A Probabilistic Model Development to Predict Traffic Accident Tendency under Snowy Conditions for Expressway

Bui Tien Manh

Supervisor: Prof. **Kazushi SANO**

Urban Transport Laboratory

1. Introduction

1.1. Background

Road traffic accidents occurred 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. According to statistics from the US, Canada, and Japan and recent studies have shown that the frequency of traffic accidents tends to increase under snowy conditions. The study aims to develop a probabilistic model to analyse the factors and predict traffic accident under snowy conditions on expressways.

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). According to the analysis of the accident dataset on five expressways in Niigata prefecture (2012-2020), 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 and are shown in Figure 1.

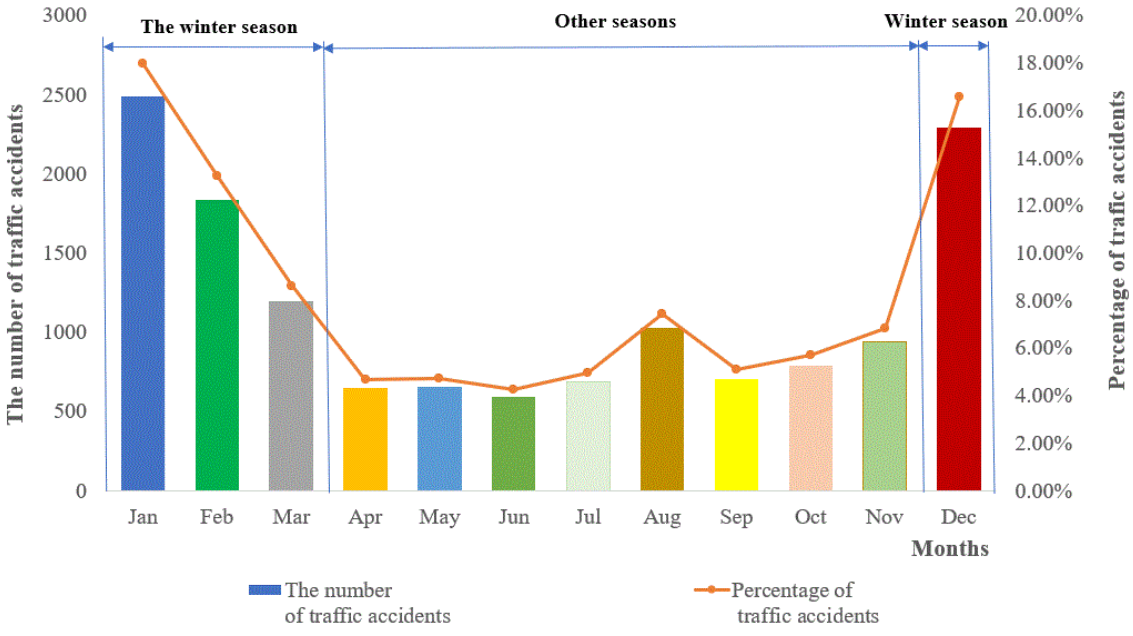

Figure 1. Number and percentage of accidents occurring on expressways in Niigata prefecture, Japan (2012-2020)

At present, there have not been many studies on developing models to analyse influencing factors and predict traffic accident trends for expressways that consider snowy conditions in the winter season. Besides, these studies have not considered factors such as average flow speed, and segment type, and identified the factors affecting accidents in snowy conditions. In addition, the influencing factors are aggregated over a long period of time (years or days), which leads to decrease accuracy in forecasting and analysing the impact of factors on accidents under winter conditions.

1.2. Novelty of Research

The novelty of this study will develop a probabilistic model based on traditional statistic model and machine learning to analyse factors effecting the tendency of predict the tendency of hourly accident and predict accident frequency occurring on expressways in short-term (hour) that consider snow conditions. Besides, the study also considers factors affecting accidents in a short time (hour and minutes) to ensure the accuracy of the probabilistic model.

1.3. Research objective

This study includes some objectives as follows:

- Determines multicollinearity in the model based on correlation analysis and analyses factors affecting the traffic accident tendency based on NB models.

- Predicts the number of accidents occurring in the short-term (hour) on expressways under snowy conditions based on the NB model and ANN model.

- Finally, compare the performance between the NB model and ANN model to propose a suitable model.

2. Methodology

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 the probabilistic model that includes Negative Binomial Regression Model (NB) in the traditional statistic model and Artificial Neural Network (ANN) in Machine Learning algorithms to analyse factors effecting the accident tendency and predict the accident frequency occurring on expressways in short-term (hour) under snowy conditions. Finally, the study evaluates and compares the fit of two models based on the mean absolute error (MAE), the mean squared error (MSE), and the root mean square error (RMSE).

2.1. Correlation analysis

This study uses correlation analysis to determine the correlations between the variables that used the probabilistic model to avoid multicollinearity. The sample correlation coefficient of X and Y, denoted r, is calculated by the following formula:

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

Correlation coefficient values range -1 to +1. The closer to 1 the correlation coefficient gets the "stronger" the correlation. If the absolute value of the correlation coefficient is close to 0.8, collinearity is likely to exit.

2.2. Negative binomial regression model (NB)

The negative binomial regression model with probability density function is as follows:

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

Where:

- $\Pr\left(Y = y_i|\mu_i,\alpha\right)$ is the probability of accidents y_i occurring segment i^{th} over the hour j^{th}.

- y_i: is the number of accidents occurring to a given segment i^{th} during the hour j, y_i = 0,1, 2.

- μ_i is the expected number of accidents of segment i^{th} during the hour j. In this study, μ_i follows the below function:

$$\mu_{ij} = e^{\beta_0} L_i \times TV_i^{\beta_1} \times e^{\left(\beta_2 X_{2i} + \ldots\ldots\ldots + \beta_k X_{ki} + \varepsilon_i\right)}$$

- L_i: is the length of segment i^{th}.

- TV_i and X_{ki}: is the hourly traffic volume and other explanatory variables at given segment i^{th} during the hour j^{th}, respectively.

- β_k and α: is coefficients of models and the dispersion parameter, respectively. These parameters are estimated by maximizing the log-likelihood function.

2.3. Artificial Neural Network (ANN)

ANN model is designed with an input layer, an output layer and two hidden layers. The input layer consists of 8 neurons that are factors effecting traffic accidents. The output layer consists of one neuron which is the hourly accident frequency. There are 2 hidden layers with 5 neurons and 3 neurons, respectively. Between layers in ANN model are connected full connections by weights, bias, and activation functions (the leaky rectified linear activation unit). The structure of the ANN model is shown in Figure 2 below.

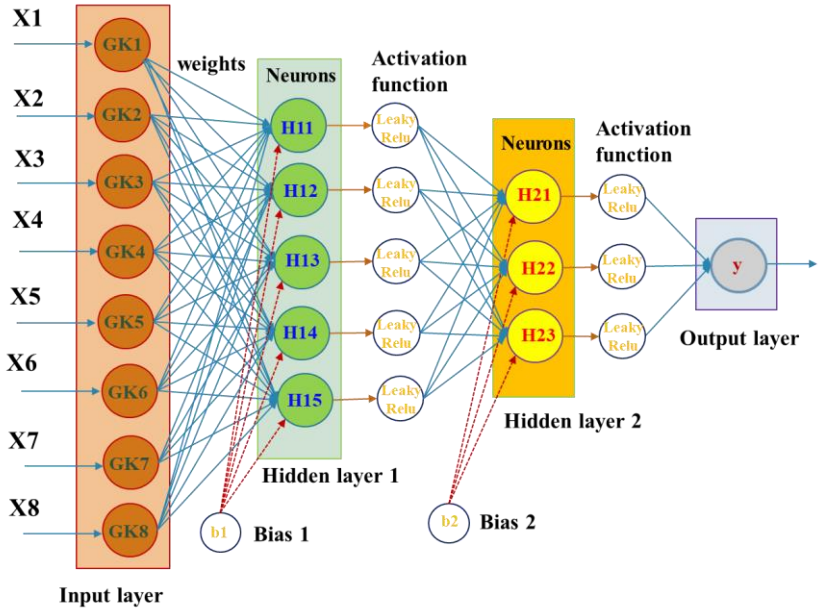

Figure 2. The structure of ANN model

The goal of the training processing is to find the weight and bias such that make the loss function (mean squared error) as small as possible. To achieve this goal, this study used stochastic gradient descent and the backpropagation algorithm in training process. However, before conducting training processing for ANN model, this study normalizes the input and output data using the minimax algorithm.

2.4. Model evaluation

To evaluate the predictive ability of the model, this study used main metrics which are the mean absolute error (MAE), the mean squared error (MSE), and the root mean square error (RMSE). The desirable model should fit the data as closely as possible, which is presented by the smaller value of MAE, MSE, and RMSE. Their formula is shown as follows:

$$MAE = \frac{\sum_{i}^{n} |y_i - y_i|}{n}, \quad MSE = \frac{\sum_{i}^{n} \left(y_i - y_i\right)^2}{n}, \quad RMSE = \sqrt{MSE}$$

Where: y_i, y_i are the observed accident frequency, the predicted accident frequency, respectively. n the size of the set test.

3. Study scope and data

The scope of this study focuses on traffic accidents and stacks occurring on five major expressways which covers the entire Niigata Prefecture. 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. After collecting the datasets, this study synthesized and matched the datasets into a common dataset based on spatial (segments) and temporal (hour). The data processing is performed by R software. Descriptive statistics of explanatory variables used in model are shown in Table 1 below.

Table 1. Descriptive statistics about explanatory variables in model

Explanatory variables	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
Truck percentage (%)	0	50	25	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
Explanatory variables	**Category**		**Number of Segments**	
Segment type	Un-divided		174 (19.98%)	
	Divided		697 (80.02%)	

4. Results and Discussions

4.1. Results of correlation analysis

The results of the correlation analysis show that the correlation coefficient between the explanatory variables in the model which are shown in heat map (Figure 4) below ranges from -0.35 to 0.376. This indicates that between explanatory variables have a weak correlation and do not cause multicollinearity for the model.

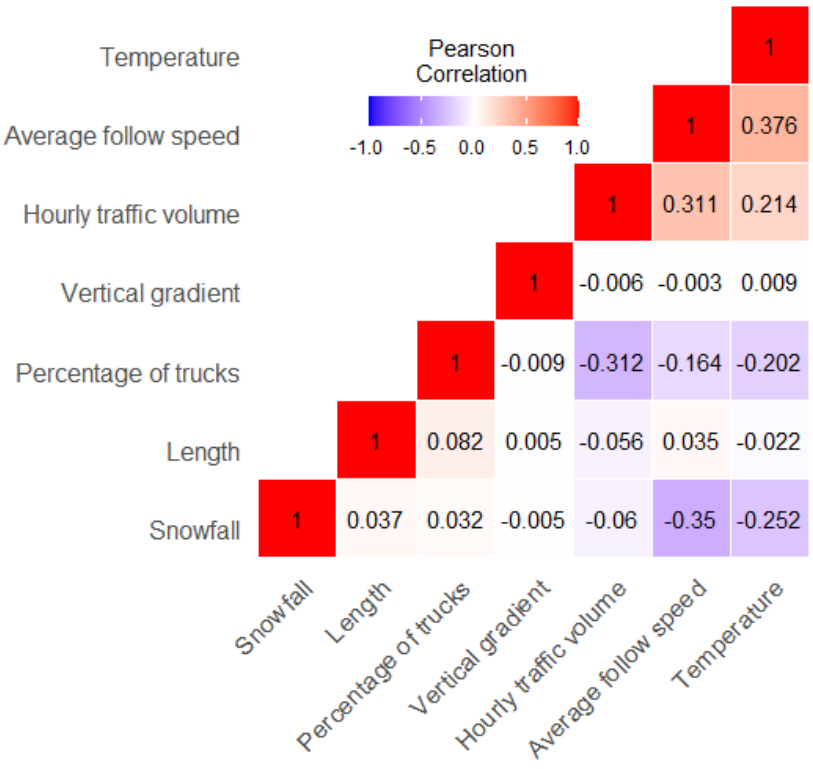

Figure 3. Correlation between factors in the model

4.2. Results of Negative binomial regression model (NB)

After analysing the correlation between the variables in the model, the study trains NB model that used the training dataset (80%). The result of training this model is presented in Table 2. In additional, this table 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 97,819, the residual deviance is 8,423, and the ratio of the log-likelihood index ($\rho 2$) is 0.19.

Table 2. Parameter estimates of the NB model.

Explanatory variables	Estimate	Std. Error	z value	Pr(>|z|)
Intercept	-7.041	0.271	-26.025	< 2e-16
Ln (Hourly traffic volume)	0.734	0.044	16.811	< 2e-16
Truck percentage	-2.281	0.247	-9.242	< 2e-16
Average snowfall	1.382	0.272	5.083	3.72e-08
Temperature	-0.073	0.008	-8.952	< 2e-16
Average flow speed	-0.652	0.003	-21.770	< 2e-16
Segment type Divided segments	0.787	0.101	7.783	7.07e-15
Vertical gradient	-0.086	0.020	-4.405	1.06e-06
2 x log-likelihood	-97,819			
Standard error	1.10e-06			
AIC	97,837			
Null deviance	6,856			
Residual deviance	8,423			
Ratio of log-likelihood index	0.19			
Observations	16,221,504			

The results of NB model show that all the factors in the model are statistically significant at the significance level of 1%. Besides, the results of this model have shown that the hourly traffic volume and average snowfall have a positive effect on the expected number of traffic accidents occurring in short-term (hour). Meanwhile, factors including truck percentage, average flow speed, temperature, and vertical gradients have a negative effect on the expected number of traffic accidents occurring in short-term. In addition, the expected number of traffic accidents occurring in short-term on the divided segments has higher than those on the undivided segments on expressways under snowy conditions. Finally, the average snowfall is one of risk factors that has

the highest effect on the expected number of traffic accidents occurs in short term on each segment under snowy conditions.

The percentage change of the expected number of traffic accidents occurring in short-term on each segment of expressways under snowy conditions when the risk factors (variables) except variables, hourly traffic volume, in the NB model change one unit is summarized in Table 3.

Table 3. The percentage change of the expected number of traffic accidents

Explanatory variables	Coefficients (β_i)	Exp(βi)	Percentage Change (%)	Std. Error	Lower 95.0% Confidence Limit	Upper 95.0% Confidence Limit
Intercept	-7.041	0.001	-99.9%	31.1%	-160.9%	-38.9%
Truck percentage	-2.281	0.102	-89.8%	28.0%	-144.7%	-34.9%
Average snowfall	1.382	3.983	298.3%	31.3%	237.0%	359.6%
Temperature	-0.073	0.930	-7.0%	0.8%	-8.6%	-5.5%
Average follow speed	-0.652	0.521	-47.9%	0.3%	-48.5%	-47.3%
Segment type Divided segments	0.787	2.197	119.7%	10.6%	98.8%	140.5%
Vertical gradient	-0.086	0.918	-8.2%	2.0%	-12.2%	-4.3%

4.3. Results of Artificial Neural Network (ANN)

ANN model is trained in training dataset and validation dataset (80%). This training process was conducted with 100 epochs with loss function (mean squared error) and mean absolute error (MAE). The results of the training process are shown in Figure 4 below.

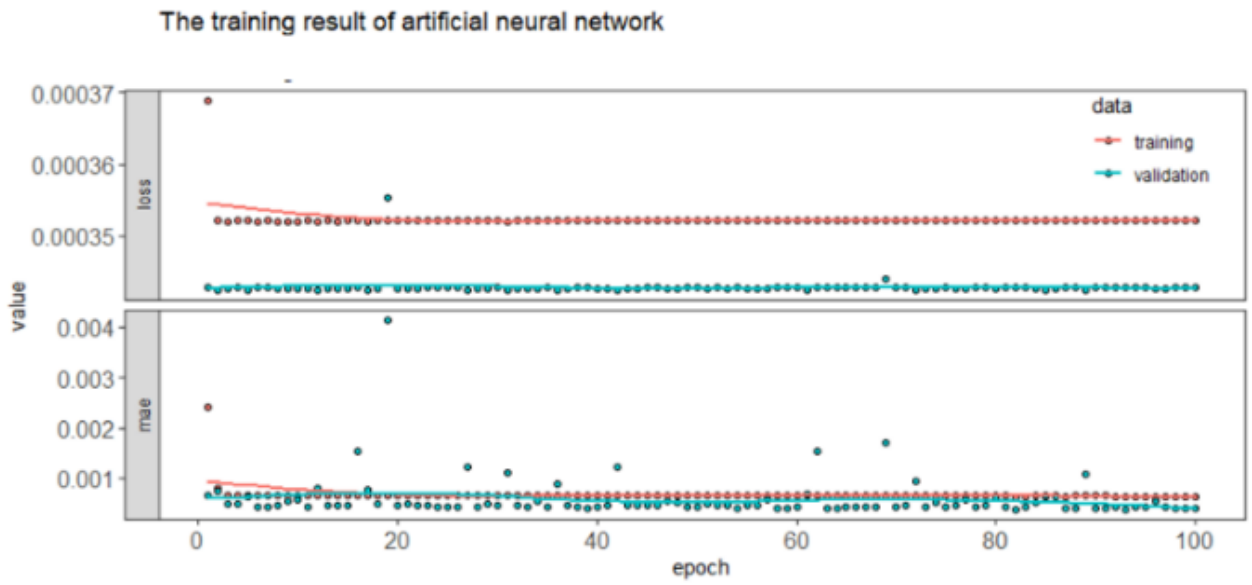

Figure 4. The training processing for ANN model

The results of the training process are as follows:

- The values of the loss function (MSE), and MAE in the training process are approximately zero. This indicates that the predicted accident frequency closes with the actual accident frequency.

- The values of the loss function (MSE) and mean square error (MAE) in the validation process are close to the value in the training process. Therefore, the training process of the ANN model is relatively good.

After the training process, ANN model was evaluated by a test dataset. The results show that the values of MAE, MSE, and RMSE of the test process are relatively small and close to the value of the training process of the model. This indicates that the model is a good model or has high accuracy.

4.4. Comparison results between two models.

The study compared the predictive performance of two models based on MAE, MSE, and RMSE on a test dataset (20%) are shown in Table 4 below.

Table 4. Comparison results between two models

Measure	NB Model	ANN Model
The Mean Absolute Error (MAE)	0.0006	0.0004
The Mean Squared Error (MSE)	0.0004	0.0003
The Root Mean Squared Error (RMSE)	0.02	0.0173

The study has shown some comparative results between the two models as follows:

- The values of MAE, MSE, and RMSE are small in both models. This indicates both models have the good performance or high accuracy.

- The values of MAE, MSE, and RMSE in ANN model is less than NB model. Therefore, ANN model is better than the NB model.

5. Conclusions and Limitations

The purpose of this study was to analyse the affected factors and predict the accident frequency occurring in short-term (hour). This helps expressway operators make appropriate policies to reduce traffic accidents on expressways under snowy conditions. The results of this study are as flows:

- Results of correlation analysis show that the correlation coefficient between variables in the model ranges from -0.35 to 0.376. This indicates between variables in the model have a weak correlation and do not cause multicollinearity for the regression model.

- The results of the NB model have shown that all the factors are statistically significant at the significance level of 1%. Besides, the hourly traffic volume and average snowfall have a positive effect on the expected number of traffic accidents occurring in short-term (hour) on each segment of expressways under snowy conditions. Meanwhile, risk factors including truck percentage, average flow speed, temperature, and vertical gradients have a negative effect on the expected number of traffic accidents occurring in short-term on each segment of expressways. In addition, the expected number of traffic accidents occurring in short-term on each divided segment tends to have higher than those on each undivided segment of on expressways under snowy conditions. Finally, the average snowfall is one of risk factors that has the highest effect on the expected number of traffic accidents occurs in short term on each segment under snowy conditions.

- Besides, this study designed a structure of ANN model with 1 input layer, an output layer, and two hidden layers. The training process of the ANN model is relatively good. After training process, ANN model is tested by MAE, MSE, and RMSE index on a test dataset. The results indicates that ANN model is a good model or has high accuracy. Finally, this study compared the performance of both models. The results show that both models have good performance or high accuracy. However, ANN model is better than NB model.

However, there are some limitations in this study related to considering other factors such as snow depth on the road, road surface condition, the driver's age, the driver's experience, and the daytime and night time effecting traffic accident under snowy conditions; solving an excess zero in dataset; and using different optimization algorithms in training ANN model.