Modelling Motorway Accidents using Negative Binomial Regression

Pan CHENGYE [a], Prakash RANJITKAR [b]

[a,b] *Department of Civil & Environmental Engineering, University of Auckland, Auckland 1142, NEW ZEALAND*
[a]*Email:leaf8473@gmail.com*
[b]*Email: p.ranjitkar@auckland.ac.nz*

Abstract: This paper investigates motorway safety by developing accident prediction models that link accident frequencies to their non-behavioural contributing factors, including traffic conditions, geometric and operational characteristics of road, and weather conditions. The study used a sample of accidents occurred from 2004 through 2010 on a 74 km long section of Auckland motorway. A number of accident prediction models were developed and assessed for their predictive ability using negative binomial regression models under three categories: first for the whole of the motorway, second for rural and urban motorway segments separately and third for motorway segments without ramp, with on-ramp and with off-ramp separately. The results uncovered the safety impacts of different non-behavioural contributing factors, in which segment length, AADT per lane and the number of lanes always have the most profound effects on accident frequency. The validation tools were applied to examine the ability of models to predict accidents.

Keywords: Accident prediction model, Auckland motorway, negative binomial model, model validation

1. INTRODUCTION

Traffic crashes cannot be absolutely eradicated, however, it is possible to prevent them to some extent as long as the contributing factors are identified and tackled appropriately. In general, factors that contribute to road crashes may include inappropriate driver behaviour, congested traffic, unanticipated roadway geometrical change and adverse weather condition. The driver behaviour plays a crucial role in the occurrence of a crash; however it is usually complex and unpredictable. Hence, it is important to figure out the role of the non-behavioural factors in traffic accidents, based on which cost effective countermeasures can be recommended to reduce the chance of accidents.

Motorway is considered as a relatively safe modern traffic facility, for its high level of design and construction standards embodied as controlled access, grade-separated intersections and median-divided carriageways. However, traffic safety on motorway continues to be of great concern for the general public. Motorway accounts for only a small proportion (0.21%) of a total 94,000 kilometres of road network in New Zealand, while it carries nearly 10% of the country's traffic. New Zealand Transport Agency (NZTA) (previously Transit NZ) have been monitoring crashes and undertaking crash reduction studies of motorway network in New Zealand since at least the early 1990s. A statistics compiled by Ministry of Transport shows that among 22 worst traffic black spots in the country, 18 are located on motorways. In the face of increasing traffic flows, the number of crashes per driver has been reducing in response to the improvements; however more work is needed to address safety and congestion problems in the motorway network. There are a number of safety issues along motorways that are not well understood e.g. the effect of closely spaced ramps on lane changing and traffic safety, the effect of congestion and resulting in queues forming on traffic safety. A greater understanding of these issues should result in more innovative safety interventions being developed，for example the use of ITS interventions (such as variable speed signs) to improve the operation of motorways.

The main objective of this paper is to develop accident prediction models for Auckland motorway. The main focus is on the motorway safety, specifically the association between motorway accident frequency and non-behavioural contributing factors including traffic conditions, geometric and operational characteristics of road, and weather conditions. The paper begins with a literature review of previous research on road accidents in terms of their relationship with potential contributing factors. Description of data collected from four different sources is covered in data collection, which is followed by data processing section that covers information on how data were disaggregated into 959 samples used for model development that follows the next. Then accident prediction models are presented with interpretation of the effects of significant variables and model validation to compare the predictive performance of different models. Finally in the last section, some concluding remarks are drawn based on the results presented in the previous sections.

2. LITERATURE REVIEW

A number of accident prediction models have been developed in the last two decades to estimate the expected accident frequencies on roads as well as to identify various factors associated with the occurrence of accidents. It is not possible for regression models to account for each and every factor that affects accident occurrence (Persaud and Dzbik, 1993). Most previous researchers have focused on non-behavioural factors such as traffic flow characteristics, road geometry and environmental conditions. As the first accident modellers who worked on multilane road, Persaud and Dzbik (1993) investigated the relationship between freeway crash data and traffic volumes. Using the hourly traffic, the model indicated that higher accident risk is associated with congestion and afternoon rush hour. Shankar et al. (1995) modelled the monthly accident frequency of a rural motorway as a function of geometrics, weather conditions and their interaction, and found it dangerous for the areas with large rainfall and snowfall to have steep grades and tight horizontal curves. They also examined the effects of geometrics and weather on the likelihood of classified accident types, e.g. rear-end, sideswipes, overturn, hitting parked-vehicle or fixed-object. The results showed that inclement weather conditions increase the tendency of most types of accidents. Hadi et al. (1995) focused on the safety effects of cross-section elements. For road median, greater width could reduce accident rate. A lane width of 4.0 m was found to be associated with low crash rate on urban freeways, so was the outside shoulder width of 3.0-3.7 m for urban freeways and inside shoulder width ranged between 1.2 and 1.8 m for rural freeways. Research conducted by Abdel-Aty and Radwan (2000) revealed that AADT per lane has the most critical impact on accident occurrence. Narrow lane width and large number of lanes, narrow shoulder width, reduced median width and sharper horizontal curves are potential geometric characteristics increasing accident frequency. Other researchers have specifically contributed to investigate the road safety factors involving horizontal curvature (Anderson et al., 1999), traffic characteristics (Abdel-Aty and Abdalla, 2004; Lord et al., 2005), design consistency (Montella, 2008), and road pavement quality (Anastasopoulos et al., 2012).

Regression models have been most commonly used to relate accident frequency with explanatory variables. The result of model strongly relies on the choice of regression technique. The earlier traffic accident studies used ordinary or normal linear regression models, which follow the assumption of a normal distribution for the dependent variable, a constant variance for the residuals, and the linear relationship existing between dependent and independent variables. However the conventional linear regression method should be used with caution because of the problems associated with non-negative and error terms (Jovanis and Chang, 1986; Abdel-Aty and Radwan, 2000). Jovanis and Chang (1986) recommended generalized linear model using Poisson distribution error structure as a mean to describe the random, discrete and non-negative accidents. Poisson regression assumes an exponential relationship between response and explanatory variables. Eenink et al. (2007) advised a basic form of accident prediction model:

$$E(\lambda) = \alpha Q^{\beta} e^{\sum r_i x_i} \tag{1}$$

where the expected number of accidents E(λ) is a function of traffic volume, Q and a set of risk factors, x_i.

Abdel-Aty and Radwan (2000) attempted to use the Poisson regression methodology and then rejected it because of different mean and variance value of the dependent variable, which indicated over dispersion in the accident data. Consequently, Poisson-gamma (or called negative binomial) model was adopted as a superior alternative to accommodate the over dispersion. The negative binomial regression model has been widely employed in vehicle accident analysis for rural highways, arterial roadways, urban motorways, and rural motorways (Shankar et al. 1995; Lord, 2005; Montella, 2008; Pemmanaboina, 2005). When considerable zeros and extremely low mean value are observed in accident numbers, negative binomial model is significantly unreliable to fit the data, and the dispersion parameter can be mis-estimated. To overcome the difficulties arising with zero accident samples, some researchers used extended Poisson and negative binomial models which can account for excessive zeros, for example, zero-inflated Poisson and zero-inflated negative binomial (Miaou, 1994; Shankar et al., 1997; Lord et al., 2005). Zero-inflated models deal with the dual-state system: the zero-accident state in which no accidents will ever be observed, and the non-zero accident state where accident frequencies follow some known distributions such as the Poisson or negative binomial distribution.

More recently, a number of innovative approaches have been proposed in the domain of accident prediction models seeking improvement from the traditional methods discussed earlier. For example, EI-Basyouny and Sayed (2006) proposed a modified negative binomial regression technique which improved goodness of fit. Caliendo et al. (2007) applied the negative multinomial distribution to model multiple observations in the same road section at different years that may not be mutually independent. The generalized estimating equation (GEE) was also used to fit data of several years (Abdel-Aty and Abdalla, 2004; Lord and Persaud, 2000).

In New Zealand, the generalized linear models have been used in accident research of different intersection types as well as for two-lane roads. The earlier work usually focused on the effect of traffic volume. The non-flow variables have not begun to be taken into account till recent years. So far the efforts to model accident likelihood on motorway have been minimal in New Zealand, except for flow-only models presented in Turner (2001) and Economic Evaluation Manual (2010).

3. DATA COLLECTION

A 74 km long motorway section along State Highway 1N namely Northern Motorway and Southern Motorway linking Auckland City with adjacent regions was selected to develop accident prediction models. Although named as two motorways, they are treated as a whole since they are physically connected. This motorway section is appropriate for this study for several reasons. First of all, it is one of the longest motorway routes in New Zealand which is long enough to produce a sufficient number of segments for modelling. Secondly, it runs through several districts, giving plenty of variance in, for example traffic volume, road design, land use, and terrain characteristics along the project, producing road samples with different features available. Thirdly, it experiences high frequency of accidents. More than 1,300 reported accidents occurred per year in the last decade, containing around 300 injury and fatal accidents. Lastly, limited infrastructure changes were observed during the study period. Efforts were taken to gather as much data as possible in order to know more about accidents and the information on factors contributing to accident occurrence.

Accident data was sourced from the Crash Analysis System (CAS), which is maintained by New Zealand Transport Agency. As a crash management, analysis and mapping tool, CAS is made to be a very exhaustive data resource, containing crash location, time, collision type, severity, etc. 10,149 crashes occurred on the selected motorway section over the seven-year period between 1st January 2004 and 31st December 2010 were retrieved.

Traffic related data was extracted from Traffic Monitoring System (TMS), which collects traffic information via various vehicle sensors and detectors installed along the State Highway. Average Annual Daily Traffic (AADT) data ranged between 18,625 and 102,420 vehicles per day on the motorway mainline. The proportion of heavy vehicles was in the range of 1.47% to 10.79%. Only accidents occurred on mainline of the motorway were considered in this study, which exclude accidents occurred on ramps. The flow through ramps was included in the analysis as a factor, which may influence traffic on mainline as it causes traffic conflicts. AADT on ramps ranged between 430 and 24,217 vehicles per day.

Road related data was sourced from Road Assessment and Maintenance Management

(RAMM) system maintained by Auckland Motorway Alliance. It records geometric data containing position of ramps, cross-section design features, and horizontal and vertical alignments. The legal speed limits along motorways sections, which were either 100 km/h or 80 km/h was sourced from Highway Information Sheets (HIS) provided by Auckland Motorway Alliance.

The diversity of weather conditions along the study motorway is not significant, but the effect of weather condition on accident occurrence is still worth examining. The meteorological data is available online from National Climate database of National Institute of Water and Atmospheric Research. The rain precipitation data from seven weather stations nearest to the motorway route was retrieved for the study area, including the annual rainfall and annual wet days which refers to numbers of days with 1mm rain or more in a year. The data was assigned to proximate segments. As a result a number of contiguous segments may share the rainfall data from the same station. The annual rainfall ranged from 802.8 to 1,445.6 mm, and annual wet days were between 103 and 165 days during 2004 to 2010.

4. DATA PROCESSING

The northbound and southbound carriageways were treated separately. A few road segments were filtered out due to their special features e.g. Auckland Harbour Bridge where the number of lanes changes during peak hours to cope with the traffic demand towards and away from Auckland CBD. Some other segments excluded from the analysis had considerable construction works carried out during the study period. Finally, the remaining 67 km of motorway were disaggregated into 137 segments, including 66 northbound segments and 71 southbound segments. These segments were homogeneous in terms of traffic features and key roadway design characteristics including number of lanes, lane width and horizontal alignments. A less common alternative segment dividing method used in previous accident modelling research is fixed-length segment which simply divides the roadway into small segments with a suitable equal length. Comparatively speaking, using homogeneous road segments greatly increase the difficulty of data collection and data processing, but is more beneficial to accounting the effects of road-related factors such as traffic volumes and geometric characteristics on accident frequency. In this study the presence of an on-ramp or off-ramp was the primary factor considered in the segmentation process. Kopelias et al. (2007) found that no significant crash number increase was observed when the distance to on/off-ramps was extended to 500m. Based on this finding, the motorway was divided into three types of segments: segment with off-ramp, segment with on-ramp, and segment without ramp. The lengths of former two are illustrated in Figure 1.

The segments without ramp varied a lot in length, which were further subdivided based on changes in other road elements such as the number of lanes and horizontal curvature. Segments shorter than 0.2 km or longer than 3 km were avoided to mitigate the heteroskedasticity problem. Short segments can potentially lead to excess zeros in the crash data. One result coming out in this stage was that a few segments contained two ramps (one on-ramp and one off-ramp) when the distance between the two interchanges were fairly short.

Each year data from a segment was treated as a sample giving a total of 959 data samples. Table 1provides as a statistical summary of continuous and discrete data used in this study.

Table 1. Statistics of data

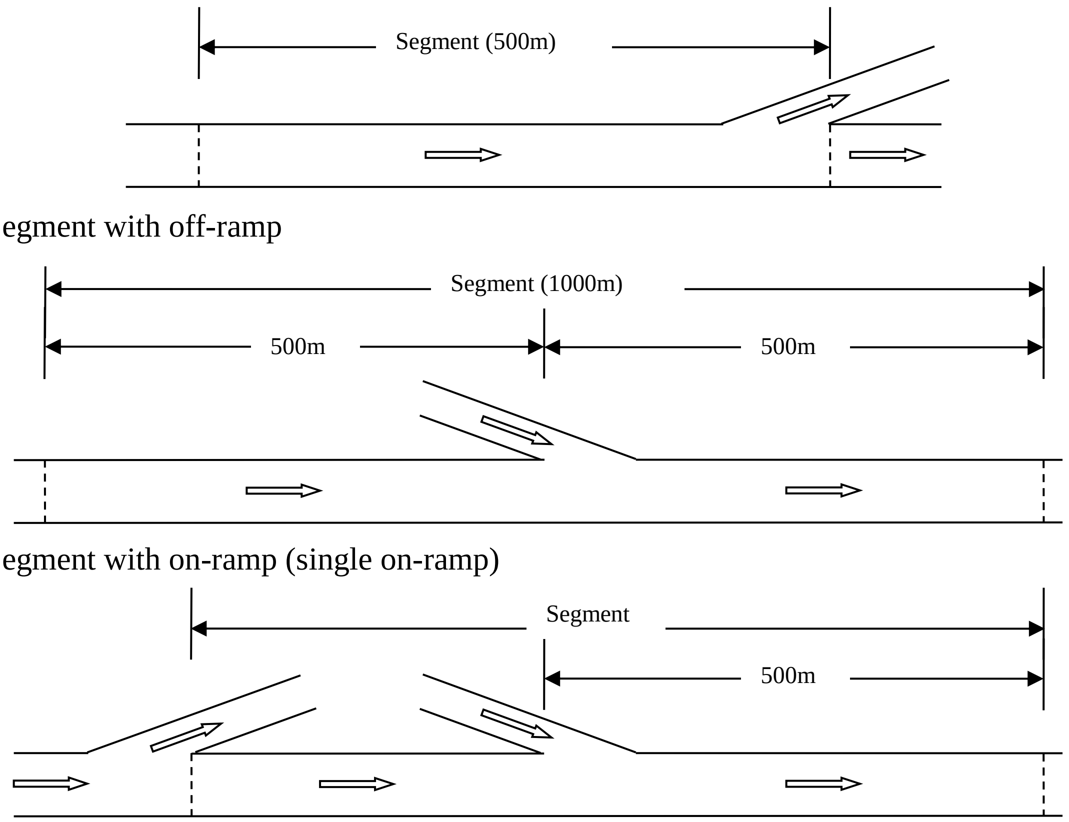
a) Statistics of continuous data

	Minimum	Maximum	Mean	Standard deviation
Accident frequency (per year)	0	69	8.77	9.85
Segment length (km)	0.27	2.94	0.97	0.53
AADT per lane (in 1000's of vehicles)	6.25	28.15	16.25	5.34
Percentage of heavy traffic (%)	1.47	11.83	6.11	2.21
On-ramp AADT	0.00	24.22	3.46	5.82
Off-ramp AADT	0.00	22.21	3.33	5.61
Average radius of horizontal curves(m)	440	20000	4155.11	3373.47
Minimum radius of horizontal curves (m)	189	5000	734.85	894.94
Average vertical up-grade (%)	0	4.95	1.22	1.12
Average vertical down-grade (%)	0	5.10	1.23	1.15
Maximum vertical up-grade (%)	0	6.40	1.98	1.60
Maximum vertical down-grade (%)	0	8.00	2.13	1.84
Number of lanes	2	5	2.76	0.81
Shoulder width (m)	0.10	7.20	3.03	1.12
Median width (m)	2.00	20.00	5.56	3.14
Annual rainfall (m)	0.80	1.45	1.13	0.13
Annual wet days	103	165	130.78	13.06

*Segment number N is 959 except for AADT and Percentage of heavy traffic with some missing figures.

b) Statistics of discrete data

	Category	Segment percentage
Lane width (m)	3.5	30%
	3.6	70%
Presence of ramp	With on-ramp	33%
	With off-ramp	34%
	Without ramp	37%
Median type	Steel	49%
	Concrete	51%
Speed limit (km/h)	80	11%
	100	89%



(a) Segment with off-ramp

(b) Segment with on-ramp (single on-ramp)

(c) Segment with on-ramp (on-ramp belonging to an interchange)

Figure 1. Typical layout of segments with on/off ramp

5. MODEL DEVELOPMENT

In an effort to come up with the most appropriate approach that relates accident frequencies with their explanatory variables, a number of regression models were tested in SPSS and STATA software including the Poisson model, negative binomial model, zero-inflated negative binomial model and GEE model. Zero-inflated Poisson model is inappropriate as there are not many sections with zero accident frequency. After evaluating the predictive performances of all of the above mentioned models, it was found that the negative binomial model is the most desirable approach for this case.

Let y_i be the random variable that represents the number of accidents occurring at a given motorway segment I during a given time interval (one year in this case), where i = 1, 2, …, n, and y_i is a non-negative integer. In a Poisson regression model, y_i follows the Poisson probability law, which takes the following form:

$$P(y_i)=\frac{e^{(-\lambda)}\lambda_i^{y_i}}{y_i\,!}, \qquad (2)$$

where $P(y_i)$ is the probability of segment i experiencing y_i accidents over one year and λ_i is the Poisson parameter for segment i, which is equal to the expected number of accident per year on segment i, i.e. the mean of accident frequency, E(y_i). The Poisson regression model commonly assumes the log-linear relationship between Poisson parameter λ_i and explanatory variables

$$\lambda_i = E(y_i) = e^{\beta X_i}, \tag{3}$$

where X_i is a vector of explanatory variables, such as traffic, road and environmental characteristics of segment i, and β is a vector of unknown regression coefficients which can be estimated by the method of standard maximum likelihood.

The major advantage of the Poisson distribution is the simplicity in calculation due to its property of the mean equalling to the variance. The relationship is termed equidispersion, which is also known as its restriction. If E(y_i) > Var(y_i), it is said that the data are under-dispersed, oppositely over-dispersed if E(y_i) < Var(y_i). The Poisson regression model is inappropriate to use when the variance of accident data is significantly different from the mean. In this case negative binomial regression model can be applied as an alternative to overcome the problem. The negative binomial technique relaxes the assumption of equality of the mean and variance, by adding a gamma-distributed error term. Equation (3) is rewrote as,

$$\lambda_i = E(y_i) = e^{\beta X_i + \varepsilon_i}, \tag{4}$$

where ε_i is an error term, and $e^{\varepsilon i}$ is gamma-distributed error term with mean 1 and variance α^2. The addition of ε_i makes the variance to be different from the mean as follows:

$$VAR(y_i) = E(y_i)[1 + \alpha E(y_i)] = E(y_i) + \alpha E(y_i)^2, \tag{5}$$

where α is also called the dispersion parameter, which plays an important role in the determination of choosing the Poisson regression or the negative binomial regression model. When α is significantly different from zero, the distribution is under-dispersion or over-dispersion and the negative binomial model is appropriate. When α approaches zero, the variation is almost equal to the mean, and the distribution can be simply modelled by the Poisson regression technique.

The form of negative binomial probability distribution is given as:

$$P(y_i) = \frac{e^{(-\lambda_i e^{\varepsilon_i})}(\lambda_i e^{\varepsilon_i})^{y_i}}{y_i!}, \tag{6}$$

The formulation integrating ε_i is as follows:

$$P(y_i) = \frac{\Gamma((1/\alpha)+y_i!)}{\Gamma(1/\alpha)y_i!}\left(\frac{1/\alpha}{(1/\alpha)+\lambda_i}\right)^{1/\alpha}\left(\frac{\lambda_i}{(1/\alpha)+\lambda_i}\right)^{y_i}, \tag{7}$$

where $\Gamma(.)$ is a gamma function.

Like the Poisson model, the negative binomial model is also estimated by the standard maximum likelihood method. The corresponding likelihood function is:

$$L(\lambda_i) = \prod_i \frac{\Gamma((1/\alpha)+y_i!)}{\Gamma(1/\alpha)y_i!}\left(\frac{1/\alpha}{(1/\alpha)+\lambda_i}\right)^{1/\alpha}\left(\frac{\lambda_i}{(1/\alpha)+\lambda_i}\right)^{y_i}, \tag{8}$$

The function is maximized to obtain coefficient estimates for β and α.

Goodness of Fit

For linear regression models the coefficient of determination R^2 test is a conventional goodness of fit measure. For non-linear model (e.g. Poisson, negative binomial regression models), however, a pseudo-R^2 statistic is used. A variety of pseudo-R^2 has been formulated and the most commonly employed one is calculated as the following function:

$$\rho^2 = 1 - \mathrm{LL}(\beta)/\mathrm{LL}(0), \tag{9}$$

where LL(β) is the log likelihood at convergence having coefficient vector β, and LL(0) is the initial log likelihood having all coefficients set to zero. ρ^2 is an indication of the additional variance in accident frequency explained by the model to the constant term alone. The value of ρ^2 is bound between zero and one. A higher value represents a better fitted model.

6. ACCIDENT PREDICTION MODELS

Prior to the development of accident prediction models, Pearson's correlation analysis for aggregate data was processed to find out the linear relationship within every two independent factors. It is not appropriate for the variables representing factors with strong correlation to exist simultaneously in a model. The result of 7 pairs of strong correlations is shown in Table 2. They will be tested in the models to see which of them make significant effect on the independent variable.

A number of negative binomial regression models are developed to estimate accident frequency as a function of traffic conditions, geometric and operational characteristics of road, and weather conditions. The functional form is transformed into an explicit structure as follows:

$$E(y_i) = e^{\beta_0 + LnL + \beta_1 \times LnAADTperlane + \sum \beta_i \times x_i}, \tag{10}$$

Table 2 Candidate explanatory factors having significant correlation based on Pearson's correlation analysis

Factors affecting accident		R
Presence of on-ramp	On-ramp AADT	0.778
Presence of off-ramp	Off-ramp AADT	0.802
Average horizontal curvature	Maximum horizontal curvature	0.660
Average up-grade gradient	Maximum up-grade gradient	0.934
Average down-grade gradient	Maximum down-grade gradient	0.930
Annual rainfall	Annual wet days	0.833
Speed limit	Maximum horizontal curvature	0.748

where $E(y_i)$ = expected accident frequency in segment i,
L = segment length,
β_0 = intercept,
β_1 and β_i = model coefficients, and
x_i = independent variables in additional to *Ln L* and *LnAADTperlane.*

A 5-year period (2004 to 2008) dataset was employed in the model development, and a 2-year period (2009 and 2010) dataset was used for testing the prediction performance. The regression analysis can be performed by using the function of Generalized Linear Models in SPSS statistical software. The function can also evaluate the goodness of fit of model. An exclusive process was used to identify significant independent variables. Variables insignificant at 80% confidence level were excluded along with those having strong pairwise correlation

Table 3 shows the estimation results for overall accident frequency in the motorway. The results reveal that there are 11 variables that significantly contribute to the variance in number of accidents at different locations. T-statistics reveal some important variables that include the log of segment length, log of AADT per lane and the number of lanes. Motorway segments with higher percentage of heavy vehicle suffer from more frequent vehicle accidents, attributed to the fact that trucks occupy wider lane and have speed variance with cars. With regard to the ramps, both on-ramp and off-ramp traffic volumes are found to have positive effects on accidents. Generally, a large number of merging and diverging occur near the ramps that increase chances of accident due to speed variance and lane changing activities. A cluster of crashes occurred every year in the motorway sections near Auckland CBD where ramps are quite densely spaced. It is important to pay special attention to the spacing between motorway ramps. Turning to the road cross section features, as anticipated, the amount of accidents increases with the increase in number of lanes, as conflicts between vehicles changing lanes increase. With regard to median type, fewer accidents are observed on segments divided by steel median compared to those with concrete median. The median type (steel or concrete barrier) actually suggests the type of region. The higher accident frequency associated with concrete type is not because of the effect of median type, but because concrete medians are generally located in urban motorway where more accidents happen. Width of shoulder has increasing effect on the likelihood of accidents as indicated by positive sign in its computed coefficient value, which seems to be counterintuitive, and demands further investigations to explore possible reasons.

Table 3. Negative binomial regression model for overall accident frequency

Factor	Independent Variable	Coefficient	t-statistics
	Constant	-5.105	-11.357
Segment length	Ln length	0.806	12.231
Traffic conditions	Ln AADT per lane	2.006	19.307
	Percentage of heavy vehicle	0.039	2.294
Ramp	On-ramp AADT	0.013	2.766
	Off-ramp AADT	0.018	3.673
Cross section features	Number of lanes	0.634	15.890
	Median type (1 = steel, 0 = concrete)	-0.205	-3.198
	Width of shoulder	0.058	2.533
Road alignment	Average curvature	-0.200	-2.488
	Maximum up-grade	0.047	2.701
Weather condition	Annual rainfall	-0.574	-3.312
Over dispersion	α	0.183	9.196
Number of observation	N	652	
Log-likelihood at zero	LL(0)	-1732.431	
Log-likelihood at convergence	LL(β)	-1525.638	
Goodness of fit	ρ^2	0.119	

In terms of the road alignment features, the horizontal curvature appears to have a negative impact on accident frequency. A possible interpretation of this result is that visual effect created by curves becomes stronger as the curvature increases and drivers tend to be more cautious at locations with sharp horizontal curves. The result is consistent with some previous findings (Shankar et al., 1995; Anderson et al., 1999; Chang, 2005). In Anastasopoulos et al. (2012), the effect was called risk compensation. The maximum up-grade is found to be a significant vertical alignment factor. As the up-grade becomes steeper, accident frequency is expected to increases. The explanation is that up-grade lead to high variability in speed of vehicles. Lastly, the negative coefficient of rainfall indicates that motorway segments with more rainfall have less accident possibility. The finding is probably attributed to the fact that traffic volume is reduced in rainy days, and drivers are more cautious in driving and keep greater following distance. The relationship between precipitation and road accident is a more complicated issue that needs more specific investigation not covered in this study. The estimated α suggests existence of over dispersion of data, confirming that the negative binomial formulation was favoured rather than the Poisson.

The model development methodology was also applied separately to 58.3 km rural and 74.4 km urban motorway sections, as different model results are expected because of distinct traffic characteristics, geometric characteristics and accident frequency they have. Table 4 shows the results of rural and urban accident models using negative binomial regression. There are three most important factors for both models: segment length, traffic volume and number of lanes, and they all have positive effects on crash occurrence, which is in accord with common sense. Some interesting results are found on the variance of other explanatory variables in the two models. First of all, traffic volumes entering and exiting motorway (from on-ramps and off-ramps respectively) have significant impact on the safety of urban motorway sections.

Table 4. Negative binomial regression model for rural and urban accident frequency

Factor	Independent Variable	Rural segments		Urban segments	
		Coefficient	t-statistics	Coefficient	t-statistics
	Constant	-5.711	-3.439	-2.346	-4.422
Segment length	Ln length	0.863	6.674	0.904	12.151
Traffic conditions	Ln AADT per lane	1.751	3.369	1.393	8.907
Ramp	Presence of on-ramp (1 = with on-ramp, 0 = without on-ramp)	0.310	2.089	/	/
	On-ramp AADT	/	/	0.012	2.667
	Off-ramp AADT	/	/	0.019	3.958
Cross section features	Number of lanes	1.103	4.547	0.531	13.375
	Width of lane (1 = 3.6m, 0 = 3.5m)	0.584	2.983	/	/
	Median type (1 = steel, 0 = concrete)	/	/	-0.167	-2.546
	Width of median	-0.059	-1.879	/	/
Road alignment	Average curvature	/	/	-0.133	-1.692
	Maximum curvature	-0.232	-1.528	/	/
Weather condition	Annual rainfall	/	/	-0.698	-3.964
Over dispersion	α	0.167	2.323	0.170	8.852
Number of observation	N	227		425	
Log-likelihood at zero	LL(0)	-392.912		-1321.3	
Log-likelihood at convergence	LL(β)	-329.019		-1205.3	
Goodness of fit	ρ^2	0.163		0.088	

The accident frequency increases with increase in the AADT of ramps. However, they have no significant effect on rural motorway sections. The factor that matters for rural motorway sections is the presence of on-ramp. The insignificance of off-ramp variable for these sections might be related to lower traffic density in rural sections allowing a relatively safer deceleration and lane changing operations near off-ramps. Wider central median is found to have beneficial effects on the safety of rural motorway sections.

As to urban motorway sections, accidents are more likely to occur on segments installed concrete median barrier. Both rural and urban accident frequencies are negatively associated with horizontal road curvature, although it is represented by different variables in these models, that is, maximum horizontal curvature is used for rural sections while average horizontal curvature is used for urban sections. The effect rainfall on accidents is noticed only for urban sections. The goodness of fit statistic is noticeably better for rural model (0.163) than urban model (0.088).

Characteristics of traffic flows on segments without ramp, with on-ramp or with off-ramp are not supposed to be the same. Merging traffic and diverging traffic have predominant effects on on-ramp segments and off-ramp segments safety. Table 5 shows the estimation results for the models developed separately for segments without ramp, with on-ramp and with off-ramp. Segment length, AADT per lane and number of lanes are found especially significant in all the three models. As for heavy vehicle percentage, it has a significant impact on segments without ramp (t = 4.235) and a minor impact on segments with on-ramp (t = 1.777), but insignificant impact on segments with off-ramp. The results on the safety impacts of ramp traffic volume are interesting. For segments with off ramp, as anticipated the ramp AADT affects accident frequency positively; whilst for segments with on-ramp, the ramp AADT has a negative impact on accidents, which is in contrast with the previous results. This result can be attributed to congested traffic flow conditions near motorway sections on upstream side of the on-ramps with heavy entering traffic. This contributes to significantly slow moving traffic, which has less likelihood of accident occurrence.

For all the three types of segments, accidents are more likely to occur on places with concrete median barriers (mostly on urban motorway). Besides, width of median is only significant for accidents on off-ramp segments, where wider median decreases accident frequency. As to shoulder, the negative sign in the model of segments without ramp demonstrates that wider shoulder reduces accident number on these segments; however, segments with on-ramp or off-ramp have more accidents when shoulder is wider. The accident frequency on segments with on-ramp is not significantly affected by any horizontal or vertical alignment factor. Greater horizontal curves result in fewer numbers of crashes on non-ramp segments and off-ramp segments.

Motorway segments without ramp tend to experience added accident risks when they have steeper average up-grade and maximum down-grade. Uphill grades influence vehicle speed, especially for large trucks. The speed differentials on up-grades increase the likelihood of accidents. Breaking distance needed on steep down-grades is significantly longer compared to the one for flat road, which substantially increases the risk of rear end collisions.

Conversely, the two variables have negative effects on accident occurrence on off-ramp segments, which means accident likelihood is greater on off-ramp segments with relatively level vertical alignments. The positive coefficient value of speed limit shows accident frequency is expected to be higher on segments with 80km/h speed limit. This does not mean speed reduction is an attribute raising accident occurrence, however, it is because the speed limit is introduced as an improvement measure on motorway segments that already have high accident rates. The examination of statistic ρ^2 for the three models indicates the best fit goodness of fit for on-ramp segment model (ρ^2 = 0.194), then non-ramp segment model (ρ^2 = 0.131). Generally speaking, the goodness of fit is improved through the process of developing models separately for the three types of segments, compared with the overall model for which ρ^2 = 0.119.

Table 5. Negative binomial regression model for accident frequency on segments without ramp/with on-ramp/with off-ramp

Factor	Independent Variable	Segments without ramp		Segments with on-ramp		Segments with off-ramp	
		Coefficient	t-statistics	Coefficient	t-statistics	Coefficient	t-statistics
	Constant	-6.517	-8.202	-4.270	-5.936	-2.743	-4.022
Segment length	Ln length	0.814	5.839	1.578	9.900	0.536	3.027
Traffic conditions	Ln AADT per lane	2.311	13.244	1.906	10.702	1.111	5.073
	Percentage of heavy vehicle	0.166	4.235	0.043	1.777	/	/
Ramp	On-ramp AADT	/	/	-0.020	-2.198	/	/
	Off-ramp AADT	/	/	/	/	0.065	5.804
Cross section features	Number of lanes	0.775	5.831	0.509	7.366	0.539	8.370
	Median type (1 = steel, 0 = concrete)	-0.522	-3.387	-0.236	-2.258	-0.236	-2.258
	Width of median	/	/	/	/	-0.066	-4.151
	Width of shoulder	-0.094	-2.009	0.094	1.808	0.102	3.218
Road alignment	Average curvature	/	/	/	/	-0.448	-2.089
	Maximum curvature	-0.302	-3.301	/	/	/	/
	Average up-grade	0.107	1.935	/	/	-0.081	-1.633
	Maximum down-grade	0.154	3.929	/	/	-0.124	-3.272
Traffic operation	Speed limit (1 = 80km/h, 0 = 100km/h)	0.912	1.814	0.399	2.694	/	/
Weather condition	Annual rainfall	/	/	-0.446	-1.711	/	/
	Annual wet days	-0.006	-1.935	/	/	/	/
Over dispersion	α	0.130	4.167	0.106	4.344	0.121	4.172
Number of observation	N	241		192		190	
Log-likelihood at zero	LL(0)	-542.230		-551.177		-469.407	
Log-likelihood at convergence	LL(β)	-471.256		-444.216		-417.626	
Goodness of fit	ρ^2	0.131		0.194		0.110	

7. MODEL VALIDATION

The accident predictive abilities of the above mentioned models were examined according to the method described in Oh et al. (2003), Lord et al. (2008) and Couto and Ferreira (2011), where two measures were considered.

7.1 Mean Absolute Deviance (MAD)

MAD is an estimator of the average mis-prediction of the model, calculated as:

$$\text{Mean absolute deviance } \left(\text{MAD}\right) = \frac{1}{n}\sum_{i=1}^{n}\left|y_i' - y_i\right| \tag{11}$$

7.2 Mean-Squared Predictive Error (MSPE)

MSPE is used to measure the error associated with a validation or external dataset. It is computed as:

$$\text{Mean-squared predictive error } \left(\text{MSPE}\right) = \frac{1}{n}\sum_{i=1}^{n}\left(y_i' - y_i\right)^2 \tag{12}$$

where y'_i is the predicted accident frequency on segment i, and y_i is the observed accident frequency on segment i. The desirable model should fit the data as closely as possible, which is represented by the smaller value of MAD and MSPE.

The 5 years data between 2004 and 2008 were used to fit the model, and 2 years data of 2009 and 2010 were applied to test the predictive performance of model. MAD and MSPE of the two dataset were computed and presented in Table 6.

In order to justify the general accuracy level of the prediction results presented in this paper, we refer similar studies conducted by previous researchers. Lord et al. (2008) obtained MAD values of 2.4 and 4.1 and MSPE values of 21 and 33 for samples with 4.9 and 11.6 average accident frequencies respectively. The smallest MAD and MSPE values presented by Couto and Ferreira (2011) are 1.8 and 7.2 for a mean accident frequencies of 2.85 respectively. Given the average accident frequency of 8.77 in this study, it could be seen that the models have produced fairly satisfactory predictive performance as presented in Table 6. Furthermore, considering the randomness of accident occurrence, which is largely due to the unexpectable risky driver situations including fatigue, inattention and traffic violation, as well as poor vehicle conditions, it was acceptable that the factors involved in the prediction models could just partly explain the motorway crashes happened in reality.

Table 6 also proves statistical improvements from the models for various segments in terms of the computed MAD and MSPE values using both 5 years fitting data and 2 years predicting data. The improvement is not much for rural and urban models; however it is significant for the models separately for segments without ramp, with on-ramp and with off-ramp. Therefore a conclusion is drawn that applying different models for motorway segments without ramp, with on-ramp and with off-ramp can obtain more precise accident frequency prediction.

Table 6. Predictive performances evaluation measures

Measure	Overall model	Rural + urban section models	Non-ramp + on-ramp + off-ramp segment models
MAD_{Fit}	3.71	3.62	3.21
$MSPE_{Fit}$	33.25	31.55	24.92
MAD_{Pred}	4.07	3.98	3.70
$MSPE_{Pred}$	36.60	34.23	27.87

Prediction model was also developed for overall accident frequency by applying GEE technique and the predictive performance of GEE model was evaluated (MAD_{Fit} =3.74, $MSPE_{Fit}$ =34.46). This outcome indicated that negative binomial model slightly surpass GEE model in terms of predictive ability in this case.

8. CONCLUDING REMARKS

This paper presents a number of accident prediction models developed using negative binomial regression models under three categories: first for the whole of the motorway, second for rural and urban motorway segments separately and third for motorway segments without ramp, with on-ramp and with off-ramp separately. These models relate accident frequency to their non-behavioural contributing factors. The results showed that the third category of models obtained considerably improved accident prediction results compared with the other two categories of the models.

This study also captured explanatory variables that have significant effects on motorway safety. The variables associated with traffic conditions, road geometric characteristics, operational characteristics and weather conditions were selected through the process of model development. The segment length, AADT per lane and number of lanes were identified as the most critical factors in all the models. Motorway segments with higher percentage of heavy vehicles are more likely to have accidents than other segments. The effects of ramps were investigated in terms of the presence of ramp and AADT of ramp. On rural motorway, the presence of on-ramp rather than off-ramp was identified as a contributor to accident numbers, while on urban motorway the traffic volume from both on-ramp and off-ramp create more accidents. An interesting finding is that for segments with on-ramp, the ramp AADT has a negative effect on accidents. Wider central median could improve the rural motorway safety, and wider shoulder is able to benefit the segments away from ramps. Generally greater horizontal curvature of a segment, either average or maximum curvature, tends to decrease accident likelihood by virtue of the visual effect that can caution the drivers. As to vertical features, accidents on segments without ramp are expected to increase as the up-grade and down-grade slopes increase. Oppositely the segments with off-ramp appear to be riskier when flatter terrains are observed. The results in most models manifest that an increase in rainfall or wet days will reduce accident frequency, which contradicts common sense and need further research.

The accident prediction models can be used to assess the safety performance and identify hazardous locations needing treatment for both existing and proposed roads. In terms of future work, it is suggested to extend the accident prediction models by taking into account the combined effect of different explanatory variables. The safety impact of interaction between different factors has always been a focus of interest for traffic engineers, for example, safety impact of interaction between horizontal and vertical alignments, and interaction between weather conditions and road geometry. The variables representing combination of factors are worth testing in future models.

REFERENCES

Abdel-Aty, M.A., Radwan, A.E. (2000) Modelling traffic accident occurrence and involvement. *Accident Analysis and Prevention*, Vol. 32, pp. 633-642.

Abdel-Aty, M., Abdalla, M.F. (2004) Linking roadway geometrics and real-time traffic characteristics to model daytime freeway crashes: generalized estimating equations for correlated data. *Transportation Research Record*, Vol. 1897, pp. 106-115.

Anastasopoulos, P.C., Mannering, F. L., Shankar, V. N., Haddock J. E. (2012) A study of factors affecting highway accident rates using the random-parameters tobit model. *Accident Analysis and Prevention*, Vol. 45, pp.628-633.

Anderson I.B., Bauer, K., Harwood D.W., Fatzpatrick, K. (1999) Relationship to Safety of Geometric Design Consistency Measures for Rural Two-Lane Highways. *Transportation Research Record*, Vol. 1658, pp. 43-51.

Caliendo, C., Guida, M., Parisi, A. (2007) A crash-prediction model for multilane roads. *Accident Analysis and Prevention*, Vol. 39, pp. 657-670.

Chang L. (2005) Analysis of freeway accident frequencies: Negative binomial regression versus artificial neural network. *Safety Science*, Vol. 43, Issue 8, pp. 541-557.

Couto, A., Ferreira, S. (2011) A note on modelling road accident frequency: a flexible elasticity model. *Accident Analysis and Prevention*, Vo. 43, pp. 2104-2111.

Economic evaluation manual (2010) New Zealand Transport Agency, Web link: http://www.nzta.govt.nz/resources/economic-evaluation-manual/volume-1/index.html.

Eenink, R., Reurings, M., Elvik, R., Cardoso, J., Wichert, S., Stefan, C. (2007) Accident prediction models and road safety impact assessment: recommendations for using these tools. RIPCORD-iSEREST Deliverable D2. BASt.

EI-Basyouny, K., Sayed, T. (2006) Comparison of Two Negative Binomial Regression Techniques in Developing Accident Prediction Models. *Transportation Research Record*, Vol. 1950, pp. 9-16.

Hadi, M.A., Aruldhas, J., Chow, L., Wattleworth, J.A. (1995) Estimating safety effects of cross-section design for various highway types using negative binomial regression.

Transportation Research Record, Vol.1500, pp. 169-177.

Jovanis, P.P., Chang, H. (1986) Modelling the relationship of accidents to miles travelled. *Transportation Research Record*, Vol. 1068, pp. 42-51.

Kopelias, P., Papadimitriou, F., Papandreou, K., Prevedouros, P. (2007) Urban Freeway Crash Analysis: Geometric, Operational, and Weather Effects on Crash Number and Severity. *Transportation Research Record*, Vol. 2015, pp. 123-131.

Lord, D., Guikema, S.D., Geedipally, S.R. (2008) Application of the Conway-Maxwell-Poisson generalized linear model for analyzing motor vehicle crashes. *Accident Analysis and Prevention*, Vol. 40, pp. 1123-1134.

Lord, D., Manar, A., Vizioli, A. (2005) Modelling crash-flow-density and crash-flow-V/C ratio relationships for rural and urban freeway segments. *Accident Analysis and Prevention*, Vol. 37, pp. 185-199.

Lord, D., Persaud, B.N. (2000) Accident prediction models with and without trend: application of the generalized estimating equations procedure. *Transportation Research Record*, Vol. 1717, pp. 102-108.

Lord, D., Washington, S.P., Ivan, J.N. (2005) Poisson, Poisson-gamma and zero-inflated regression models of motor vehicle crashes: balancing statistical fit and theory. *Accident Analysis and Prevention*, Vol. 37, pp. 35-46.

Miaou, S. (1994) The relationship between truck accidents and geometric design of road sections: Poisson versus Negative Binomial Regressions. *Accident Analysis and Prevention*, Vol. 26, pp. 471-482.

Montella, A. (2008) Crash prediction models for rural motorways. *Transportation Research Record*, Vol. 2083, pp. 180-189.

Oh, J., Lyon, C., Washinton, S.P., Persaud, B.N., Bared, J. (2003) Validation of the FHWA crash models for rural intersections: lessons learned. *Transportation Research Record*, Vol. 1840, pp. 41-49.

Pemmanaboina, R. (2005) Assessing occurrence on urban freeways using static and dynamic factors by applying a system of interrelated equations. M.E. thesis, University of Central Florida.

Persaud, B., Dzbik, L. (1993) Accident prediction models for freeways. *Transportation Research Record*, Vol. 1401, pp. 55-60.

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

Shankar, V., Milton, J., Mannering, F. (1997) Modelling accident frequencies as zero-altered probability processes: an empirical inquiry. *Accident Analysis and Prevention*, Vol. 29, pp. 829-837.

Turner, S.A. (2001) Accident prediction models. *Transfund New Zealand Research Report 192*.