Available online at www.sciencedirect.com
SCIENCE DIRECT
Safety Science 43 (2005) 541–557
SAFETY SCIENCE
www.elsevier.com/locate/ssci

# Analysis of freeway accident frequencies: Negative binomial regression versus artificial neural network

Li-Yen Chang *

*Graduate Institute of Transportation and Logistics, National Chia-Yi University, 300 University Road, Chia-Yi 600, Taiwan*

Received 27 October 2004; received in revised form 13 April 2005; accepted 17 April 2005

**Abstract**

The Poisson or negative binomial regression model has been employed to analyze vehicle accident frequency for many years. However, these models have the pre-defined underlying relationship between dependent and independent variables. If this assumption is violated, the model could lead to erroneous estimation of accident likelihood. In contrast, the artificial neural network (ANN), which does not require any pre-defined underlying relationship between dependent and independent variables, has been shown to be a powerful tool in dealing with prediction and classification problems. Thus, this study employs a negative binomial regression model and an ANN model to analyze 1997–1998 accident data for the National Freeway 1 in Taiwan. By comparing the prediction performance between the negative binomial regression model and ANN model, this study demonstrates that ANN is a consistent alternative method for analyzing freeway accident frequency.


*Keywords:* Accident frequency; Artificial neural network; Negative binomial regression; Freeway

## 1. Introduction

The impact that traffic accidents have on society is significant. For example, there are approximately 3000 people are killed and thousands more injured by traffic accidents in

* Tel.: +886 5 271 7982.
*E-mail address:* liyen@mail.ncyu.edu.tw

0925-7535/$ - see front matter © 2005 Elsevier Ltd. All rights reserved.
doi:10.1016/j.ssci.2005.04.004

Taiwan each year. Individuals injured (or killed) in traffic accidents must deal with pain and suffering, medical costs, wage loss, higher insurance premium rates, and vehicle repair costs. For society as a whole, traffic accidents result in enormous costs in terms of lost productivity and property damage. Therefore, public agencies have put much effort into preventive measures, such as illumination and enforcement. However, the annual number of traffic accidents has not yet significantly decreased. Therefore, there should be further research studies on the risk factors for traffic accidents. This study focuses on the non-behavioral factors of freeway accident risk, specifically highway geometric characteristics and environmental conditions. Although past statistics indicated that most traffic accidents resulted from drivers' errors (behavioral factors), non-behavioral factors also play an important role in traffic safety. Not only can they contribute to certain types of driver errors (e.g., speeding often occurs at downgrades), but accidents will be likely to occur at the same location repeatedly if the problem is not mitigated. In addition, with a better understanding of non-behavioral factors of freeway accidents, transportation engineers will be able to design freeways to higher safety standards.

Past research analyzing accident frequencies mainly relied on statistical models such as linear regression models, Poisson regression or/and negative binomial regression models because the occurrence of accidents on a highway section can be regarded as a random event. Another major advantage of applying these statistical models is the ability to identify a broad range of risk factors that can contribute significantly to accidents. However, most of the statistical models have their own model assumptions and pre-defined underlying relationships between dependent and independent variables. If these assumptions are violated, the model could lead to erroneous estimation of accident likelihood. Artificial neural networks (ANN) which do not require any pre-defined underlying relationship between dependent and independent variables have been widely employed in financial analysis, decision problems, and pattern recognition. The ANN has been shown to be a powerful tool, particularly in dealing with prediction and classification problems. There has also been an increased interest in applying ANN in the field of transportation since the 1990s, such as driver behavior analysis, pavement maintenance, vehicle detections, and so on (Dougherty, 1995). However, the applications of ANN to analyze traffic safety problems have been relatively few. Therefore, this study examines whether ANN can be used to analyze the relationship between risk factors and accidents. This is done by evaluating the prediction performance between the negative binomial regression model and ANN model. The paper begins with a review of previous literature on modeling accident frequencies and then presents the methodological approach. A description of the available data and an assessment of the model estimation results follow this. The paper concludes with a summary and directions for future research.

## 2. Literature review

Past research on modeling accident frequencies has been diverse, both empirically and methodologically. From an empirical standpoint, most research studies (Shankar et al., 1995; Milton and Mannering, 1998; Carson and Mannering, 2001) have focused on non-behavioral risk factors of accidents on the freeway or arterial roadways. These non-behavioral factors included highway geometry (e.g., horizontal and vertical alignments, and shoulder width), traffic characteristics (e.g., average annual daily traffic (AADT) and percentage of trucks) and weather conditions (e.g., rain or snow). The

findings indicated that number of lanes, narrow shoulder width, vertical grade, horizontal curves, AADT, amount of snowfall and their interaction can have significant influence on vehicle accidents. McCarthy (1999) focused on the effectiveness of public policy (e.g., traffic regulations, alcohol availability and enforcement) in reducing fatal accidents in individual cities in California. The results indicated that fatal accidents were significantly reduced due to traffic enforcement, but little effect was found from speed limit and seat belt use laws.

From a methodological standpoint, the most common approach to analyze accident frequencies for a specified roadway segment is to apply Poisson or negative binomial regression models because of the distributional property (i.e., random, discrete and non-negative) of vehicle accidents (Milton and Mannering, 1998). Although the Poisson regression model has desirable statistical properties for describing vehicle accidents, it has an important constraint, which is that the mean and variance of the accident data are constrained to be equal. To overcome this constraint, the negative binomial regression model, which allows this constraint to be relaxed, has been widely employed to analyze vehicle accidents (Miaou, 1994; Poch and Mannering, 1996; Hadi et al., 1995; Shankar et al., 1995; McCarthy, 1999; Carson and Mannering, 2001). The findings suggested that most of vehicle accident data tend to be overdispersed (i.e., the variance will likely be significantly greater than the mean) and negative binomial modeling is an appropriate technique for exploring the frequency of accidents. In addition, zero-inflated Poisson and zero-inflated negative binomial models were also employed to analyze accident frequencies by Shankar et al. (1997), Lee and Mannering (2002) and Lee et al. (2002) to deal with the overdispersion problem potentially caused by the excessive zeroes (i.e., no accidents being observed) in traffic accident data. The application of zero-altered counting processes allows modeling roadway section accident frequencies in two states: the zero-accident state (where no accidents will be ever observed), and the accident state (where accident frequencies follow some known distribution, such as the Poisson or negative binomial distribution). The findings showed that the zero-altered probability process provides great flexibility in uncovering processes affecting accident frequencies on roadway sections with zero accidents and those observed with accidents. In terms of model selection for analyzing accident frequencies, Miaou (1994) and Lee et al. (2002) recommended that the Poisson regression model is estimated as an initial model. If the overdispersion of accident data is found, both negative binomial and zero-inflated count models could be considered.

ANN has also been employed to analyze transportation problems for many years. According to the review by Dougherty (1995), most studies have focused on modeling driver behavior, parameter estimation, pavement maintenance, and vehicle detection. More recent applications in the transportation field using ANN included traffic prediction (Yin et al., 2002; Zhong et al., 2004), traffic parameters estimation (Tong and Hung, 2002), traffic signal control (Zhang et al., 2001), incident detection (Jin et al., 2002; Yuan and Cheu, 2003), travel behavior analysis (Subba Rao et al., 1998; Hensher and Ton, 2000; Vythoulkas and Koutsopoulos, 2003), vehicle emissions (Shiva Nagendra and Khare, 2004) and traffic accident analysis (Mussone et al., 1996; Mussone et al., 1999; Sohn and Lee, 2003; Abdel-Aty and Pande, 2005). For example, Tong and Hung (2002) employed a three-layer neural network to estimate vehicle discharge headway. The results showed that the ANN model could produce reasonably discharge headway estimates for individual vehicles. Subba Rao et al. (1998) used ANN to model the choice behavior with respect to access mode for transit. The performance of ANN was found to be

superior to multinomial logit model in both calibration and prediction. These studies have provided general insight into the performance of ANN models. For analyzing traffic safety problems, there have been relatively few applications using ANN. Mussone et al. (1996) employed a three-layer neural network to estimate the accident probability using accident data in Italy. Accident sites, road features, weather and roadbed conditions, human error, as well as vehicular and environmental factors were used as input variables. "Carelessness" and "excessive speed" were found to be the risk factors in defining accident probability. Mussone et al. (1999) employed ANN modeling approach to analyze the degree of danger of urban intersections and demonstrated that ANN is a good alternative method for analyzing the factors contributing to intersection accidents. Abdel-Aty and Pande (2005) applied a probabilistic neural network (PNN) model to predict crash occurrence on the Interstate-4 corridor in Orlando. Average and standard deviation of speed around crash sites were extracted from loop data as input variables. The analysis results showed that at least 70% of the crashes can be correctly identified by the proposed PNN model.

## 3. Empirical setting

The study area for this paper is National Freeway 1, which is the most important transportation corridor in Taiwan. National Freeway 1 is a 373 km tolled freeway with 47 interchanges and 10 mainline toll plazas. Illumination is provided only at interchange areas, toll plaza areas and locations with severe geometric changes such as severe downhill or uphill grades.

To investigate the relationship between vehicle accidents and highway geometry, traffic characteristics and environment conditions, data from a number of resources were collected. The vehicle accident data were taken from the National Traffic Accident Investigation Reports provided by the Ministry of Transportation and Communications. The data were obtained in a computer-ready form, which included coded information on reported accidents on National Freeway 1. Information on the accidents that occurred in the period from 1997 to 1998 was extracted for this study. The primary resources of highway geometric design information and traffic data were obtained from the Taiwan Area National Freeway Bureau. The highway geometric design information includes number of lanes, lane width, horizontal curvature, vertical grade and others; while traffic information includes average daily traffic (ADT) of various vehicle types, peak hour factors, and traffic distribution over lanes. Weather information was taken from the annual report of climatological data. This report records detailed weather information of cities and towns along National Freeway 1 including pressure, temperature, humidity, precipitation, wind speed, and cloudiness.

With these data, the next step is how to divide the study area into manageable roadway sections. The most common alternatives adopted in previous studies for determining roadway section length include the use of fixed-length sections or homogeneous sections (in terms of geometric characteristics). In order to account directly for the effects of highway geometric characteristics on accident frequencies, homogeneous sections in terms of number of lanes, horizontal curvature, and vertical grade were considered in this study. A more detailed discussion on the advantages and disadvantages of these two alternatives can be found in Shankar et al. (1995). According to this approach, 373 km of freeway were first disaggregated into 498 sections. Because of the opposite vertical alignment and different traffic conditions in the northbound and southbound roadway sections, these 498 sections

Table 1
Sample summary of statistics of characteristics of road sections

| | Minimum | Maximum | Mean | Standard deviation |
|---|---|---|---|---|
| Accident frequency (per year) | 0 | 7 | 0.67 | 1.00 |
| Section length (km) | 0.1 | 4.2 | 0.75 | 0.52 |
| Degree of horizontal curve (angle, in degree, subtended by a 100 m arc, equal to 18,000/($\pi \times$ radius)) | 0 | 15.68 | 2.18 | 2.46 |
| Vertical grade (%) | −5.3 | 5.3 | 0 | 1.58 |
| ADT per lane (in 1000's of vehicles) | 12.88 | 42.13 | 20.96 | 4.14 |
| Trucks percentage | 0.86 | 18.45 | 10.13 | 4.16 |
| Bus percentage | 0.67 | 11.31 | 3.53 | 1.27 |
| Peak hour factor (PHF) | 0.77 | 0.97 | 0.91 | 0.04 |
| Number of days with precipitation | 56 | 224 | 111.1 | 37.5 |
| Annual precipitation (mm) | 1439 | 5773 | 2165.6 | 790.7 |

were further disaggregated into 996 sections (i.e., northbound and southbound roadway sections considered separately). During the 1997–1998 study period, there were 1338 accidents resulting in deaths and/or injuries. The summary statistics of these 1992 highway sections (i.e., each section produces two observations) are presented in Table 1. The observed accident frequency on these freeway sections ranges from 0 to 7, and the average frequency is 0.67. In order to be able to compare the prediction performance between the statistical model and ANN model, the collected data were randomly divided into two subsets, one for training and the other for testing. The number of sections used for model estimation is 1500, and the number of sections used for testing is 492. A Chi-squared test shows that the accident frequency distributions between the two sub-samples are not significantly different.

## 4. Negative binomial modeling approach to freeway accident frequencies

This study models the number of accidents that occurred on a highway section over a one-year time period. Because accident frequencies on a highway section are discrete and non-negative integer values, the Poisson regression technique is a natural first choice for modeling such data. However, past research has indicated that accident frequency data are likely to be overdispersed and has suggested using the negative binomial regression model as an alternative. Deriving the negative binomial regression model can start with a Poisson model, which is defined by the following equation:

$$P(n_i) = \frac{\lambda_i^{n_i} \exp(-\lambda_i)}{n_i!} \tag{1}$$

where $P(n_i)$ is the probability of n accidents occurring on a highway section $i$ over a one-year time period, and $\lambda_i$ is the expected accident frequency (i.e., $E(n_i)$) for highway section $i$. When applying the Poisson model, the expected accident frequency is assumed to be a function of explanatory variables such that

$$\lambda_i = \exp(\boldsymbol{\beta} \mathbf{X}_i) \tag{2}$$

where $\mathbf{X}_i$ is a vector of explanatory variables that could include the geometry, traffic characteristics, and weather conditions of highway section $i$ that determine accident frequency;

and $\boldsymbol{\beta}$ is a vector of estimable coefficients. With this form of $\lambda_i$, the coefficient vector $\boldsymbol{\beta}$ can be estimated by the maximum likelihood method with the likelihood function being

$$L(\beta) = \prod_i \frac{\exp[-\exp(\boldsymbol{\beta}\mathbf{X}_i)][\exp(\boldsymbol{\beta}\mathbf{X}_i)]^{n_i}}{n_i!} \tag{3}$$

To overcome the overdispersion problem, negative binomial regression can be applied by relaxing the assumption that the mean of accident frequencies equals the variance. To do this, an error term is added to the expected accident frequency ($\lambda_i$) such that Eq. (2) becomes

$$\lambda_i = \exp(\boldsymbol{\beta}\mathbf{X}_i + \varepsilon_i) \tag{4}$$

where $\exp(\varepsilon_i)$ is a gamma-distributed error term with mean one and variance $\alpha$. This gives a conditional probability.

$$P(n_i|\varepsilon) = \frac{\exp[-\lambda_i \exp(\varepsilon_i)][\lambda_i \exp(\varepsilon_i)]^{n_i}}{n_i!} \tag{5}$$

Integrating $\varepsilon$ out of this expression produces the unconditional distribution of $n_i$. The formulation of this distribution (the negative binomial) is

$$P(n_i) = \frac{\Gamma(\theta + n_i)}{[\Gamma(\theta) \cdot n_i!]} \cdot u_i^{\theta}(1 - u_i)^{n_i} \tag{6}$$

where $u_i = \theta/(\theta + \lambda i)$ and $\theta = 1/\alpha$, and $\Gamma(\cdot)$ is a value of gamma distribution. The corresponding likelihood function is

$$L(\lambda_i) = \prod_{i=1}^{N} \frac{\Gamma(\theta + n_i)}{\Gamma(\theta) n_i!} \left[\frac{\theta}{\theta + \lambda_i}\right]^{\theta} \left[\frac{\lambda_i}{\theta + \lambda_i}\right]^{n_i} \tag{7}$$

where $N$ is the total number of highway sections. This function is maximized to obtain coefficient estimates for $\beta$ and $\alpha$. This model structure allows the mean to differ from the variance such that,

$$\mathrm{var}[n_i] = E[n_i][1 + \alpha E[n_i]] \tag{8}$$

where $\alpha$ is the variance of the gamma-distributed error term and is used as a measure of dispersion. The choice between the negative binomial model and the Poisson model can be determined by the statistical significance of the estimated coefficient $\alpha$. If $\alpha$ is not significantly different from zero, the negative binomial model simply reduces to a Poisson model with $\mathrm{var}[n_i] = E[n_i]$. If $\alpha$ is significantly different from zero, the negative binomial model is the correct choice. A more detailed description of negative binomial regression analysis can be found in Washington et al. (2003).

The estimation results of the negative binomial model of freeway accident frequencies are presented in Table 2. The model has a reasonable overall statistical fit, as indicated by the $\rho^2$ statistic. Fifteen variables are found statistically significant or marginally significant in determining accident likelihood. It is noteworthy that the dispersion parameter, $\alpha$, is significantly different from zero. This confirms the appropriateness of the negative binomial model relative to the Poisson model. As shown in Table 2, a number of highway geometric variables that can significantly influence the accident occurrence were found. The positive sign of the number of lanes variable indicates that the increase in number of lanes will also

Table 2
Negative binomial estimation results

| Variable | Estimated coefficient | *t*-Statistic |
|---|---|---|
| Constant | −2.338 | −7.49 |
| Number of lanes | 0.367 | 5.21 |
| Descent grade | −0.064 | −1.46 |
| Level indicator (1 if $-1\% \leqslant$ grade $\leqslant 1\%$, 0 otherwise) | −0.161 | −1.64 |
| Severe upgrade indicator (1 if grade $\geqslant 3\%$, 0 otherwise) | 0.353 | 1.76 |
| Severe horizontal curve indicator (1 if degree of horizontal curve $\geqslant 6°$, 0 otherwise) | −0.538 | −2.81 |
| Military section indicator (1 if section is a military section, 0 otherwise) | 0.326 | 1.73 |
| Interchange indicator (1 if section contains an interchange, 0 otherwise) | 0.091 | 1.05 |
| ADT per lane (in 1000's of vehicles) | 0.028 | 2.98 |
| High PHF indicator (1 if PHF $\geqslant 0.95$, 0 otherwise) | −0.346 | −3.12 |
| High truck percentage indicator (1 if truck percentage $\geqslant 30\%$, 0 otherwise) | 0.268 | 2.19 |
| Fog zone indicator (1 if section is a fog zone, 0 otherwise) | −0.094 | −1.03 |
| Annual precipitation (mm) | −0.00016 | −2.26 |
| North proportion of freeway indicator (1 if mileage post is between north end and 94.7 km, 0 otherwise) | 0.218 | 1.75 |
| Section length (km) | 0.842 | 14.66 |
| 1998 indicator (1 if accident data were from 1998, 0 otherwise) | 0.080 | 1.06 |
| $\alpha$ (dispersion coefficient) | 0.220 | 3.32 |
| Number of observations | | 1500 |
| Restricted log-likelihood (constant only) | | −1775.74 |
| Log-likelihood at convergence | | −1545.31 |
| $\rho^2$ | | 0.12 |

increase accident likelihood. As expected, when the number of lanes increases, the total amount of lane changing as well as the conflicts between traffic will increase. This result is consistent with previous findings (Milton and Mannering, 1998; Carson and Mannering, 2001). Vertical and horizontal alignments are other important elements in highway geometric design. Grade can significantly influence vehicle operation speed, particularly for large trucks and buses. The effect of speed differentials can play an important role in accident occurrence. The estimated results indicate sections with severe uphill grade (3% or more) or descent grades have increased likelihood of accident occurrence, while level sections have reduced likelihood of accidents. It is important to note that the effect of downgrades on accident likelihood is positive because both the coefficient and the value of descent grade variable are negative. At downgrades, a greater frequency of accidents is expected because the speeds on downgrades are high and it is more difficult to control the vehicle under such conditions. For the effect of horizontal curves, the estimated result shows reduced accident likelihood for the sections with degree of horizontal curve greater than six degree. This result may seem counterintuitive, but it is consistent with past findings (Milton and Mannering, 1998). An explanation for this is that drivers are more likely to drive cautiously at sharp horizontal curves. In addition to vertical and horizontal alignments, lane width, median types and shoulder width were also identified by the past studies to have impact on accident occurrence (Carson and Mannering, 2001). Because most of the highway sections have the same lane width and the detailed information on shoulder width and median types was unavailable, this study could not examine their effects on accident frequencies.

In addition to the horizontal and vertical alignments, location-related indicator variables were also tested to investigate if accidents tend to occur at specific locations. Interchange indicator variable is intended to capture the impact of vehicles entering or leaving freeway mainline on accident risk. The estimated results show an increased likelihood of accidents at an intersection area. In addition, the estimated results show the increased likelihood for accidents in the military sections. The military sections, which provide emergency use for air force aircrafts, are a unique design feature of National Freeway 1. The design characteristics are being level, straight and with very wide shoulders, and the two traffic directions are not physically separated. An explanation for the increased likelihood of accident risk for the military sections is that drivers might easily but unintentionally speed under such design conditions. Because military sections are the only places where two directions of traffic are not physically separated, this finding implicitly indicates that physical separation between traffic for a freeway is crucial for preventing severe accidents. Another section location indicator variable shows that accidents are more likely to occur in the north portion of this freeway. These findings might imply differences of vehicle use and driver behavior associated with level of urbanization because the north proportion of the freeway runs through several metropolitan areas. In addition, this study also investigated if the mainline toll plazas have any effect on accident occurrence because mainline toll plazas can significantly interrupt freeway operations. A toll plaza indicator variable was examined, but the results show that the effect of mainline toll plazas on accident occurrence is statistically insignificant.

As for the traffic characteristic variables, AADT is typically used to indicate the traffic conditions. However, these long-term traffic data were not available. In this study, the interpolated daily traffic volumes for 1997 and 1998 based on traffic surveys conducted in 1994 and 1999 were used. The positive coefficient of the ADT variable and the high truck percentage indicator variable imply that conflicts between vehicles and the exposure to potential risk of accidents increase with increasing number of vehicles and trucks. In addition, for freeway sections with the peak hour factor higher than 0.95, the accident likelihood is reduced. Peak hour factor is used to measure the fluctuation of traffic flow. When the peak hour factor is greater than 0.95, the traffic condition is congested but relatively stable. A slow and stable traffic flow is less likely to increase the traffic accident likelihood.

In addition to geometric and traffic factors, environmental factors, such as snow and rain, were identified by past studies to have significant impacts on accident occurrence. Among these environmental factors, fog has been taken as a greater contributor to vehicle accidents because it can significantly reduce drivers' visibility. For example, there was a major accident caused by fog on this freeway in 1996 involving 99 vehicles, and resulting in three deaths and 23 injuries. The Taiwan Area National Freeway Bureau has identified specific freeway sections as fog zones and installed more traffic safety facilities such as illumination, flashing lights, and warning signs. To identify if there is higher accident frequency in fog zones, an indicator variable is selected for these zones. The negative coefficient indicates that accident frequency tends to decrease in fog zones. An explanation for this is that fog often occurs at early mornings and less number of vehicles is expected to be driving at such times compared to non-foggy conditions. In addition, the effect of precipitation was also investigated and the negative coefficient of the annual precipitation variable shows that an increase in annual precipitation will reduce accidents. A possible explanation for this is that drivers are more likely to drive at lower speeds and also keep longer car-following distances under the wet pavement and reduced visibility conditions.

However, detailed data on rain intensity and time of raining are unavailable. To have a better understanding of the effect of rain on accidents, a more detailed study is suggested. Overall, the environmental variables do not appear to be risk factors for vehicle accidents.

To gain a better understanding of the marginal effects of the variables on accident frequency, elasticities were computed. In general, the direct elasticity is defined as

$$E_{x_{ij}}^{\lambda_i} = \frac{\partial \lambda_i}{\partial x_{ij}} \cdot \frac{x_{ij}}{\lambda_i} \tag{9}$$

where $E$ represents the elasticity, $x_{ij}$ is the value of variable $j$ being considered for highway section $i$, and $\lambda_i$ is the mean of accident frequency on highway section $i$. Applying this to Eq. (4) gives

$$E_{x_{ij}}^{\lambda_i} = \beta_j x_{ij} \tag{10}$$

where $\beta_j$ is the coefficient corresponding to variable $j$.

The elasticity defined in Eq. (10) is used to measure the effect that 1% change in an accident covariate will have on accident frequency. Thus, it is only valid when the accident covariates are continuous variables, and it is not applicable for indicator variables (i.e., variables that take on values of zero or one). For indicator variables, a "pseudo-elasticity" can be used to give an approximate elasticity of the variable. The pseudo-elasticity gives the incremental change in accident frequency caused by the indicator variable in the model. In this case, the pseudo-elasticity is defined as

$$E_{x_{ij}}^{\lambda_i} = \frac{\exp(\beta_j) - 1}{\exp(\beta_j)} \tag{11}$$

The elasticities for the independent variables are shown in Table 3. None of them are elastic (i.e., absolute value of elasticity greater than one). The values in the table can be readily interpreted. For example, the elasticity for ADT is 0.58. This means that a 1% increase in ADT will result in a 0.58% increase in accidents. It is also important to note that elasticity estimates can only be applied to examine the effect of a small change of independent variable (e.g., 1–5% increase of ADT) on the expected accident frequency.

Table 3
Elasticity estimates of key variables

| Variables | Elasticity |
|---|---|
| Number of lanes | 0.81 |
| Descent grade | 0.03 |
| Level indicator | −0.17 |
| Severe upgrade indicator | 0.30 |
| Severe horizontal curve indicator | −0.71 |
| Interchange indicator | 0.09 |
| Military section indicator | 0.28 |
| ADT per lane (in 1000's of vehicles) | 0.58 |
| High PHF indicator | −0.41 |
| High percentage of truck indicator | 0.24 |
| Fog zone indicator | −0.10 |
| Annual precipitation (mm) | −0.35 |

## 5. Artificial neural network approach to freeway accident frequencies

The structure of the artificial neural network used for this study is a three-layer neural network, as shown in Fig. 1. The basic elements are the artificial neurons, and each neuron is interconnected with all the neurons in the next layer. The first layer is the input layer, where the data are presented to the neural network. The values of the input variables can be either a numerical value (generally normalized) or a binary code (e.g., gender). The intermediate layer is the hidden layer. The function of the hidden layer is to compute the complicated pattern associations. A single hidden layer has been found to be satisfactory in most applications, while the number of neurons in the hidden layer is generally determined through experimentation. The third layer is the output layer, representing the network response to the corresponding input (i.e., accident frequencies). The neural network can then be trained through a training algorithm. Currently, there are a number of training algorithms available for artificial neural network models, and the back-propagation rule, which is one of the most widely used training algorithms, is adopted in this study. The principle of this rule is to minimize the total output error described in Eq. (12).

$$\mathrm{MSE} = \frac{1}{N \times K} \sum_{i=1}^{N} \sum_{j=1}^{k} (t_{ij} - a_{ij})^2 \tag{12}$$

where MSE is the mean squares error, $t$ is the target output value, $a$ is the model output value, $K$ is the number of output neurons and $N$ is the number of testing data. More detailed description on the back-propagation rule can be found in Hagan et al. (1996).

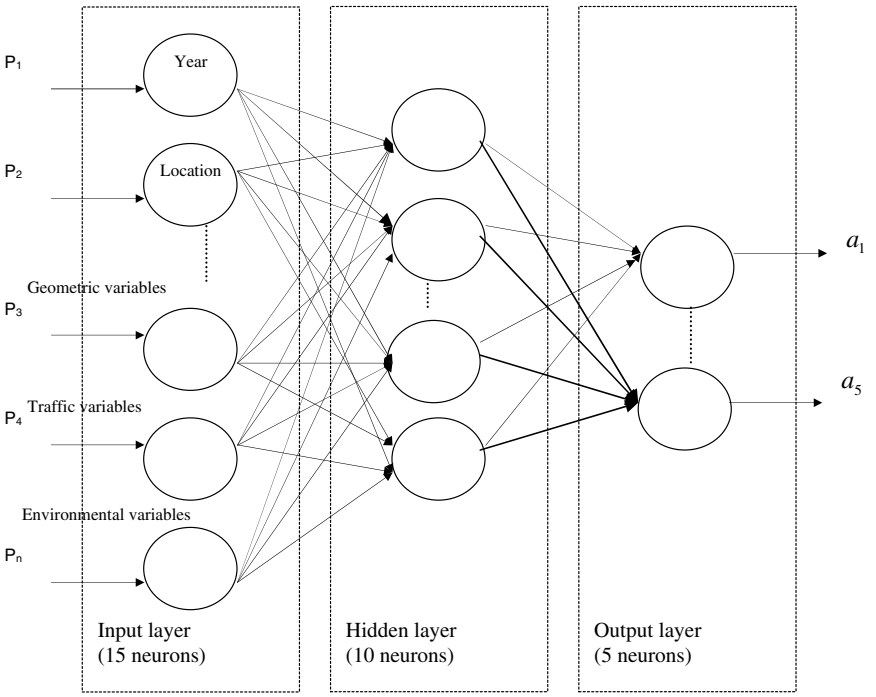


Fig. 1. The structure of the ANN model.

To estimate the ANN model, there are a number of software packages ready to perform the back-propagation algorithm, and MATLAB was chosen for this study. Here, the input layer contains 15 neurons which are the input variables representing the potential risk factors for accidents. Currently, there is no systematic way to select the number of neurons in the input layer. Thus 15 neurons, the statistically significant variables found in the negative binomial regression model, are used. The other reason for using these 15 input variables is to allow the comparisons of the model prediction performance and the marginal effects of each independent variable between the negative binomial regression and ANN models. Table 4 shows the definition of input variables. Because the effectiveness of the back propagation training algorithm depends on the number of neurons in the hidden layer, various numbers of neurons (ranging from 1 to 29) in the hidden layer were tested. The "optimal" number of neurons in the hidden layer was found to be 10. In this study, five neurons in the output layer were taken to represent the accident frequencies for the freeway sections because relatively few highway sections had more than three accidents. Treating freeway sections with accident frequencies greater than three as a group can reduce the complexity of the network. By minimizing the MSE calculated on the testing data, the best solution was found after 800 learning cycles. The change of MSE in relation to learning cycles is shown in Fig. 2. The MSE of the best solution for the training and for the testing data are 0.097 and 0.112, respectively.

To evaluate the marginal effects of input neurons (variables) on output neurons for an ANN model, sensitivity analysis is commonly applied (Mussone et al., 1999; Tong and Hung, 2002). The sensitivity analysis conducted by this study is to examine the effect of a particular independent variable change on accident frequency distribution by holding all other variables fixed. Table 5 shows the analysis results. For example, of the level highway sections, about 95.4% had no accident occurrence, 2.8% had one accident, 1.2% had two accidents, 0% had three accidents, and 0.6% had four or more accidents. When the highway section is not level (i.e., the grade is greater than 1% or less than −1%), the distribution of accident frequencies of highway sections are 95.2%, 2.4%, 0.6%, 1.8%, and 0% for zero, one, two, three and four or more accidents, respectively. Thus the average

Table 4
Definition of the input variables

| Variable | Definition | Binary/numerical code |
|---|---|---|
| X1 | Number of lanes | Numerical value |
| X2 | Vertical alignment 1 | Numerical value |
| X3 | Vertical alignment 2 | 1 for section with grade between −1% and 1%, 0 for others |
| X4 | Vertical alignment 3 | 1 for section with grade $\geqslant$ 3%, 0 for others |
| X5 | Horizontal alignment 1 | 1 for section with degree of horizontal curve $\geqslant$ 6°, 0 for others |
| X6 | Military section | 1 for section within a military section; 0 for others |
| X7 | Interchange | 1 for section containing an interchange; 0 for others |
| X8 | ADT per lane | Numerical value |
| X9 | High truck percentage | 1 for section with truck percentage $\geqslant$ 30%, 0 for others |
| X10 | High PHF | 1 for PHF $\geqslant$ 0.95, 0 for others |
| X11 | Fog zone | 1 for section located in fog zone, 0 for others |
| X12 | Annual precipitation | Numerical value |
| X13 | Section location | 1 for section located in northern Taiwan, 0 for others |
| X14 | Section length | Numerical value |
| X15 | Year of 1998 | 1 for accident data from 1998, 0 for others |

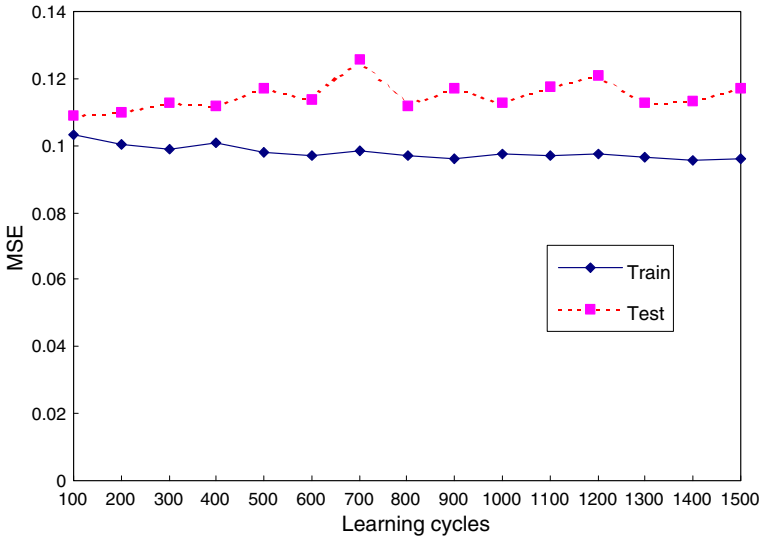


Fig. 2. The training curve for the ANN model.

accident frequency for level highway sections is about 19% lower than that in the highway sections with grades. The shift of accident frequency distribution indicates that highway sections with grades have greater likelihood of having an accident. This finding is consistent with the elasticity analysis from the negative binomial regression model. However, it is important to note that not all of the sensitivity analysis results are consistent with those from the negative binomial regression model. For example, the average accident frequency is 0.22 for the highway sections with an interchange and 0.06 for those without an interchange in the sensitivity analysis. The presence of an interchange in a highway section increases the accident likelihood by 270%, while the elasticity analysis shows only a 9% increase in the accident likelihood. It should also be noted that for the continuous independent variables such as ADT per lane, only a certain range of changes was analyzed. Although the results of sensitivity analysis and elasticity are not quite similar, the sensitivity analysis still provides valuable insight into the relationship between risk factors and accident frequency.

## 6. Comparisons of the prediction performance of neural networks and negative binomial regression models

In order to examine the performance of the ANN model, the comparison of model prediction performance between the negative binomial regression model and ANN model is examined. Tables 6 and 7 show the comparison results. For the negative binomial regression model, the overall model prediction accuracy for the training data is about 58.3%, while that for the testing data is about 60.8%. For the ANN model, the overall model prediction performances for the training data and the testing data are 64% and 61.4%, respectively. The proposed ANN model performs slightly better than the negative binomial regression model in analyzing the training data. In predicting the accident frequency on

Table 5
Sensitivity analysis of ANN model

| | Variable | $N = 0$ (%) | $N = 1$ (%) | $N = 2$ (%) | $N = 3$ (%) | $N \geqslant 4$ (%) |
|---|---|---|---|---|---|---|
| 1 | Number of lanes = 2 | 97.30 | 2.70 | 0.00 | 0.00 | 0.00 |
| | Number of lanes = 3 | 92.31 | 3.55 | 1.78 | 1.18 | 1.18 |
| | Number of lanes = 4 | 96.52 | 1.99 | 1.00 | 0.50 | 0.00 |
| 2 | Level indicator = 0 | 95.15 | 2.42 | 0.61 | 1.82 | 0.00 |
| | Level indicator = 1 | 95.41 | 2.75 | 1.22 | 0.00 | 0.61 |
| 3 | Severe upgrade indicator = 0 | 95.14 | 2.75 | 1.06 | 0.63 | 0.42 |
| | Severe upgrade indicator = 1 | 100.00 | 0.00 | 0.00 | 0.00 | 0.00 |
| 4 | Severe horizontal curve indicator = 0 | 95.31 | 2.77 | 0.85 | 0.64 | 0.43 |
| | Severe horizontal curve indicator = 1 | 95.65 | 0.00 | 4.35 | 0.00 | 0.00 |
| 5 | Interchange indicator = 0 | 96.81 | 1.72 | 1.23 | 0.00 | 0.25 |
| | Interchange indicator = 1 | 88.10 | 7.14 | 0.00 | 3.57 | 1.19 |
| 6 | Military section indicator = 0 | 95.67 | 2.47 | 1.03 | 0.62 | 0.21 |
| | Military section indicator = 1 | 71.43 | 14.29 | 0.00 | 0.00 | 14.29 |
| 7 | ADT per lane < 15,000 | 97.30 | 2.70 | 0.00 | 0.00 | 0.00 |
| | ADT per lane < 20,000 | 92.31 | 3.55 | 1.78 | 1.18 | 1.18 |
| | ADT per lane < 25,000 | 96.52 | 1.99 | 1.00 | 0.50 | 0.00 |
| | ADT per lane $\geqslant$ 25,000 | 97.65 | 2.35 | 0.00 | 0.00 | 0.00 |
| 8 | High PHF indicator = 0 | 94.72 | 2.88 | 1.20 | 0.72 | 0.48 |
| | High PHF indicator = 1 | 98.67 | 1.33 | 0.00 | 0.00 | 0.00 |
| 9 | High percentage of truck indicator = 0 | 89.09 | 7.27 | 1.82 | 1.82 | 0.00 |
| | High percentage of truck indicator = 1 | 95.77 | 2.82 | 0.00 | 0.00 | 1.41 |
| 10 | Fog zone indicator = 0 | 94.63 | 3.32 | 0.77 | 0.77 | 0.51 |
| | Fog zone indicator = 1 | 98.02 | 0.00 | 1.98 | 0.00 | 0.00 |
| 11 | Annual precipitation < 2000 | 97.65 | 1.57 | 0.78 | 0.00 | 0.00 |
| | Annual precipitation < 3000 | 92.23 | 3.63 | 1.55 | 1.55 | 1.04 |
| | Annual precipitation $\geqslant$ 3000 | 95.45 | 4.55 | 0.00 | 0.00 | 0.00 |

Table 6
Prediction performance for the negative binomial regression model

| | Training data | | | Testing data | | |
|---|---|---|---|---|---|---|
| | Observed frequency | Predicted frequency | Correctly predicted | Observed frequency | Predicted frequency | Correctly predicted |
| $N = 0$ | 862 | 1354 | 833 (97%) | 295 | 452 | 286 (97%) |
| $N = 1$ | 389 | 105 | 32 (8%) | 127 | 26 | 10 (8%) |
| $N = 2$ | 154 | 0 | 0 (0%) | 46 | 0 | 0 (0%) |
| $N = 3$ | 58 | 0 | 0 (0%) | 18 | 0 | 0 (0%) |
| $N \geqslant 4$ | 37 | 41 | 9 (24%) | 6 | 14 | 3 (50%) |

The overall prediction accuracy is 58.3% for training data and 60.8% for testing data.

a highway section, the proposed ANN model performs better for the highway sections with one or more accidents, while the negative binomial regression model performs slightly better for the sections with zero accidents.

Table 7
Prediction performance for the ANN model

| | Training data | | | Testing data | | |
|---|---|---|---|---|---|---|
| | Observed frequency | Predicted frequency | Correctly predicted | Observed frequency | Predicted frequency | Correctly predicted |
| $N = 0$ | 862 | 1142 | 780 (90%) | 295 | 469 | 290 (98%) |
| $N = 1$ | 389 | 312 | 149 (38%) | 127 | 13 | 6 (5%) |
| $N = 2$ | 154 | 21 | 13 (8%) | 46 | 5 | 1 (2%) |
| $N = 3$ | 58 | 4 | 4 (7%) | 18 | 3 | 3 (17%) |
| $N \geqslant 4$ | 37 | 21 | 14 (38%) | 6 | 2 | 2 (33%) |

The overall prediction accuracy is 64% for training data and 61.4% for testing data.

## 7. Discussion

In this particular application, the negative binomial regression model and ANN model provide similar results in terms of prediction performance on the training data and testing data. This demonstrates that the ANN model is an appropriate methodology for analyzing traffic accidents. Although it is difficult to distinguish which modeling approach is better according to the analysis results of this study, there are some aspects might be of great interest for future research.

In past research, the Poisson or negative binomial model was commonly employed for traffic accident analysis because of the nature of random, discrete, and non-negative characteristics of vehicle accidents and their capacity of identifying effectively a broad range of risk factors for accidents. In addition, the elasticity of each risk factor can be mathematically defined and ready to compute. These analysis results not only can be easily interpreted, but also provide clear and valuable information for traffic engineers to perform mitigation. In constrast, the advantage of the ANN model is that it requires no assumptions of underlying relationship between risk factors and traffic accidents. In this application, if the underlying relationship between risk factors and traffic accidents does not follow a gamma distribution, the relationship estimated by the negative binomial regression could be erroneous. Another advantage of the ANN model is that it can effectively handle interrelation problems between independent variables. When a serious correlation exists between independent variables, the variability of estimated coefficients will be inflated and interpretation of relationship between independent variables and dependent variable will also be difficult. But when the ANN model is applied, the correlation problems between independent variables would not be a great concern. Compared to the commonly applied regression models in traffic accident analysis, this is an important advantage of employing ANN models because an accident is rarely due to a single risk factor, but is rather the outcome of a series of factors.

Despite these advantages, the ANN model has its own drawbacks. Firstly, developing an ANN model is very time-consuming. The time required to develop an ANN model depends on the size of training data and network structure. As discussed earlier, there is no general rule in determining the network structure and it can only be done by experimentation. Therefore, it always takes a great deal of time to determine the model structure, including the network structure (number of hidden layers and number of neurons in the hidden layer), transfer functions, and so on. Once a network is specified, it usually takes hours to complete an experiment especially when the size of training data is large because a

training algorithm usually needs to go through several hundred of iterations to obtain an "optimal" weighting for the network. Secondly, unlike the elasticity analysis for most statistical models, the sensitivity analysis for the ANN model cannot be mathematically defined. Therefore, the sensitivity analysis for continuous variables is difficult to perform. For example, to analyze the relationship between accidents and precipitation, the ANN model can only examine the distribution change of accident frequency against a certain amount of annual precipitation. In addition, the selection of input neurons is also a critical issue in developing an ANN model. Because the input neurons are usually regarded as the input variables, only the variables with cause–effect relationship should be selected. Having more input neurons in the input layer can significantly increase the computing time, but does not guarantee a better prediction performance of the network. As stated above, various combinations of input variables were tested during model development. The performance of the proposed model was better than that of all tested models. The selection of input neurons can significantly influence the model performance and so selection should be done with caution.

## 8. Conclusions

A negative binomial regression and an ANN model were proposed to establish the empirical relationship between traffic accidents and highway geometric variables, traffic characteristics and environmental factors. The results of this study can eventually be employed to identify locations of high accident frequency for the most important transportation corridor in Taiwan. This study also demonstrated that ANN is a consistent alternative for analyzing freeway accident frequency by comparing the prediction performance with negative binomial regression analysis. This represents an important methodological step in studying traffic accident frequency. The results obtained here, by exploring a broad range of variables including highway geometry, traffic and environmental characteristics, provide valuable insight into the underlying relationship between risk factors and vehicle accidents. In terms of future work, an application of the methodological approaches used in this paper to different roadway types, such as interchange ramps, would be worthwhile. The accident database revealed that quite a number of accidents occurred at interchange ramps. Further exploration could provide a better understanding of the characteristics for accidents occurred at interchange ramps and safer designs for the freeway systems. In addition, the overall prediction performance of the proposed ANN model is approximately 60%. Future work might focus on how to improve the prediction performance of ANN models. First of all, because the structure of ANN can significantly influence the prediction performance, it would be worthwhile for future studies to develop a new ANN model for predicting accident rate (i.e., the neurons in the output layer will be reduced to one) instead of accident frequency and to check if the resulting model can give a better prediction. Secondly, training an ANN model using a different ratio of testing to training data or the same ratio of testing to training data for a different number of observations could result in different results. Further investigation on the effects of the ratio of testing to training data on the model performance would also be a good direction. Finally, it would also be interesting for future studies to employ different training algorithms such as the PNN to explore the factors that affect accident frequency and to see if the prediction performance could be improved.

## Acknowledgements

The author would like to thank the Editor and referee for their constructive comments and the National Science Council of Taiwan for financially supporting this study under Contract No. NSC91–2211–E–415–002.

## References

Abdel-Aty, M., Pande, A., 2005. Identifying crash propensity using traffic speed conditions. Journal of Safety Research 36 (1), 97–108.

Carson, J., Mannering, F., 2001. The effect of ice warning signs on ice-accident frequency and severity. Accident Analysis and Prevention 33 (1), 99–109.

Dougherty, M., 1995. A review of neural networks applied to transport. Transportation Research Part C 3 (4), 247–260.

Hadi, M., Aruldhas, J., Chow, L.-F., Wattleworth, J., 1995. Estimating safety effects of cross-section design for various highway types using negative binomial regression. Transportation Research Record 1500, 169–177.

Hagan, M.T., Demuth, H.B., Beale, M., 1996. Neural Network Design. Thomson Learning Publishing Inc.

Hensher, D.A., Ton, T.T., 2000. A comparison of the predictive potential of artificial neural networks and nested logit models for commuter mode choice. Transportation Research Part E 36 (3), 155–172.

Jin, X., Cheu, R.L., Dipti, S., 2002. Development and adaptation of constructive probabilistic neural network in freeway incident detection. Transportation Research Part C 10 (2), 121–147.

Lee, J., Mannering, F., 2002. Impact of roadside features on the frequency and severity of run-off-roadway accidents: an empirical analysis. Accident Analysis and Prevention 34 (2), 149–161.

Lee, A.H., Stevenson, M.R., Wang, K., Yau, K., 2002. Modeling young driver motor vehicle crashes: data with extra zero. Accident Analysis and Prevention 34 (4), 515–521.

McCarthy, P.S., 1999. Public policy and highway safety: a city-wide perspective. Regional Science and Urban Economics 29 (3), 231–244.

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

Milton, J., Mannering, F., 1998. The relationship among highway geometrics, traffic-related elements and motor-vehicle accident frequencies. Transportation 25, 395–413.

Mussone, L., Rinelli, S., Reitani, G., 1996. Estimating the accident probability of a vehicular flow by means of an artificial neural network. Environment and Planning B 23, 667–675.

Mussone, L., Ferrari, A., Oneta, M., 1999. An analysis of urban collisions using an artificial intelligence model. Accident Analysis and Prevention 31 (6), 705–718.

Poch, M., Mannering, F., 1996. Negative binomial analysis of intersection—accident frequencies. Journal of Transportation Engineering 122 (2), 105–113.

Shankar, V.N., Mannering, F., Barfield, W., 1995. Effect of roadway geometrics and environmental factors on rural freeway accident frequencies. Accident Analysis and Prevention 27 (3), 371–389.

Shankar, V.N., Milton, J., Mannering, F., 1997. Modeling accident frequencies as zero-altered probability processes: an empirical inquiry. Accident Analysis and Prevention 29 (6), 829–837.

Shiva Nagendra, S.M., Khare, M., 2004. Artificial neural network based line source models for vehicular exhaust emission predictions of an urban roadway. Transportation Research Part D 9 (3), 199–208.

Sohn, S., Lee, S., 2003. Data fusion, ensemble and clustering to improve the classification accuracy for the severity of road traffic accident in Korea. Safety Science 41 (1), 1–14.

Subba Rao, P.V., Sikdar, P.K., Krishna Rao, K.V., 1998. Another insight into artificial neural networks through behavioral analysis of access mode choice. Computers, Environment and Urban Systems 22 (5), 485–496.

Tong, H.Y., Hung, W.T., 2002. Neural network modeling of vehicle discharge headway at signalized intersection: model descriptions and the results. Transportation Research Part A 36 (1), 17–40.

Vythoulkas, P.C., Koutsopoulos, H.N., 2003. Modeling discrete choice behavior using concepts from fuzzy set theory, approximate reasoning and neural networks. Transportation Research Part C 11 (1), 51–73.

Washington, S., Karlaftis, M.G., Mannering, F.L., 2003. Statistical and Econometric Methods for Transportation Data Analysis. Chapman & Hall/CRC Press.

Yin, H., Wong, S.C., Xu, J., Wong, C.K., 2002. Urban traffic flow prediction using fuzzy-neural network. Transportation Research Part C 10 (2), 85–98.

Yuan, F., Cheu, R.L., 2003. Incident detection using support vector machines. Transportation Research Part C 11 (3-4), 309–328.

Zhang, H.M., Ritchie, S.G., Jayakrishnan, R., 2001. Coordinated traffic-responsive ramp control via nonlinear state feedback. Transportation Research Part C 9 (5), 337–352.

Zhong, M., Lingras, P., Sharma, S., 2004. Estimation of missing traffic counts using factor, genetic, neural and regression techniques. Transportation Research Part C 12 (2), 139–166.