THE PROBABILISTIC MODEL DEVELOPMENT TO PREDICT TRAFFIC ACCIDENT TENDENCY UNDER SNOWY CONDITIONS FOR EXPRESSWAYS

Bui Tien MANH [1], KPD Frank PERERA[2], Kazushi SANO[3], Takumi TAKAKURA[4], Teppei KATO[5]

[1] Graduate student, Dept. of Civil Eng., Nagaoka University of Technology
(Kamitomioka 1603-1, Nagaoka, Niigata, 940-2188, Japan)
E-mail: buitienmanh91@gmail.com

[2] Graduate student, Dept. of Science of Technology Innovation, Nagaoka University of Technology
(Kamitomioka 1603-1, Nagaoka, Niigata, 940-2188, Japan)
E-mail: s205001@stn.nagaokaut.ac.jp

[3]Professor, Dept. of Civil Eng., Nagaoka University of Technology
(Kamitomioka 1603-1, Nagaoka, Niigata, 940-2188, Japan)
E-mail: sano@nagaokaut.ac.jp

[4] Masters, New Civil Engineering (NCE) Company
(Misakichyo 1-7-25, Chuo-ku, Niigata-shi, Niigata, 950-0954, Japan)
E-mail: T-Takakura@nceinc.co.jp

[5] Associate Professor, Dept. of Civil Eng., Nagaoka University of Technology
(Kamitomioka 1603-1, Nagaoka, Niigata, 940-2188, Japan)
E-mail: tkato@vos.nagaokaut.ac.jp

Traffic accidents are a global challenge that countries in the world have been facing. According to statistics from the US, Canada, and Japan and recent studies have shown that the frequency of traffic accidents tends to increase in snowy conditions. The study aims to develop a probabilistic model to predict hourly traffic accident tendency under snowy conditions on expressways. The study conducts a correlation analysis between the factors to detect multicollinearity occurring in the regression models. Then, the study utilized a negative binomial regression model to predict the effects of factors on the tendency of hourly traffic accidents and evaluated this model based on residual analysis, the goodness of fit, and predictive performance evaluation. The results of the model have shown that hourly traffic volume and average snowfall have a positive effect on the likelihood of hourly traffic accidents. Meanwhile, factors including the percentage of trucks, average flow speed, temperature, and vertical gradients have a negative effect on the likelihood of hourly traffic accidents. In addition, the hourly accident frequency on the divided segments tends to have a higher accident frequency than on the un-divided segments on expressways under snowy conditions. Finally, the results of the model evaluation also show that this model has good predictive performance.

Key Words: *Traffic accidents, Stacks, Negative binomial regression Model, Snowy conditions, Expressways in Japan*

1. INTRODUCTION

Road traffic accidents are one of the global challenges that countries in the world have been facing. According to a global status report on road safety, 2018, the number of road traffic accident deaths reach a high of 1.35 million in 2016; and road traffic injury is the 8th leading cause of death for all age groups, up from the 9th leading cause of death.

Road traffic accidents also occur with the change between seasons. The frequency of traffic accidents in the winter season is three times higher than in summer because drivers are affected by adverse factors such as heavy snowfall, blowing snow, high winds, and fog, which lead to snow pavement, poor visibility, and rust is increasing *(Saha et al. 2015, Gaweesh et al. 2019)*.

In snowy conditions, studies show that traffic accidents have an increased tendency. Practically, in the USA, 10 – the year running percentage of crashes involving snow/sleet snow/sleet, icy, or slushy pavement was around 11% of all reported crashes (*Wong et al. 2021*). In Canada, the number of collisions showed a seasonality whereby the number of

property damage only collisions (no injuries or fatalities) were higher over the winter months (*Pennelly et al. 2018*). In Japan, the most traffic accidents in the winter season have a dramatic increase in November and December corresponding to the first snowfall, and the number of fatalities for accidents in the winter season has been largely stable (*Asano 2003*).

According to statistics about accidents on expressways in Japan, over a 10-year period (2010-2020), the number of accidents per 100,000 persons on expressways decreased by 57% but the number of accidents remained at a high level (309,178 in 2020). The study utilized an accident dataset on expressways in Niigata prefecture in Japan. The Niigata prefecture has five main corridors in the expressway network that covers the prefecture and experience heavy snowfall in the winter season. According to the analysis of the accident dataset on five expressways in Niigata prefecture, this study has shown that accidents occurring on expressways during the winter season (from December to March) are higher compared to other seasons, in which accidents are concentrated in January and tend to decrease gradually until the end of March. Figure 1 shows the number of accidents and the percentage of accidents occurring on the expressways in Niigata prefecture from 2012 to 2020.

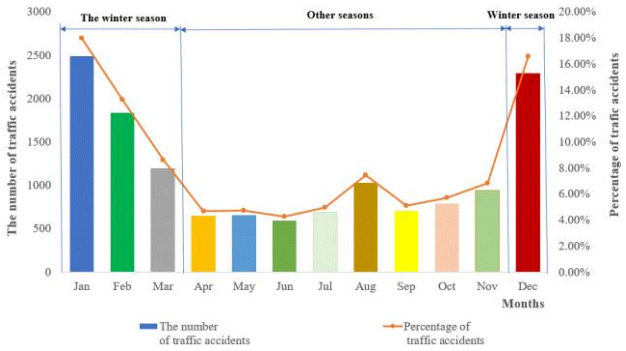

Fig.1. Number and percentage of accidents occurring on expressways between months in Niigata prefecture, Japan.

In order to reduce accidents in the winter season under snow conditions, studies have developed different methods and models to analyze the influencing factors and predict accidents related to accident severity *(Sano et al. 2010, Heqimi et al. 2018, Hyodo et al. 2021)*, the accident frequency within snowy conditions *(Usman et al. 2010, Usman et al. 2011, Gaweesh et al. 2019)*. This helps to provide policies and solutions for the management and maintenance of state road agencies.

With the studies about accidents in snow conditions for expressways, the studies focused on analyzing the factors affecting the accident severity and accidents risk *(Sano et al. 2010, Kim et al. 2021)*, and estimated rates of winter collisions in snow road conditions based on weather warning type and other meteorological factors (*Wong et al. 2021*). In addition, some studies have developed models to analyze factors and predict the accident frequency occurring on expressways, but did not consider factors related to snow conditions (*Ma et al. 2017*). Therefore, at present, there have not been many studies on developing models to analyze influencing factors and predict traffic accident trends for expressways that consider snowy conditions in the winter season.

Besides, studies about accidents in snowy conditions focused on factors such as traffic volume, snowfall, road surface condition, visibility, precipitation, rainfall, the gap between two snow events, the number of rainy days, etc *(Usman et al. 2010, Usman et al. 2011, Seeherman et al. 2015, Gaweesh et al. 2019)*. However, these studies have not considered factors such as average flow speed, and segment type, and identified the factors affecting accidents in snowy condistions. Besides, the influencing factors are aggregated over a long period of time (years or days), which leads to decrease accuracy in forecasting and analyzing the impact of factors on accidents under winter conditions.

Therefore, the objective of this study selects the factors based on correlation analysis and develops a probabilistic model to predict the trend of accidents occurring on the expressway under snowy conditions in the winter season. In addition, this study also considers accidents and factors affecting accidents in a short time (hours and minutes) to ensure the accuracy of the probabilistic model.

The paper is structured as follows, section 2 reviews previous studies related to the factors that affect accidents in snow conditions and the statistical models used for accident prediction. Section 3 describes the used methodologies in this study including correlation analysis, negative binomial regression model, and model evaluation. Next, the study will describe the study site, data processing, and data structure for developing the probabilistic model. The results of the study are presented and discussed in section 5. Finally, Section 6 provides the conclusion and limitations of this study.

2. LITERATURE REVIEW

Traffic accidents are the global challenges as mentioned in the previous section. Currently, there is a number of studies around the world that have studied this research area *(Abdel-Aty et al. 2000, Chengye et al. 2013)*. This aids state management organizations in developing policies that effectively lower accident frequency and severity. Studies often focus on analyzing factors that affect traffic acci-

dents and developing models to predict traffic accidents using traditional statistical models and machine learning algorithms. For the studies of traffic accidents in snowy conditions, there are a number of studies that analyze factors affecting the accident severity *(Sano et al. 2010, Heqimi et al. 2018, Hyodo et al. 2021)* and predict the risk of accidents on expressways *(Kim et al. 2021)* and the frequency of traffic accidents *(Usman et al. 2010, Usman et al. 2011)*.

(1) Factors affecting the likelihood of accidents in snowy conditions

The studies analyzed factors related to accident frequency and severity in snowy conditions. Studies have shown that factors such as annual average daily traffic (AADT), segment lengths, percentage of heavy vehicles, absence of a variable speed limit, snowfall, the longitudinal gradient, and road surface condition index have positive effects on the likelihood of accidents *(Qiu et al. 2008, Sano et al. 2010, Usman et al. 2010, Usman et al. 2011, Seeherman et al. 2015, Heqimi et al. 2018, Hyodo et al. 2021)*. In addition, factors such as temperature, and visibility have negative effects on the likelihood of accidents *(Usman et al. 2010, Usman et al. 2011)*.

However, in accident prediction models, these factors are considered over a long period of time such as annual average daily traffic (AADT), average speed and percentage of trucks considered by day, snowfall, and average temperature by month. This will not accurately reflect the impact of factors affecting accidents in snowy conditions. In addition, the above studies have not identified the effect between factors in the prediction model. This study will detail these factors in hours for traffic volume and percentage of trucks and minutes for other factors.

(2) The statistical models to predict traffic accidents

Currently, the studies use different statistical models to predict the frequency of traffic accidents such as the Poisson regression models *(Jovanis et al. 1986, Joshua et al. 1990)*, Negative binomial regression model *(Abdel-Aty et al. 2000, Chengye et al. 2013, Ma et al. 2017)*, Poisson-lognormal regression model *(Miaou et al. 2005*), Gama regression model *(Oh et al. 2006, Daniels et al. 2010)*, Random- effects model *(Johansson 1996, Shankar et al. 2003)*. However, each of these statistical models has different advantages and disadvantages such as over-dispersion or under-overdispersion, a large number of zero-crash observations, and temporal correlation (*Lord et al. 2010*).

The negative binomial regression model is applied widely for modeling traffic accident prediction models compared to other statistical models *(Abdel-Aty et al. 2000, Usman et al. 2010, Chengye et al. 2013, Seeherman et al. 2015, Ma et al. 2017, Heqimi et al. 2018)*. The advantages of this regression model are easy to estimate and can handle overdispersion because its estimate variance separates parameter from mean and rate. However, the disadvantage of this model can't handle under dispersion and can be detrimentally influenced by the low-sample mean and small sample size bias (*Lord et al. 2010*).

In these models, the dependent variables often consider the frequency of accidents on segments by the long yearly interval *(Abdel-Aty et al. 2000, Chengye et al. 2013, Ma et al. 2017, Gaweesh et al. 2019)* without considering the short hourly interval. This leads to a loss of information and models of distorted risk factors and effect size. Also, the effects of these insignificant variables could be distributed to the significant variables, distorting their parameter estimates *(Usman et al. 2011)*. Therefore, the model develops in this study predicts the frequency of accidents in the segments by the hour to reflect the impact of factorss on traffic accidents accurately.

Besides, the spatial is considered in these models. Studies often divided roadways into different segments based on homogenous characteristics or fixed length. Each study showed different results regarding the performance of these models related to the spatial. Specifically, *Gaweesh et al. 2019* showed that subdividing roadways into segments with homogenous characteristics could be more suitable for prediction in these models. Meanwhile, *Ma et al. 2017* showed that the model performance with the fixed-length segment method is superior to that with the homogeneous characteristics. In this study, we will use segments that are divided into homogenous characteristics by the data accuracy of the longitudinal gradients of each segment.

Finally, the development of these models in studies usually focuses on highways, roadways in urban areas, motorways, and roadways in mountainous areas (*Usman et al. 2010, Usman et al. 2011, Chengye et al. 2013, and Seeherman et al. 2015)*. A few studies have developed predictive models about the accident frequency for expressways but have not taken into snowy conditions during the winter season (*Ma et al. 2017*). Besides, a few researchers have studied the influencing factors of head-to-head collisions on undivided segments on the expressway, but a prediction model for accident frequency for expressways has not been developed *(Sano et al. 2010)*. Literature shows that there have not been many studies on developing probabilistic models to

predict the accident tendency under snow conditions for expressways. This study will focus on solving this problem.

3. METHODOLOGY

Firstly, the study used correlation analysis to determine the correlation between factors used in the model. That helps to detect the multicollinearity in the regression model. Then, the study develops a probabilistic model using the negative binomial regression model to predict accident trends on expressways under snowy conditions. Finally, the study evaluates the model's fit based on residual analysis, the goodness of fit test, and the evaluation of the predictive performance. The details of these methodologies are as follows:

(1) Correlation analysis

This study uses correlation analysis to determine the correlations between the variables that used the probabilistic model to avoid multicollinearity. The correlation coefficient is a coefficient to measure the intensity of the linear association between variables. Besides this coefficient is also possible to have nonlinear associations. Suppose X and Y are two random variables with variance V(X) > 0, and V(Y) > 0. The correlation coefficient of two random variables X and Y, denoted ρ (X, Y), is calculated as follows:

$$\rho(X,Y)=\frac{Cov(X,Y)}{\sqrt{V[X]}\sqrt{V[Y]}}=\frac{E[XY]-E[X]E[Y]}{\sqrt{V[X]}\sqrt{V[Y]}} \quad (1a)$$

However, if the distribution of (X, Y) is not known, it is very difficult to calculate the theoretical correlation coefficient ρ. Therefore, we need to find estimates of the correlation coefficient ρ through a random sample. Suppose we conduct n independent observations for a pair of random variables (X, Y), we have a random sample of size n: (X_1, Y_1), (X_2, Y_2), (X_n, Y_n). The sample correlation coefficient of X and Y, denoted r, is calculated by the following formula:

$$r=\frac{\sum_{i=1}^{n}\left(X_i-\overline{X}\right)\left(Y_i-\overline{Y}\right)}{\sqrt{\sum_{i=1}^{n}\left(X_i-\overline{X}\right)^2}\sqrt{\sum_{i=1}^{n}\left(Y_i-\overline{Y}\right)^2}}=\frac{\overline{XY}-\overline{X}\,\overline{Y}}{S_X S_Y} \quad (1b)$$

Correlation coefficient values range -1 to +1. The closer to 1 the correlation coefficient gets the "stronger" the correlation. If $r \geq 0$, variable X and variable Y are positive correlations, and if $r < 0$, then variable X and variable Y are negative correlations. If $0.7 \leq |r| \leq 1$, then variable X and variable Y are strongly correlated, and if $0.3 \leq |r| \leq 0.7$, variable X and variable Y are weakly correlated.

The correlation is higher with the pairwise variables, it indicates the possibility of collinearity. If the absolute value of the correlation coefficient is close to 0.8, collinearity is likely to exit *(Shrestha 2020)*.

(2) Model development

The probabilistic model used in this study is the negative binomial regression model to predict the trend of accidents for expressways in snowy conditions. The negative binomial regression model is one of the generalized linear models *(Zwilling 2013)*. The dependent variable, Y, is assumed to follow a negative binomial distribution with the mean μ, which is dependent on explanatory variables. The negative binomial regression model with probability density function is as follows:

$$\Pr(Y=y_i|\mu_i,\alpha)=\frac{\Gamma\left(y_i+\alpha^{-1}\right)}{\Gamma\left(\alpha^{-1}\right)\Gamma\left(y_i+1\right)}\left(\frac{1}{1+\alpha\mu_i}\right)^{\alpha^{-1}}\left(\frac{\alpha\mu_i}{1+\alpha\mu_i}\right)^{y_i} \quad (2a)$$

Where:

- $\Pr\left(Y=y_i|\mu_i,\alpha\right)$: is the probability of accidents y_i occurring segment j^{th} over the hour i (i = 1, …., n).
- y_i: is the number of accidents occurring to a given segment j^{th} during the hour i, y_i = 0,1, 2.
- α: is the dispersion parameter.
- μ_i is the expected number of accidents of segment j^{th} during the hour i. In this study, μ_i follows the below function:

$$\mu_i=e^{\beta_0}L_i\times TV_i^{\beta_1}\times e^{(\beta_2 X_2+.....+\beta_n X_n+\varepsilon_i)} \quad (2b)$$

- L_i: is the length of segment j^{th}.
- TV_i: is the hourly traffic volume of segment j^{th}.
- X_i: is the explanatory variables at given segment j^{th} during the hour i, such as snowfall, temperature, vertical gradients, etc.
- ε_i: is the gamma distributed error term with mean 1 and variance α^2.
- β_i: is the regression coefficients.

The coefficients β_i and the overdispersion parameter α are estimated by maximizing the log-likelihood function. This function was shown as follows:

$$L(\alpha,\beta)=\prod_{i=1}^{n}\Pr\left(Y=y_i|\mu_i,\alpha\right)=\prod_{i=1}^{n}\left(\frac{\Gamma\left(y_i+\alpha^{-1}\right)}{\Gamma\left(\alpha^{-1}\right)\Gamma\left(y_i+1\right)}\left(\frac{1}{1+\alpha\mu_i}\right)^{\alpha^{-1}}\left(\frac{\alpha\mu_i}{1+\alpha\mu_i}\right)^{y_i}\right) \quad (2c)$$

Besides, in the negative binomial regression model, the mean and variance of dependent variables are:

$$E\left(y_i|\mathrm{X}_i\right)=\mu_i \quad (2d)$$

$$Var\left(y_i|\mathrm{X}_i\right)=\mu_i+\alpha\mu_i^2 \quad (2e)$$

In the negative binomial regression model, the variance of dependent variables is always greater than the mean. If α is zero, the negative binomial regression model is essentially a Poisson regression model (*Lord and Mannering 2010*).

(3) Model evaluation

To evaluate the performance of the model in fitting individual observation and the overall performance of the model, this study will conduct residual analysis, the goodness of fit of the model, and the evaluation of predictive performances as follows:

(3.1) Residual analysis

The residual is used to measure the departure of the fitted value from the actual values of the dependent variable. They are used to detect model misspecification; outliers, observation with the poor fit; and influential observation. The study will use three residuals to evaluate the model which are raw residual, Pearson residual, and deviance residuals. The formulas for these residues are expressed as follows:

Raw residual

$$r_i = \left(y_i - \mu_i\right) \quad (3a)$$

Where y_i is actual values, and the fitted mean μ_i is the conditional mean.

Pearson residual

$$p_i = \frac{\left(y_i - \mu_i\right)}{\sqrt{\omega_i}} \quad (3b)$$

Where: ω_i is an estimate of variance of ω_i and y_i. For negative binomial regression models, one use:

$$\omega = \mu + \alpha\mu^2 \quad (3c)$$

Deviance residual

$$d_i = sign\left(y_i - \mu_i\right)\sqrt{2\left\{y_i \ln\left(\frac{y_i}{\mu_i}\right) - \left(y_i + \alpha^{-1}\right)\ln\left(\frac{y_i + \alpha^{-1}}{\mu_i + \alpha^{-1}}\right)\right\}} \quad (3d)$$

(3.2) Goodness of fit

It is an obligatory requirement to check the overall fit and quality of the fit between the observed value and the predicted value. This study uses some of the following measures of goodness of fit:

Deviance statistic

One of the important statistics in measuring the model fit is the deviance which is twice the difference between the maximum log-likelihood achievable and the log-likelihood of the fitted model. In the negative binomial regression model, the deviance is calculated as follows:

$$D_{NB2} = \sum_{i=1}^{n}\left\{-y_i \ln\left(\frac{y_i}{\mu_i}\right) - \left(y_i + \alpha^{-1}\right)\ln\left(\frac{y_i + \alpha^{-1}}{\mu_i + \alpha^{-1}}\right)\right\} \quad (3e)$$

For the best model, one must expect a smaller value of the deviance. Therefore, the smaller the value of the deviance in a specific model, better the model or more statistically significant the model becomes *(Omari-Sasu et al. 2016)*.

Akaike's information criterion (AIC)

Akaike's information criterion (AIC) is proposed as a model selection index based on the fitted log-likelihood function. For any statistical model, the AIC index is calculated as follows:

$$AIC = -2\ln L + 2k \quad (3f)$$

Where L is the maximized value of the likelihood function and k is the number of parameters in the model. Based on the AIC index, a model is considered good, when the AIC index is the smallest.

Starting with a full set of explanatory variables, other link functions, and interactions, a stepwise procedure has been used to select the model based on minimizing the AIC.

Ratio of log-likelihood index (ρ2)

The log-likelihood ratio value of the model (ρ^2) which is an indication of the additional variation in accident frequency explained by the obtained model to the constant term *(Abdel-Aty and Radwan 2000)* was also used in this study. This index is defined according to the following formula:

$$\rho^2 = 1 - \frac{L(\beta)}{L(0)} \quad (3g)$$

Where L(β) is the log-likelihood value of the fitted model, and L (0) is log-likelihood value of the zero model.

(3.3) The evaluation of predictive performances

In order to evaluate the predictive ability of the model, the previous studies (*Li et al. 2008*, *Chengye et al. 2013)* used main metrics which are the mean absolute deviation (MDA), the mean squared prediction error (MSPE), and the root mean square error (RMSE). The MDA evaluates the prediction deviation after calculating the mean absolute error but does not consider the error direction. The MSPE is the same as the MDA, but it is used to determine the variance of the difference between predicted and observed results. Meanwhile, the RSME shows how dispersed the data is compared to the model, or how close the data points are to a fitted line (*Silva et al. 2020*). Their formula is shown as follows:

$$MAD = \frac{\sum_{i=1}^{n} |y_i - y_i|}{n} \quad (3h)$$

$$MSPE = \frac{\sum_{i=1}^{n} \left(y_i - y_i\right)^2}{n} \quad (3k)$$

$$RMSE = \sqrt{MSPE} \quad (3l)$$

Where: y_i and y_i is the observed accident frequency on segment j^{th} over the hour i, and the predicted accident frequency on segment j^{th} over the hour i, respectively, and n the size of the set training or test. The desirable model should fit the data as closely as possible, which is presented by the smaller value of MAD, MSPE, and RMSE (*Chengye et al. 2013)*.

4. DATA DESCRIPTION

(1) Study scope and data

The study scope is Niigata Prefecture of Japan which has a population of 2,227,496 (1 July 2019) and by geographic area of 12,584.18 km^2. Besides, Niigata is one of the prefectures with heavy snowfall in the winter season from late November to the end of March every year in Japan.

The scope of this study focuses on traffic accidents and stacks occurring on five major expressways that are Hokuriku Expressway, Kan-etsu Expressway, Ban-etsu Expressway, Joushin-etsu Expressway, and Nihonkai-Tohoku Expressway. The total length of the five expressways is 404.8 kilometers and covers the entire Niigata Prefecture. Fig.2 below shows the scope of this study.

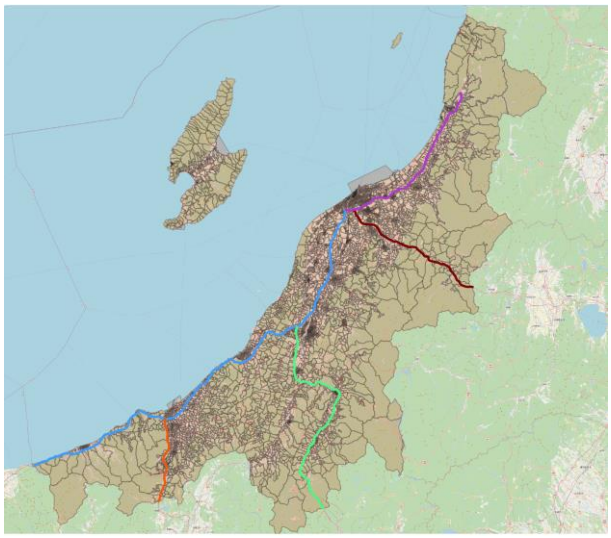

Fig.2. Map about Niigita Prefecture and expressway network

The data source in this study was collected from the Nippon Expressway Company Limited (NEXCO) at five expressways in Japan's Niigata Prefecture. The dataset was collected over 8 winter seasons (from December 2013 to March 2021), and over 4 months for each winter season (from December to March). The dataset used in this study includes subdivided datasets which are an accident and stack dataset, a traffic condition dataset, a roadway dataset, and a weather condition dataset.

The information in the dataset regarding accidents and stacks that occurred on expressways includes locations, times, severity (fatal, injury, and minor injury), road surface conditions, vehicle types, and pavement types. The total number of accidents and stacks occurring on expressways in the winter season is 5306 causes for the development of the probabilistic model to predict the accident tendency for expressways in this study.

In addition, the traffic condition dataset includes information related to the collected time, segments, directions, the number of cars, the number of trucks, and the average follow speed. The roadway dataset represents information about the geometric design features of five expressways that include the length of the slope, the longitudinal gradient, the longitudinal types, divided segments, and un-divided segments. Finally, the weather condition dataset includes stations, the collected time, the amount of snowfall, and temperature.

(2) Data processing

After collecting the datasets presented in the previous section. This study synthesized and matched the datasets into a common dataset based on spatial (segments) and temporal (hour) to select factors and develop the probabilistic model in snow conditions. For the spatial problem, this study has divided each expressway into different segments based on the homogenous characteristic of slopes along each direction. After dividing the expressway into different segments, the total number of sections is 871 segments for both directions on five expressways in this study. Meanwhile, the temporal issue has been divided by hour to reflect the accurate impact of factors on accidents under snowy conditions.

The data processing is performed by R software, version 4.2.1. R software is a language used for statistical computing and graphics. The study used some main packages that include the MASS package, ggplot 2 package, tidyverse package, dplyr package to process the data and develop the probabilistic model.

Table 1 below shows descriptive statistics of data related to accident/stack (dependent variable) and factors (explanatory variables) used for developing the probabilistic model in this study.

Table 1 Descriptive statistics about data in the model

Variable name	Min	Max	Mean	SD
Accident frequency (per hour)	0.00	2.00	0.0003	0.019
Segment Length (Km)	0.21	3.91	0.95	0.56
Hourly traffic volume	1	4,435	426	369.48
Percentage of trucks (%)	0.00	0.50	0.25	0.13
Average flow speed (km/h)	0.17	150.42	84.54	11.18
Average Snowfall (cm/10minutes)	0.00	4.93	0.03	0.10
Temperature (°C)	-11.42	31.28	4.53	4.36
Vertical gradient (%)	-4.5	4.50	-0.02	1.50
Variable name	Category		Number of Segments	
Segment type	Un-divided		174 (19.98%)	
	Divided		697 (80.02%)	

5. RESULTS AND DISCUSSIONS

(1) Correlation analysis between variables

Before developing the model, the study performed a correlation analysis between the explanatory variables (factors) in the model. The correlation coefficient and the level of correlation between the explanatory variables in the model are shown in the heat map in figure 3 below. The results of the correlation analysis show that explanatory variables in the model have a weak correlation. This shows that the explanatory variables in the model will not cause multicollinearity for the model.

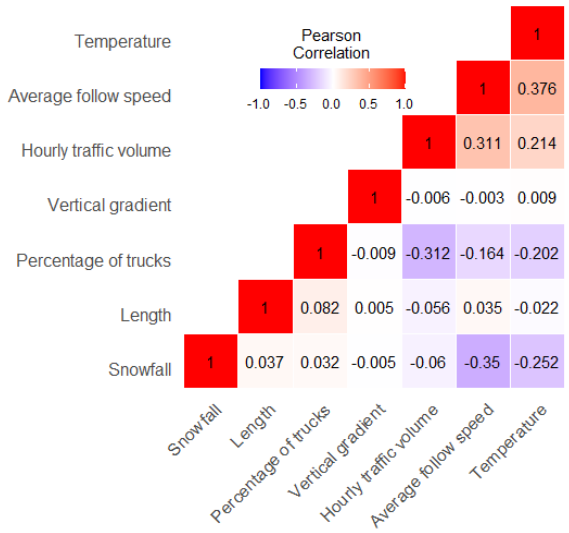

Fig.3 Heat map about correlation between explanatory variables

(2) Estimated results of model

After analyzing the correlation between the variables in the model, this study used the R software to develop the negative binomial regression model. The result of this model for the hourly accident frequency under snow conditions for the expressway is presented in Table 2. This table contains explanatory variables, parameter estimates, z values, and the respective p values of variables. Besides, this table also shows that all variables have the expected sign (with positive signs indicating an increase in the hourly accident frequency, and negative signs indicating a decrease). In addition, the log-likelihood, and standard error are also shown in this table.

Firstly, the results of this model showed that all the variables used in the model were statistically significant at a 0.001 level of significance. This means that all the factors are statistically significant for the hourly accident frequency occurring in snowy conditions for the expressways. Next, this study analyzes the effect of each specific factor on the hourly accident frequency occurring on segments of the expressway under snow conditions as follows:

- The hourly traffic volume has a positive statistical effect on the hourly accident frequency occurring under snowy conditions for segments of the expressways. This means that when the hourly traffic volume on segments of the expressway increases, the hourly accident frequency increases under snowy conditions. This finding is similar to that in previous studies of accidents in snowy conditions (*Usman et al. 2010, Usman et al. 2011, and Seeherman et al. 2015)*.

- The percentage of trucks is also an important factor affecting the hourly accident frequency in snow conditions ($P < 2e\text{-}16$). The proportion of trucks' negative coefficient suggests that the number of trucks has a detrimental impact on the hourly accident frequency on expressway segments. In fact, truckers are often experienced and careful in driving on expressways in snowy conditions. This led to reducing the hourly accident frequency related to trucks.

- Average snowfall is also one of the factors in this model that have a positive effect on the likelihood of accidents in snow conditions on segments of the expressway. This result is completely consistent with the fact, that average snowfall will make the road surface covered with snow and reduce visibility to increase the risk of accidents. In addition, this result is similar to that found in other studies on traffic accidents in snowy conditions (*Qiu and Nixon 2008, Seeherman et al. 2015, and Heqimi et al. 2018)*.

- The negative coefficient of temperature has shown that the increase of temperature leads to a decrease in the hourly accident frequency on segments of the expressway. This finding is probably attributed to the fact that the amount of snowfall will increase when the temperature is reduced. This led to having good conditions for drivers in driving on expressways such as increased visibility and reduced snow cover on the road surface.

- In snow conditions, the average flow speed on each segment of the expressway can be increased with good weather conditions such as a decrease in snowfall, an increase in temperature, and a decrease of snow cover on the pavement. Therefore, it can reduce the likelihood of stacks and accidents on each segment of expressways. The result in the model also shows that the average flow speed also has a negative effect on the hourly accident frequency occurring on each segment of the expressway.

- In this model, segment type is a dummy variable that includes the divided segments and the undivided segments. This study compares the hourly accident frequency occurring between the divided segments and undivided segments. The result of this model shows that the hourly accident frequency occurring in the divided segment is higher than in the undivided segment. Because drivers often go slower and are more careful when driving through the undivided segments on the expressway under snowy conditions.

- Finally, the vertical gradient is also one of the factors found in the model to be statistically significant. The results show that the vertical gradient has a negative influence on the hourly accident frequency on each segment of the expressway in snowy conditions. This means that the higher the slopes on the segments, the smaller the hourly accident frequency of this segment.

Table 2 Parameter estimates of the model for hourly accident frequency on the expressway under snowy condition

Explanatory variables		Estimate	Std. Error	z value	Pr(>|z|)
Intercept		-6.985	0.242	-28.898	< 2e-16
Ln (hourly traffic volume)		0.727	0.039	18.613	< 2e-16
Percentage of trucks		-2.299	0.221	-10.411	< 2e-16
Average snowfall		1.381	0.243	5.687	1.29e-08
Temperature		-0.071	0.007	-9.755	< 2e-16
Average flow speed		-0.066	0.003	-24.587	< 2e-16
Segment type	Divided segments	0.821	0.091	9.013	< 2e-16
Vertical gradient		-0.083	0.018	-4.739	2.15e-06
2 x log-likelihood		-121,594			
Standard error		9.87e-07			

(3) Model evaluation

Firstly, the study conducts the residuals analysis of the probabilistic model. The results of descriptive statistics of residuals in this study are shown in Table 3 below. The results show that the mean of the raw residual and the Pearson residual is zero and the mean of the deviance residual is 0.014. In addition, the standard deviation of these residuals is relatively small. This shows that the predicted accident frequency is close to the observed accident frequency in this model.

Table 3 Descriptive Statistics about Residuals in the model

Residual	Mean	Standard deviation	Min	Max
r	0.000	0.019	-1.365	2.000
p	0.000	0.709	-0.012	36.392
d	-0.014	0.015	-0.049	2.639
r, Raw; p, Pearson; d, Deviance				

Next, Table 4 below displays the goodness of fit test of the fitted model for the hourly accident frequency occurring on segments of the expressway under snowy conditions. In this model, Akaike's information criterion (AIC) is 121,612, the residual deviance is 8,549 and the ratio of the log-likelihood index (ρ^2) is 0.19. This goodness of fit in the model helps determine how quality fit or appropriate the fitted model is as compared to other models.

Table 4 Goodness of fit for model

Criterion	Value
Akaike's information criterion (AIC)	121,612
Null deviance	10,512
Residual deviance	8,549
Ratio of log-likelihood index (ρ^2)	0.19

Finally, the study evaluates the predictive performance of the model using the MAD, MSPE, and RMSE indexes. Table 5 below shows the values of MAD, MSPE, and RMSE in the probabilistic model. These values are relatively small. This indicates that this model fits the data, or this model has good predictive performance for the hourly accident frequency occurring on segments of the expressway.

Table 5 Predictive performances evaluation measure

Measure	Value
The Mean Absolute Deviance (MAD)	0.0007
The Mean Squared Prediction Error (MSPE)	0.0004
The Root Mean Squared Error (RMSE)	0.0189

In addition, the study also considers the distribution of the expected number of accidents (μ_i) according to the different groups of the observed number of accidents (y_i) shown in Figure 4. The result shows that the distributions of the expected number of accidents don't have the difference between groups of the observed number of accidents. Therefore, the model in this study is a good model to predict hourly traffic accidents on expressways under snow conditions.

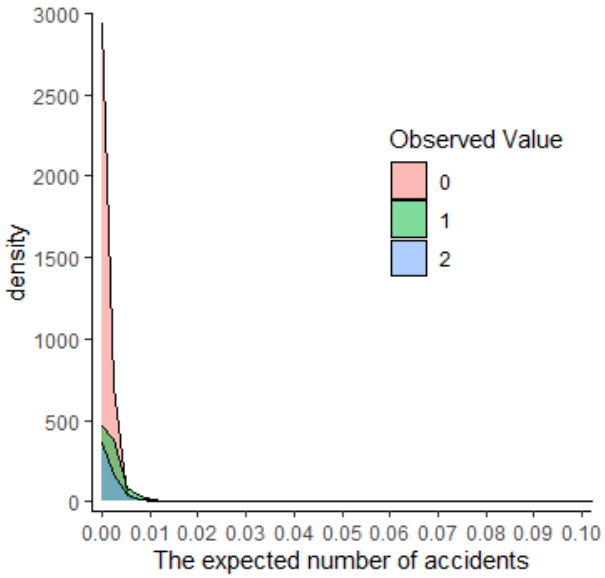

Fig.4 The distributions of the expected number of accidents (μi) according to groups of the obeserved number of accidents (*yi*)

6. CONCLUSIONS

Traffic accidents and stacks on expressways under snowy conditions are a challenge for state management agencies and expressway operators. This study has developed a probabilistic model based on a negative binomial regression model to help make appropriate policies and predict the hourly accident frequency on each segment of the expressway. This study used data on accidents and stacks occurring on expressways in Niigata Prefecture, Japan over 8 winter seasons (from December 2013 to March 2021).

The results of the model show that all the factors in the model are statistically significant at a 0.001 level of significance for the hourly traffic accident frequency on the expressway. Besides, the results of the model have shown that the hourly traffic volume and average snowfall have a positive effect on the likelihood of hourly traffic accidents. Meanwhile, factors including the percentage of trucks, average flow speed, temperature, and vertical gradients have a negative effect on hourly traffic accidents. In addition, the hourly accident frequency on the divided segments has a higher accident frequency than on the undivided segments on expressways under snowy conditions.

The study also evaluated the model based on the goodness of fit test, predictive performance evaluation, and residual analysis. The results show that the mean absolute deviation (MAD), the mean squared prediction error (MSPE) and the root mean square error (RMSE) are relatively small, therefore it can be concluded that this model has good predictive performances. In addition, the mean and standard deviation of the residuals is very small, which indicates that the predicted hourly accident frequency is close to the observed hourly accident frequency occurring on segments of expressways under snowy condition.

With the developed model, management agencies and expressway operators able to get more accurate and proactive actions with the purpose of mitigating accident occurring possibilities on the expressway, secondly can optimize resource allocation in case of an incident for rescue purposes. In conclusion, introducing the model to predict accident trends in harsh weather like winter can save human life and operation cost with huge margins.

However, there are some limitations to this study. Specifically, this study has not considered other factors that affect hourly traffic accident frequency on expressways such as snow depth on the road and road surface condition on each segment of the expressway by the hour. Therefore, future works will study the methodology of determining these factors by hour on each segment of the expressway to improve the accident prediction model for the expressway under snowy conditions.

ACKNOWLEDGMENT: The authors would like to thank the Japanese Government and the Japan International Cooperation Agency (JICA) for sponsoring this research in the master's program at the Nagaoka University of Technology. In addition, the authors would also like to thank Nippon Expressway Company Limited (NEXCO) of Japan for providing the data for this research.

REFERENCES

1) Abdel-Aty, Mohamed A., and A. Essam Radwan. 2000. “Modeling Traffic Accident Occurrence and Involvement.” *Accident Analysis & Prevention* 32(5):633–42. doi:

10.1016/S0001-4575(99)00094-9.
2) Asano, Motoki. 2003. “Characteristics of Traffic Accidents in Cold, Snowy Hokkaido, Japan.”
3) CHENGYE, Pan, and Prakash RANJITKAR. 2013. “Modelling Motorway Accidents Using Negative Binomial Regression.” *Journal of the Eastern Asia Society for Transportation Studies* 10:1946–63.
4) Daniels, Stijn, Tom Brijs, Erik Nuyts, and Geert Wets. 2010. “Explaining Variation in Safety Performance of Roundabouts.” *Accident Analysis and Prevention* 42(2):393–402. doi: 10.1016/j.aap.2009.08.019.
5) Gaweesh, Sherif M., Mohamed M. Ahmed, and Annalisa V. Piccorelli. 2019. “Developing Crash Prediction Models Using Parametric and Nonparametric Approaches for Rural Mountainous Freeways: A Case Study on Wyoming Interstate 80.” *Accident Analysis and Prevention* 123(March 2018):176–89. doi: 10.1016/j.aap.2018.10.011.
6) Heqimi, Gentjan, Timothy J. Gates, and Jonathan J. Kay. 2018. “Using Spatial Interpolation to Determine Impacts of Annual Snowfall on Traffic Crashes for Limited Access Freeway Segments.” *Accident Analysis & Prevention* 121:202–12. doi: 10.1016/J.AAP.2018.09.014.
7) Hyodo, Satoshi, and Kenta Hasegawa. 2021. “Factors Affecting Analysis of the Severity of Accidents in Cold and Snowy Areas Using the Ordered Probit Model.” *Asian Transport Studies* 7(December 2020):100035. doi: 10.1016/j.eastsj.2021.100035.
8) Johansson, Per. 1996. “Speed Limitation and Motorway Casualties: A Time Series Count Data Regression Approach.” *Accident Analysis and Prevention* 28(1):73–87. doi: 10.1016/0001-4575(95)00043-7.
9) Joshua, Sarath C., and Nicholas J. Garber. 1990. “Estimating Truck Accident Rate and Involvements Using Linear and Poisson Regression Models.” *Transportation Planning and Technology* 15(1):41–58. doi: 10.1080/03081069008717439.
10) Jovanis, Paul P., and Hsin Li Chang. 1986. “Modeling the Relationship of Accidents To Miles Traveled.” *Transportation Research Record* 42–51.
11) Kim, Daeseong, Sangyun Jung, and Sanghoo Yoon. 2021. “Risk Prediction for Winter Road Accidents on Expressways.” *Applied Sciences (Switzerland)* 11(20):1–12. doi: 10.3390/app11209534.
12) Li, Xiugang, Dominique Lord, Yunlong Zhang, and Yuanchang Xie. 2008. “Predicting Motor Vehicle Crashes Using Support Vector Machine Models.” *Accident Analysis & Prevention* 40(4):1611–18. doi: 10.1016/J.AAP.2008.04.010.
13) Lord, Dominique, and Fred Mannering. 2010. “The Statistical Analysis of Crash-Frequency Data: A Review and Assessment of Methodological Alternatives.” *Transportation Research Part A: Policy and Practice* 44(5):291–305. doi: 10.1016/j.tra.2010.02.001.
14) Ma, Zhuanglin, Honglu Zhang, Steven I. J. Chien, Jin Wang, and Chunjiao Dong. 2017. “Predicting Expressway Crash Frequency Using a Random Effect Negative Binomial Model: A Case Study in China.” *Accident Analysis & Prevention* 98:214–22. doi: 10.1016/J.AAP.2016.10.012.
15) Miaou, Shaw Pin, and Joon Jin Song. 2005. “Bayesian Ranking of Sites for Engineering Safety Improvements: Decision Parameter, Treatability Concept, Statistical Criterion, and Spatial Dependence.” *Accident Analysis and Prevention* 37(4):699–720. doi: 10.1016/j.aap.2005.03.012.
16) Oh, Jutaek, Simon P. Washington, and Doohee Nam. 2006. “Accident Prediction Model for Railway-Highway Interfaces.” *Accident Analysis and Prevention* 38(2):346–56. doi: 10.1016/j.aap.2005.10.004.
17) Omari-Sasu, A. Y., Adjei Mensah Isaac, and R. K. Boadi. 2016. “Statistical Models for Count Data with Applications to Road Accidents in Ghana.” *International Journal of Statistics and Applications* 6(3):123–37. doi: 10.5923/j.statistics.20160603.05.
18) Pennelly, Clark, Gerhard W. Reuter, and Stevanus Tjandra. 2018. “Effects of Weather on Traffic Collisions in Edmonton, Canada.” *Atmosphere - Ocean* 56(5):362–71. doi: 10.1080/07055900.2018.1548344.
19) Qiu, Lin, and Wilfrid A. Nixon. 2008. “Effects of Adverse Weather on Traffic Crashes: Systematic Review and Meta-Analysis.” *Transportation Research Record* (2055):139–46. doi: 10.3141/2055-16.
20) Saha, Promothes, Mohamed M. Ahmed, and Rhonda Kae Young. 2015. “Safety Effectiveness of Variable Speed Limit System in Adverse Weather Conditions on Challenging Roadway Geometry.” *Transportation Research Record: Journal of the Transportation Research Board* 2521(1):45–53. doi: 10.3141/2521-05.
21) SANO, Kazushi, Touru INAGAKI, Jouji NAKANO, and Nguyen Cao Y. 2010. “An Analysis on Traffic Accidents on Undivided Expressway in Cold and Snow Area.” *Journal of the Eastern Asia Society for Transportation Studies* 8:2048–61. doi: 10.11175/EASTS.8.2048.
22) Seeherman, Joshua, and Yi Liu. 2015. “Effects of Extraordinary Snowfall on Traffic Safety.” *Accident Analysis and Prevention* 81:194–203. doi: 10.1016/j.aap.2015.04.029.
23) Shankar, Venkataraman N., Gudmundur F. Ulfarsson, Ram M. Pendyala, and Mary Lou B. Nebergall. 2003. “Modeling Crashes Involving Pedestrians and Motorized Traffic.” *Safety Science* 41(7):627–40. doi: 10.1016/S0925-7535(02)00017-6.
24) Shrestha, Noora. 2020. “Detecting Multicollinearity in Regression Analysis.” *American Journal of Applied Mathematics and Statistics* 8(2):39–42. doi: 10.12691/ajams-8-2-1.
25) Silva, Philippe Barbosa, Michelle Andrade, and Sara Ferreira. 2020. “Machine Learning Applied to Road Safety Modeling: A Systematic Literature Review.” *Journal of Traffic and Transportation Engineering (English Edition)* 7(6):775–90. doi: 10.1016/j.jtte.2020.07.004.
26) Usman, Taimur, Liping Fu, and Luis F. Miranda-Moreno. 2010. “Quantifying Safety Benefit of Winter Road Maintenance: Accident Frequency Modeling.” *Accident Analysis and Prevention* 42(6):1878–87. doi: 10.1016/j.aap.2010.05.008.
27) Usman, Taimur, Liping Fu, and Luis F. Miranda-Moreno. 2011. “Accident Prediction Models for Winter Road Safety: Does Temporal Aggregation of Data Matter?” *Transportation Research Record* (2237):144–51. doi: 10.3141/2237-16.
28) Wong, Andy H., and Tae J. Kwon. 2021. “Development and Evaluation of Geostatistical Methods for Estimating Weather Related Collisions: A Large-Scale Case Study.” *Transportation Research Record: Journal of the Transportation Research Board* 036119812110200. doi: 10.1177/03611981211020008.
29) Zwilling, Michael. 2013. “Negative Binomial Regression.” *The Mathematica Journal* 15:1–18. doi: 10.3888/tmj.15-6.

(Received September 30, 2022)