZIG; −0.228 in ZILN). KORAIL organization has a higher probability of causing railway accidents with no costs. Railway accidents resulting from human-related factors are more likely to increase the amount of costs rather than those due to non-human-related factors. In the case of train types, each logit part says that urban trains have a higher probability of railway accidents with no cost (−1.551 in ZIG; −1.549 in ZILN) compared with the KTX and general trains. However, railway accidents involving either the KTX (0.790) or the urban trains (0.534) result in more costs than the general trains under the ZIG model while railway accidents involving only KTX (1.593) cost more than the general trains under the ZILN model. In terms of the accident type, traffic accidents result in less costs than rolling-stock related accidents but safety-related accidents do not differ from the rolling-stock related accidents. As the number of delayed trains increases, due to railway accidents, both the probability of incurring costs (0.022 in ZIG; 0.022 in ZILN) and its amount (0.094 in ZIG; 0.053 in ZILN) increase. Further, amount of delayed time is positively related to the amount of cost as well as the probability of incurring cost. As the number of casualties increases, the amount of cost also increases. Along with the main factors explained so far, some interactions too, have statistically meaningful effects on explaining the amount of costs. For example, as more subsequent trains are delayed due to railway accidents involving urban trains, the costs are higher compared with the general trains. As the delayed time increases because of railway accidents in KTX, the probability of incurring costs rises. From the comparison of the two TPMs, the ZILN model slightly outperforms the ZIG model in that the log-likelihood value of the former is larger and its AICC is smaller even though the latter has more independent variable.

5. Conclusions

Although railroad accident evaluation is a very critical issue, not many studies related to accident evaluation have been conducted. In this study, we propose appropriate statistical models that handle the non-negative nature of accident data with respect to the damages from railway accident. As for the damage data of railroad accidents, there are many cases where the damage did not occur even though the railroad accident occurred, so a statistical approach that reflects this is necessary. To do this, we employed the two-part regression models for evaluating the damages caused by railway accidents such as the train delay time, the number of trains delayed and the costs incurred in handling accidents. For the data set, we extracted all the recorded accidents data with respect to the railway types (urban, general and high-speed railway), organization types (metro and KORAIL) and accident factors (human-related and non-human related). Further, we analyzed the statistical results to identify the variables that are highly correlated to the accident types and the magnitude of accidents reflecting the train delay and the costs.

Overall, we found that the railway accidents in South Korea continued to decrease with the passage of time in terms of the number of accidents occurred and the amount of costs incurred during the period 2008 to 2016. During this analysis period, new rolling stocks replaced the old ones at the

general railway system and PSDs were installed at the urban railway system in order to reduce suicide attempts. It seems to have played a positive role in reducing the railway accidents.

From the statistical analyses, we found that the number of trains delayed tends to increase during the period of our interest to the railway accidents but KORAIL had smaller number of trains delayed than the other organizations. The KTX caused the least delay in trains, followed by the general trains and urban railway resulted in the most delay in trains. Human-related accidents resulted in more delays in trains compared to the non-human related accidents. It was also found that with increase in derailed trains more trains got delayed. Considering the amount of delay time due to railway accidents, KORAIL results in less delay time than the other organizations. Both the KTX and urban trains tend to decrease the amount of delay time compared to that of the general trains. Human-related accidents reduced the probability of having delay time rather than non-human related accidents. What is also important here is that, as the number of casualties increased, it took longer time for train operations to normalize.

We included the number of delayed trains as the dependent variable and the amount of time delayed as potential independent variable in the TPMs, to comprehensively evaluate the amount of costs from railway accidents. Some potential interactions between independent variables were also included in the TPMS. With the passage of time, the probability of having railway accidents with no costs generally increased. Especially, KORAIL tends to reduce the probability of causing railway accidents with costs compared with metro. Railway accidents caused by human-related factors are more likely to increase the amount of costs. Urban railways have the least chance of getting involved in railway accidents with costs and KTX results in more costs than the general railways. It is interesting that, as the number of delayed trains increases, both the probability of incurring costs and its amount of costs increase. This is also the case for amount of delay time. The number of casualties has a positive impact on the amount of costs. When an accident was involved in urban railway and more subsequent trains are delayed, the costs are likely to increase compared to an accident involving general railway. If an accident is human-related and has occurred in KTX, then the costs also escalate.

KORAIL operates both the KTX and the general railroads to provide regional transport services between those cities that are quite far ranging from 50 km to 450 km. KTX operates on exclusive routes, so the impact of accidents caused by external factors is very low. On the other hand, in the general railway system, there are accidents due to collisions with vehicles and people, at the railroad-crossings. This implies that there exist many factors in the external environment such as automobiles, pedestrians, animals, weather and so forth, which cause railway accidents. Regarding the delay and the costs associated with it due to accidents, general railways have a low frequency of operation, so the headways are relatively long compared to the KTX and the urban railways. Hence the costs, due to accidents, on the general railway system, are relatively low and the delay time of subsequent trains is also short. In the case of urban railways, the interval between the trains is very short, so the number of delayed trains and the delay time of the subsequent trains that follow in the event of an accident can be quite large. Unlike the general and the urban railway system, KTX has a large cost to recover, whenever there is a malfunctioning of the trains or the railroad systems due to the accident. Moreover, the compensation costs for the passengers due to the delayed time, caused from accidents, are very high due to the relatively higher fare.

Although the railway accident data provide detailed information regarding the number of trains delayed and the costs incurred for accident recovery based on the railway type, operator and the accident type, we cannot specify the full impact of the accidents and evaluate them. To enable more detailed analysis in future, the railway accident data should contain more accurate information, including the load of railway service by line, the location of the accident, causes, the accident recovery costs and the time of accident and so forth. For example, the information about the load factors of railway lines such as, the number of train operations and volume of passengers should be provided so that relative comparisons of the magnitude of damages for each railway route can be evaluated.

Even though the data are limited for the public, this study is informative for transit agencies to evaluate the impact of railway accidents and to create better railway services for the public.

Author Contributions: Conceptualization, M.S.P., J.K.E., J.C., and T.-Y.H.; methodology, M.S.P., J.C., and T.-Y.H.; data analysis, M.S.P., J.C., and T.-Y.H.; investigation, M.S.P., J.K.E., J.C., and T.-Y.H.; data curation, J.K.E., M.S.P., J.C.; writing—original draft preparation, M.S.P., J.K.E., J.C., and T.-Y.H.; writing—review and editing, M.S.P., J.K.E., J.C., and T.-Y.H.; supervision, J.K.E. All authors have read and agreed to the published version of the manuscript.

Funding: This research was supported by a grant from R&D Program of the Korea Railroad Research Institute (PK2001C1). This work was partially supported by the research fund of the Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education (2018R1D1A1B07047712, 2019R1|1A3A01057696).

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Haleem, K. Investigating risk factors of traffic casualties at private highway-railroad grade crossings in the United States. *Accid. Anal. Prev.* **2016**, *95*, 274–283. [CrossRef] [PubMed]
2. Marković, N.; Milinković, S.; Tikhonov, K.S.; Schonfeld, P. Analyzing passenger train arrival delays with support vector regression. *Transp. Res. Part C Emerg. Technol.* **2015**, *56*, 251–262. [CrossRef]
3. European Railway Agency. *Railway Safety Performance in the European Union 2016.* European Railway Agency: Valenciennes, France. Available online: https://erail.era.europa.eu/documents/SPR.pdf (accessed on 27 July 2020).
4. Austin, R.D.; Carson, J.L. An alternative accident prediction model for highway-rail interfaces. *Accid. Anal. Prev.* **2002**, *34*, 31–42. [CrossRef]
5. Hu, S.-R.; Li, C.-S.; Lee, C.-K. Investigation of key factors for accident severity at railroad grade crossings by using a logit model. *Saf. Sci.* **2010**, *48*, 186–194. [CrossRef] [PubMed]
6. Djordjević, B.; Krmac, E.; Mlinarić, T.J. Non-radial DEA model: A new approach to evaluation of safety at railway level crossings. *Saf. Sci.* **2018**, *103*, 234–246. [CrossRef]
7. Iranitalab, A.; Khattak, A.J. Probabilistic classification of hazardous materials release events in train incidents and cargo tank truck crashes. *Reliab. Eng. Syst. Saf.* **2020**, *199*, 106914. [CrossRef]
8. Pasha, J.; Dulebenets, M.A.; Abioye, O.F.; Kavoosi, M.; Moses, R.; Sobanjo, J.; Ozguven, E.E. A comprehensive assessment of the existing accident and hazard prediction models for the highway-rail grade crossings in the state of Florida. *Sustainability* **2020**, *12*, 4291. [CrossRef]
9. Davey, J.; Wallace, A.; Stenson, N.; Freeman, J. The experiences and perceptions of heavy vehicle drivers and train drivers of dangers at railway level crossings. *Accid. Anal. Prev.* **2008**, *40*, 1217–1222. [CrossRef]
10. Federal Railroad Administration. *Highway-Rail Grade Crossing Safety Research*; Office of Research and Development: Washington, DC, USA, 1996.
11. Eluru, N.; Bagheri, M.; Miranda-Moreno, L.F.; Fu, L. A latent class modeling approach for identifying vehicle driver injury severity factors at highway-railway crossings. *Accid. Anal. Prev.* **2012**, *47*, 119–127. [CrossRef]
12. Hao, W.; Daniel, J. Motor vehicle driver injury severity study under various traffic control at highway-rail grade crossings in the United States. *J. Saf. Res.* **2014**, *51*, 41–48. [CrossRef]
13. Hao, W.; Kamga, C.; Daniel, J. The effect of age and gender on motor vehicle driver injury severity at highway-rail grade crossings in the United States. *J. Saf. Res.* **2015**, *55*, 105–113. [CrossRef]
14. Zhao, S.; Khattak, A. Motor vehicle drivers' injuries in train-motor vehicle crashes. *Accid. Anal. Prev.* **2015**, *74*, 162–168. [CrossRef]
15. Yan, X.; Han, L.D.; Richards, S.; Millegan, H. Train-vehicle crash risk comparison between before and after stop signs installed at highway-rail grade crossings. *Traffic Inj. Prev.* **2010**, *11*, 535–542. [CrossRef]
16. Oh, J.; Washington, S.P.; Nam, D. Accident prediction model for railway-highway interfaces. *Accid. Anal. Prev.* **2006**, *38*, 346–356. [CrossRef]
17. Miranda-Moreno, L.F.; Fu, L. A comparative study of alternative model structures and criteria for ranking locations for safety improvements. *Netw. Spat. Econ.* **2006**, *6*, 97–110. [CrossRef]
18. Lambert, D. Zero-inflated Poisson regression, with an application to defects in manufacturing. *Technometrics* **1992**, *34*, 1–14. [CrossRef]

19. Ridout, M.; Demétrio, C.G.B.; Hinde, J. Models for count data with many zeros. In Proceedings of the 19th International Biometric Conference, Cape Town, South Africa, 14–18 December 1998.
20. Joe, H.; Zhu, R. Generalized Poisson distribution: The property of mixture of Poisson and comparison with negative binomial distribution. *Biom. J.* **2005**, *47*, 219–229. [CrossRef]
21. Mwalili, S.M.; Lesaffre, E.; Declerck, D. The zero-inflated negative binomial regression model with correction for misclassification: An example in caries research. *Stat. Methods Med. Res.* **2007**, *17*, 123–139. [CrossRef]
22. Neelon, B.H.; O'Malley, A.J.; Normand, S.-L.T. A Bayesian model for repeated measures zero-inflated count data with application to outpatient psychiatric service use. *Stat. Model. Int. J.* **2010**, *10*, 421–439. [CrossRef]
23. Neelon, B.; O'Malley, A.J.; Smith, V.A. Modeling zero-modified count and semi-continuous data in health services research. Part1: Background and overview. *Stat. Med.* **2016**, *35*, 5070–5093. [CrossRef]
24. Kern, D.; Wasser, T. Analysis of health care costs containing a large proportion of $0 data using traditional and zero-inflated gamma regression models. *Value Health* **2013**, *16*, A21. [CrossRef]
25. Nobre, A.A.; Carvalho, M.S.; Griep, R.H.; Fonseca, M.D.J.M.D.; Melo, E.C.P.; Santos, I.D.S.; Chór, D. Multinomial model and zero-inflated gamma model to study time spent on leisure time physical activity: An example of ELSA-Brasil. *Rev. Saúde Públ.* **2017**, *51*, 76. [CrossRef]
26. Tong, E.N.; Mues, C.; Thomas, L.C. A zero-adjusted gamma model for mortgage loan loss given default. *Int. J. Forecast.* **2013**, *29*, 548–562. [CrossRef]
27. Risio, L.; Calama, R.; Bogino, S.M.; Bravo, F. Inter-annual variability in Prosopis caldenia pod production in the Argentinean semiarid pampas: A modelling approach. *J. Arid. Environ.* **2016**, *131*, 59–66. [CrossRef]
28. Neelon, B.; Ghosh, P.; Loebs, P.F. A spatial Poisson hurdle model for exploring geographic variation in emergency department visits. *J. R. Stat. Soc. Ser. A* **2013**, *176*, 389–413. [CrossRef]
29. Ghosh, S.K.; Mukhopadhyay, P.; Lu, J.-C. Bayesian analysis of zero-inflated regression models. *J. Stat. Plan. Inference* **2006**, *136*, 1360–1375. [CrossRef]
30. Neelon, B.; O'Malley, A.J.; Normand, S.-L.T. A bayesian two-part latent class model for longitudinal medical expenditure data: Assessing the impact of mental health and substance abuse parity. *Biometrics* **2011**, *67*, 280–289. [CrossRef]
31. Cooper, N.J.; Sutton, A.J.; Mugford, M.; Abrams, K.R. Use of Bayesian Markov chain Monte Carlo methods to model cost-of-illness data. *Med. Decis. Mak.* **2003**, *23*, 38–53. [CrossRef]
32. Cooper, N.J.; Lambert, P.C.; Abrams, K.R.; Sutton, A.J. Predicting costs over time using Bayesian Markov chain Monte Carlo methods: An application to early inflammatory polyarthritis. *Health Econ.* **2006**, *16*, 37–56. [CrossRef]
33. Hurvich, C.; Tsai, C.-L. Regression and time series model selection in small samples. *Biometrika* **1989**, *76*, 297–307. [CrossRef]