Available online at www.sciencedirect.com

ScienceDirect

Transportation Research Procedia 25C (2017) 664–676

Transportation Research Procedia

www.elsevier.com/locate/procedia

World Conference on Transport Research - WCTR 2016 Shanghai. 10-15 July 2016

Freeway crash frequency modeling under time-of-day distribution

Yu-Chiun Chiou*, Yu-Chun Sheng, Chiang Fu

Department of Transportation and Logistics Management, National Chiao Tung University, 4F, 118, Sec. 1, Chung-Hsiao W. Rd. Taipei, 100, Taiwan

Abstract

This study aims to identify key factors affecting crash frequencies under various times of the day, so as to propose effective and time-specific countermeasures. Two approaches are proposed and compared. The clustering approach combines a crash count model to predict total number of crashes and a clustering model to divide segments into clusters according to their time-of-day distribution patterns of crash frequency. The multivariate approach treats the crash frequencies of various times of the day as target variables and accommodates potential correlation among them. Crash datasets of Taiwan Freeway No.1 are used to estimate and validate the models. Four times of the day, late-night/dawn (24-06), morning/noon (07-13), afternoon/evening (14-19), and night (20-23) are formed according to crash count distribution. In terms of Adj-MAPE and RMSE, the clustering approach performs better than the multivariate approach. According to the clustering results, segments in metropolitan areas have higher crash frequency in the afternoon/evening, while those in rural areas have higher crash frequency in late-night/dawn, suggesting the time-of-day distributions of crash frequency markedly differ among segments. Time-specific countermeasures are then proposed accordingly.

Peer-review under responsibility of WORLD CONFERENCE ON TRANSPORT RESEARCH SOCIETY.

Keywords: Time-of-day crash frequency distribution; negative binomial regression; clustering; multivariate modeling approach.

1. Introduction

Many studies have developed crash frequency models to identify factors contributing to crash counts at roadway segments or at intersections during a certain time period (usually one year) (e.g., Jones et al., 1991; Miaou, 1994;

* Corresponding author. Tel: +886-2-23494940
E-mail address: ycchiou@mail.nctu.edu.tw (Y.C. Chiou)

2352-1465
Peer-review under responsibility of WORLD CONFERENCE ON TRANSPORT RESEARCH SOCIETY.
10.1016/j.trpro.2017.05.450

Fridstrom et al., 1995; Shankar et al., 1995; Poch and Mannering, 1996; Shankar et al., 1997; Milton and Mannering, 1998; Ivan et al., 1999; Ivan et al., 2000; Abdel-Aty and Radwan, 2000; Khattak et al., 2002; Wang and Nihan, 2004; Lord, 2006; Wong et al., 2007; Malyshkina and Mannering, 2010). However, only a few of studies further examined risk factors contributing to crash counts at various times of the day (e.g., morning, afternoon, and night) which can definitely provide more useful information for proposing effective and time-specific countermeasures. For example, Doherty et al. (1998) studied the distribution of crash frequency at times of the day. They showed a very high crash frequency for drivers aged 16 to 19 to drive at night. Clarke et al. (2006) investigated how age, driving experience, and time of day affect the crash frequency of young drivers and the results suggested that the problems of accidents in darkness are not a matter of visibility, but a consequence of the way young drivers use the roads at night. Qin et al. (2006) examined the relationship between crash occurrence and hourly traffic. The results revealed how the relationship between crashes and hourly traffic varies by time of day, thus improving the accuracy of crash occurrence predictions. The results show that even accounting for time of day, the hourly traffic is indeed non-linear of crash occurrence, implying that at any time of day, the crash occurrence is not proportional to the hourly traffic. Marquis (2014) analyzed the truck-related crash occurrences in Manhattan, New York over four time blocks: the morning peak (6:00-10:00), the mid-day (10:00-15:00), the afternoon peak (15:00-19:00), and the night time (19:00-6:00) by using zero-inflated negative binomial models. The study found that both the built environment and the traffic flows contribute to the temporal variation of truck-related crash occurrence.

Most of the above studies examined the effect of time of day on crash counts by introducing a time variable into the model or by modelling crash counts at various time periods separately. The former has difficulty in investigating the different effects of risk factors on crash counts at various time periods and the latter ignores the potential correlation among crash counts at various time periods. Thus, this study aims to simultaneously model the crash counts at various time periods. The model framework is similar to those modelling crash frequencies by severity levels (e.g., property damage only, injury, and fatality) and collision types (e.g., rear-end, head-on, sideswipe, and right angle) or by collision types (Milton et al., 2008; Ye et al., 2009; Naderan and Shahi, 2010; Aguero-Valverde, 2013; Chiou and Fu, 2013; Ye et al., 2013; Chiou et al., 2014; Venkataraman et al., 2014; Chiou and Fu, 2013). The remainder of this paper is organized as follows. Section 2 presents the proposed models. Section 3 addresses data collection and descriptive statistics of the study crash dataset. Section 4 presents the model estimation results and comparisons. Concluding remarks and suggestions are then given in Section 5.

2. Model

To identify the key factors contributing to crash counts at various times of day, two approaches are proposed, clustering and multivariate modeling approaches, as narrated below, respectively.

2.1. Clustering approach

The proposed clustering approach contains two stages. The first stage is to predict total crashes on each segment by using commonly adopted Poisson (PO) and Negative binomial (NB) models. The second stage is to divide freeway segments into finite clusters according to their time-of-day crash frequency distribution patterns. The average time-of-day crash frequency distribution of each cluster is then used to represent the segments belonging to it.

The PO model is the most fundamental count model. The probability function of PO model (Miaou, 1994; Jones et al., 1991) is expressed as Eq. (1).

$$PO(y_i) = \frac{\lambda_i^{y_i} e^{-\lambda_i}}{y_i!} \tag{1}$$

where PO (y_i) is the probability of y_i accidents occurring on roadway segment i under a specific time period. λ_i is the expected number of accidents on roadway segment i at the time period, which is defined in non-negative numbers. For estimation purposes, λ_i is usually specified as Eq. (2):

$$\lambda_i = \exp(\beta' X_i) \tag{2}$$

where X_i and β are the vectors of explanatory variables (such as road geometry, traffic, and weather conditions) and associated estimated coefficients, respectively.

The PO distribution assumes the equality of mean and variance. However, the empirical studies usually show that crash counts of some road segments are extremely high, while others are rather low in contrast, suggesting variance may exceed mean (over-dispersion) and violating the underlying assumption of the PO model. Consequently, the NB model is an appropriate alternative option to crash frequency modelling (Milton and Mannering, 1998; Poch and Mannering, 1996).

The NB model adds an error term ε_i to the expected cash frequency (λ_i) as shown in Eq. (3):

$$\lambda_i=\exp(\beta' X_i+\varepsilon_i) \tag{3}$$

where the error term ε_i in Eq. (3) is a gamma-distributed error term with mean one and variance α which represents the degree of over-dispersion. Thus, the formulation of NB probability density distribution, $NB(y_i)$, can be expressed as:

$$NB(y_i)=\frac{\Gamma[(1/\alpha)+y_i]}{\Gamma(1/\alpha)y_i!}\left(\frac{1/\alpha}{(1/\alpha)+\lambda_i}\right)^{1/\alpha}\left(\frac{\lambda_i}{(1/\alpha)+\lambda_i}\right)^{y_i} \tag{4}$$

Secondly, according to our field observations, the time-of-day distributions of various freeway segments remarkably differ due to the traffic, weather, lighting or other unknown reasons. Based on this, the crash percentages of 24 hours is used as the clustering variables (a total of 24 variables) to divide road segments into clusters. The freeway segments belonging to the same cluster exhibit the similar time-of-day distribution of crash counts. The choice of cluster algorithms depends on samples sizes. It is appropriate for a sample size below 200 cases to use the layering method, while the *K*-means method is usually recommended for a sample size more than 200 cases (Karlaftis and Tarko, 1998; Anderson, 2009; Mohamed et al., 2013).

Afterward, the estimated crash counts of a freeway segment i at time period t can be determined by multiplying the crash counts (λ_i) estimated by the crash frequency model and the average crash percentages of time period t (p_{tc}) of cluster c which the segment is classified to.

2.2. Multivariate modeling approach

In order to reflect the correlations among various types of crashes, Ye et al. (2009; 2013) simultaneously estimated T expected crash counts ($\lambda_{it,}$ i=1, 2, …, N; t=1,2,…, T) at segment i by specifying error components coefficient (δ), which assumes the random error term follows a normal distribution. This conceptual framework can be applied to multivariate modeling of crash frequencies at various time periods of a day. Taking a case of four time periods (i.e. n=4) for instance, the crash counts of four time periods can be expressed as Eq. (5).

$$\begin{aligned}
&ln(\lambda_{i1})=\beta_1' X_{i1}+\varepsilon_{i1}=\beta_1' X_{i1}+\delta_1\mu_1\\
&ln(\lambda_{i2})=\beta_2' X_{i2}+\varepsilon_{i2}=\beta_2' X_{i2}+\delta_2\mu_1+\delta_3\mu_2\\
&ln(\lambda_{i3})=\beta_3' X_{i3}+\varepsilon_{i3}=\beta_3' X_{i3}+\delta_4\mu_1+\delta_5\mu_2+\delta_6\mu_3\\
&ln(\lambda_{i4})=\beta_4' X_{i4}+\varepsilon_{i4}=\beta_4' X_{i4}+\delta_7\mu_1+\delta_8\mu_2+\delta_9\mu_3+\delta_{10}\mu_3
\end{aligned} \tag{5}$$

where X_{it} denotes the explanatory variables vectors of different time periods, t=1, 2, 3, 4. β_t is the coefficient vector of corresponding explanatory variables to be estimated. ε_{it} is the error term. μ_j is a standard normal distributed random variable, j=1, 2, …, 10. δ_j are the coefficients corresponding to μ_j to be estimated. The variance-covariance matrix of crash frequencies of four time periods can then be expressed as:

$$Var(\varepsilon_1)=\delta_1^2;\ Var(\varepsilon_2)=\delta_2^2+\delta_3^2;\ Var(\varepsilon_3)=\delta_4^2+\delta_5^2+\delta_6^2;\ Var(\varepsilon_4)=\delta_7^2+\delta_8^2+\delta_9^2+\delta_{10}^2 \tag{6}$$

$$Cov(\varepsilon_1,\varepsilon_2)=\delta_1\delta_2;\ Cov(\varepsilon_1,\varepsilon_3)=\delta_1\delta_4;\ Cov(\varepsilon_1,\varepsilon_4)=\delta_1\delta_7;$$

$$Cov(\varepsilon_2, \varepsilon_3)= \delta_2\delta_4+\delta_3\delta_5;\ Cov(\varepsilon_2, \varepsilon_4)= \delta_2\delta_7+\delta_3\delta_8;$$
$$Cov(\varepsilon_3, \varepsilon_4)= \delta_4\delta_7+\delta_5\delta_8+\delta_6\delta_9 \tag{7}$$

$$Corr(\varepsilon_1,\varepsilon_2)=\frac{Cov(\varepsilon_1,\varepsilon_2)}{\sqrt{Var(\varepsilon_1)Var(\varepsilon_2)}}, Corr(\varepsilon_1,\varepsilon_3)=\frac{Cov(\varepsilon_1,\varepsilon_3)}{\sqrt{Var(\varepsilon_1)Var(\varepsilon_3)}},$$
$$Corr(\varepsilon_1,\varepsilon_4)=\frac{Cov(\varepsilon_1,\varepsilon_4)}{\sqrt{Var(\varepsilon_1)Var(\varepsilon_4)}} Corr(\varepsilon_2,\varepsilon_3)=\frac{Cov(\varepsilon_2,\varepsilon_3)}{\sqrt{Var(\varepsilon_2)Var(\varepsilon_3)}}, \tag{8}$$
$$Corr(\varepsilon_2,\varepsilon_4)=\frac{Cov(\varepsilon_2,\varepsilon_4)}{\sqrt{Var(\varepsilon_2)Var(\varepsilon_4)}} Corr(\varepsilon_3,\varepsilon_4)=\frac{Cov(\varepsilon_3,\varepsilon_4)}{\sqrt{Var(\varepsilon_3)Var(\varepsilon_4)}}$$

where Eqs. (6) and (7) are the variance and covariance of ε_1, ε_2, ε_3 and ε_4, respectively. Eq. (8) represents the correlation coefficients of ε_1, ε_2, ε_3 and ε_4. The flexible framework of assigned error components to various variates largely simplifies complicated potential correlation among crash counts of various time periods and increase computation efficiency. The estimated correlation coefficient with an absolute value close to 1 indicates perfect correlation. An absolute value of the correlation coefficient between 0.7 to 0.9 expresses high correlation, while a medium correlation would be between 0.3 to 0.6. Otherwise, it is slightly or not correlated.

Under the assumption of Poisson distribution (Ye et al., 2013) and the condition on μ_j of λ_{it}, the conditional probability function $PO_t(y_{it}|\lambda_{it})$ of four crash frequencies y_{it} is expressed as Eq.(9).

$$PO_1\left(y_{i1}\middle|\lambda_{i1}(u_1)\right)=\frac{\exp(-\lambda_{i1})\lambda_{i1}{}^{y_{i1}}}{y_{i1}!}$$
$$PO_2\left(y_{i2}\middle|\lambda_{i2}(u_{j=1,2})\right)=\frac{\exp(-\lambda_{i2})\lambda_{i2}{}^{y_{i2}}}{y_{i2}!}$$
$$PO_3\left(y_{i3}\middle|\lambda_{i3}(u_{j=1,2,3})\right)=\frac{\exp(-\lambda_{i3})\lambda_{i3}{}^{y_{i3}}}{y_{i3}!} \tag{9}$$
$$PO_4\left(y_{i4}\middle|\lambda_{i4}(u_{j=1,2,3,4})\right)=\frac{\exp(-\lambda_{i4})\lambda_{i4}{}^{y_{i4}}}{y_{i4}!}$$

The probability functions of four crash counts in Eq.(9) are then multiplied, making the MPO (multivariate Poisson) joint probability function as Eq. (10).

$$\begin{aligned}&\mathrm{MPO}(y_i|\lambda(\mu_j))\\ &=\mathrm{PO}_1(y_1\ |\lambda(\mu_1))\cdot \mathrm{PO}_2(y_2|\lambda(\mu_1,\mu_2))\cdot \mathrm{PO}_3(y_3|\lambda(\mu_1,\mu_2,\mu_3))\cdot \mathrm{PO}_4(y_4|\lambda(\mu_1,\mu_2,\mu_3,\mu_4))\\ &=\iiint \mathrm{PO}(y_k|\lambda(\mu_k))\cdot g(\mu_{k=1,2,3,4})\cdot d(\mu_1)\cdot d(\mu_2)\cdot d(\mu_3)\cdot d(\mu_4)\end{aligned} \tag{10}$$

2.3. Goodness-of-fit indices

The best number of clusters can be determined depends on the pre-set criterion of cluster analysis. This study uses *F* value to determine the optimal number of clusters, which can be expressed as (Chiou and Lan, 2001):

$$F=\frac{N-m}{m-1}\times\frac{\sum_{k=1}^{m} n_k\left[(\overline{x_{1k}}-\overline{\overline{x_1}})^2+(\overline{x_{2k}}-\overline{\overline{x_2}})^2+\quad+(\overline{x_{tk}}-\overline{\overline{x_t}})^2\right]}{\sum_{k=1}^{m}\sum_{i=1}^{n_k} n_k\left[(x_{1i}-\overline{x_{1k}})^2+(x_{2i}-\overline{x_{2k}})^2+\quad+(x_{2i}-\overline{x_{2k}})^2\right]} \tag{11}$$

$$\overline{x_{tk}}=\frac{\sum_{j=1}^{n_k} x_{tj}}{n_k} \tag{12}$$

$$\overline{\overline{x_t}}=\frac{\sum_{i=1}^{N} x_{ti}}{N} \tag{13}$$

where N is the total number of segments), m is the number of clusters, n_k is the number of segments in cluster k, and

x_t is a vector of clustering variables.

To evaluate the goodness-of-fit, two proposed approaches: Adjusted Mean Absolute Percentage Error (*Adj*-MAPE) and Root Mean Square Error (RMSE) are adopted:

$$Adj-MAPE(\%)=\frac{\sum_{j=1}^{t}\frac{\sum_{i=1}^{N}\left|\frac{A_{ij}-P_{ij}}{(A_{ij}-P_{ij})/2}\right|}{N}}{t}\times 100 \tag{14}$$

$$RMSE(\%)=\sqrt{\frac{\sum_{j=1}^{t}\frac{\sum_{i=1}^{N}(A_{ij}-P_{ij})}{N}}{t}}\times 100 \tag{15}$$

where A_{ij} refers to the actual crash counts of segment i at time period j, P_{ij} refers to the predicted crash counts of segment i at time period j. Actual crash counts of some segments at some time periods are zero, which results in a computation error according to the original MAPE formula. Thus, the adjusted-MAPE is adopted instead.

3. Data

The accident datasets for Taiwan's Freeway No.1 in 2005 and 2006 were collected. The Freeway runs north-south, is 373.3 km long, and has 63 interchanges. To facilitate model estimation, a study segment is formed by two adjacent interchanges. By separately considering north- and south-bound directions, 124 analytical samples are obtained. The accident database, which is maintained by the National Highway Police Bureau (NHPB), contains accident information, such as crash severity, location and time of an accident, and number and types of vehicles involved. Covariates regarding geometric and facility characteristics were extracted from as-constructed freeway drawings and the online website of the Taiwan National Freeway Bureau. Traffic data collected by the Taiwan National Freeway Bureau include traffic volumes and occupancy of small vehicles, large vehicles, and tractor trailers. Rainfall data were obtained directly from the Taiwan Central Weather Bureau.

Since the lengths of segments markedly differ, the dependent variable is represented by the crash count divided by the segment length (GL) to better reflect the crash risk. While constructing the time period of the Multivariate Poisson model, the response variant is the crash number in all time periods. Model estimation was facilitated by determining four time periods with distinct numbers of crash counts according to the real distribution of crash frequency. Four time periods are "Late-night/Dawn"(00-06), "Morning/Noon"(07-13), "Afternoon/Evening" (14-19), and "Night" (20-23). Table 1 shows the crash count per km for the four time periods. The table indicates that the afternoon/evening time period has the highest crash counts, followed by morning/noon time period and the late-night/dawn period has the lowest crash counts. The distribution of crash counts is consistent with the pattern of traffic flow, confirming the results of most studies. As reported in most studies, the distribution indicates that traffic is the most important exposure measure of crashes.

Table 1. Descriptive statistics for dependent variable (crash counts *per km*)

Crash counts per km (=*Crash counts/segment length*)	*Mean*	*Std.*	*Min.*	*Max.*
Total crash counts	17.32	16.55	2.19	85.29
Crash counts by time periods of days				
Night	1.89	2.07	0.00	11.58
Late-night / Dawn	1.73	1.24	0.00	5.59
Morning / Noon	6.13	6.75	0.40	37.89
Afternoon / Evening	7.58	8.48	0.52	39.13

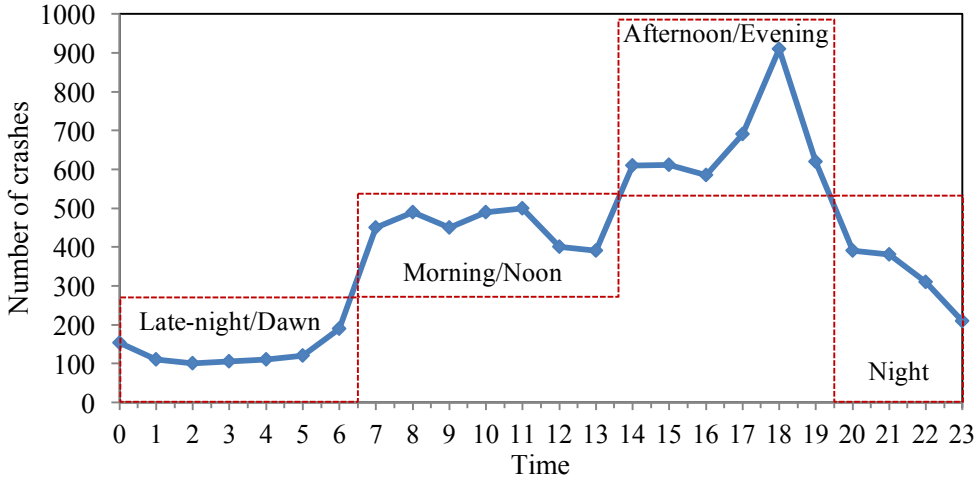


Fig. 1. Time-of-day crash distribution

Table 2 shows that the explanatory variables considered in this study can be divided into three categories: Geometric factors, Environmental factors, and Traffic factors. Geometric features of the freeway include maximum upward and downward slopes, curvature, Clothoid parameter, posted speed limit, and number of lanes. Environmental factors include the number of speeding cameras; annual rainfall; rest area or toll station; adjacent urban area, airport, seaport or industrial area; and freeway system interchanges. The traffic factors include total traffic (average daily traffic volume), traffics of various vehicles types (including small vehicles, buses and trucks, and tractor-trailers), and percentage of large vehicles (including buses, trucks, and tractor-trailers).

Table 2. Descriptive statistics of explanatory variables

Explanatory variables	Mean	Std.	Min.	Max.
Geometric factors				
Maximum upward slope (%)	1.25	2.08	0.00	13.70
Maximum downward slope (%)	1.17	1.41	0.00	5.20
Curvature (‰)	0.73	1.12	0.00	7.10
Clothoid parameter (1000 degrees)	0.93	0.94	0.00	3.25
Posted speed limit (100 km/h)	1.11	0.08	1.00	1.20
Number of lanes	2.73	.53	2.00	4.00
Environmental factors				
Number of speeding cameras	3.57	3.96	0.00	22.00
Annual rainfall (1000 millimeters)	2.80	0.57	1.48	4.24
Dummy variable(yes=1, no=0)				
Presence of rest area	0.10	0.30	0.00	1.00
Presence of toll station	0.16	0.37	0.00	1.00
Adjacent to urban area	0.47	0.50	0.00	1.00
Adjacent to airport, seaport or industrial area	0.32	0.47	0.00	1.00
Adjacent to freeway system interchanges	0.23	0.42	0.00	1.00
Traffic factors				
Total traffic (1000 pcu/hr)	3.14	1.03	1.31	6.40
Night	3.40	1.21	1.40	7.38
Late-night / Dawn	1.74	0.58	0.65	3.55
Morning / Noon	3.74	1.24	1.58	7.74
Afternoon / Evening	4.17	1.40	1.77	8.67
Traffic of buses (1000 veh/hr)	0.24	0.12	0.07	0.56
Night	0.25	0.13	0.07	0.63
Late-night / Dawn	0.16	0.08	0.03	0.41

Morning / Noon	0.28	0.13	0.08	0.62
Afternoon / Evening	0.31	0.16	0.08	0.84
Traffic of light duty vehicles (1000 veh/hr)	1.61	0.69	0.49	4.36
Night	1.92	0.88	0.55	5.46
Late-night / Dawn	0.69	0.31	0.22	2.14
Morning / Noon	1.96	0.85	0.58	5.35
Afternoon / Evening	2.28	0.97	0.73	6.40
Traffic of heavy-duty vehicles (%)	21.51	5.96	9.34	33.02
Night	18.49	5.30	7.18	29.12
Late-night / Dawn	30.23	9.50	9.33	48.36
Morning / Noon	20.86	5.88	9.43	32.64
Afternoon / Evening	19.33	5.44	7.97	30.23
Traffic of tractor-trailers (1000 veh/hr)	0.17	0.08	0.07	0.67
Night	0.15	0.07	0.06	0.56
Late-night / Dawn	0.12	0.06	0.03	0.48
Morning / Noon	0.20	0.10	0.08	0.87
Afternoon / Evening	0.20	0.10	0.07	0.77

4. Results

The clustering approach was used to classify segments into clusters according to their crash percentages in various time periods of a day. The commonly adopted *K*-means method is chosen, and the number of clusters, *K*, is determined by the *F*-value. Fig. 2 depicts the distribution of *F*-values under various numbers of clusters. Note that K=4 has the highest *F*-value and it is used to form clustering by *K*-means method.

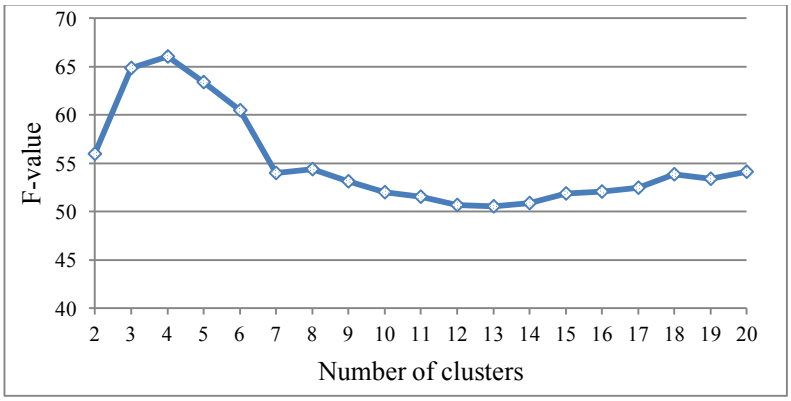


Fig. 2. Distribution of *F*-values under various numbers of clusters

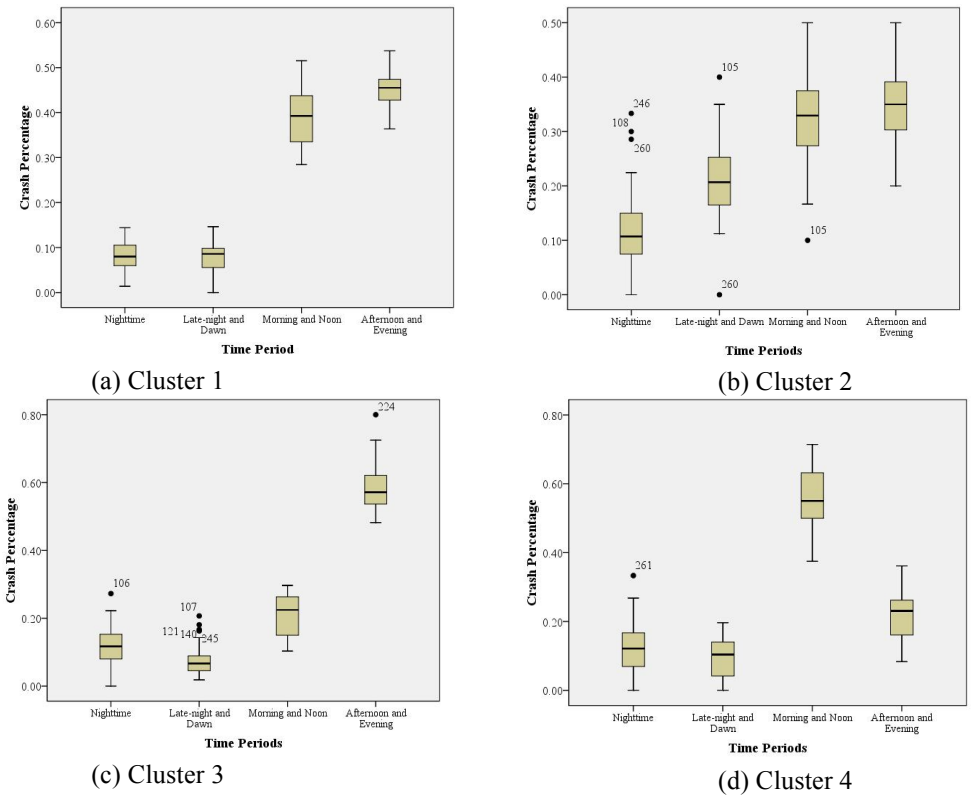
To examine the differences among four clusters, the mean values of explanatory variables of four clusters are reported in Table 3. Table 3 gives the average values of explanatory variables of four clusters. Table 3 shows that the Cluster 1 segments have the lowest downward slope, curvature, and percentage of large vehicles. These segments have no toll stations but have the highest total traffic and small vehicles, the highest percentage of segments in 100km/hr speed limits and are adjacent to urban areas, airport, port, and industrial park. These segments are mostly in metropolitan areas. The Cluster 2 segments have the highest maximum upward and downward slopes, curvature rate, Clothoid parameter values, number of speeding cameras, and percentage of large vehicles. These segments are mostly in rural areas. The Cluster 3 segments have the highest percentage of segments in 100km/hr speed limits, bus and truck traffic, presence of toll stations, but have the lowest annual rainfall and tractor-trailer traffic. The Cluster 4 segments have the lowest maximum upward slope, Clothoid parameter values, percentage of segments in 100km/hr speed limits, number of lanes, number of speeding cameras, and traffic in most vehicle types but have the highest annual rainfall and largest number of segments adjacent to system interchanges.

Table 3. Means of explanatory variables of the segments in four clusters

Means	Cluster 1	Cluster 2	Cluster 3	Cluster 4	Total
Freeway geometrics					
Maximum upward slope (%)	1.05	1.54	1.12	0.86	1.25
Maximum downward slope (%)	0.72	1.63	0.83	0.90	1.17
Curvature (‰)	0.48	0.90	0.59	0.69	0.73
Clothoid parameter (1000 degrees)	0.49	1.28	0.95	0.45	0.93
Posted speed limit (100 km/h)	1.12	1.11	1.12	1.08	1.11
Number of lanes	2.77	2.76	2.72	2.61	2.73
Environmental factors and Freeway facilities					
Number of speeding cameras	2.46	4.65	3.56	1.89	3.57
Annual rainfall (1000 millimeters)	2.87	2.73	2.68	3.08	2.80
Dummy variable(yes=1, no=0)					
Presence of rest area	0.12	0.15	0.04	0.00	0.10
Presence of toll station	0.00	0.20	0.28	0.11	0.16
Adjacent to urban area	0.58	0.40	0.44	0.56	0.47
Adjacent to airport, seaport or industrial area	0.42	0.27	0.28	0.39	0.32
Adjacent to system interchanges	0.19	0.25	0.16	0.28	0.23
Traffic characteristics					
Total traffic (1000 pcu/hr)	3.36	3.11	3.16	2.88	3.14
Night	3.65	3.34	3.52	3.09	3.40
Late-night / Dawn	1.86	1.74	1.73	1.57	1.74
Morning / Noon	4.09	3.71	3.59	3.52	3.74
Afternoon / Evening	4.35	4.13	4.38	3.77	4.17
Buses (1000 veh/hr)	0.22	0.25	0.26	0.21	0.24
Night	0.24	0.26	0.28	0.21	0.25
Late-night / Dawn	0.14	0.17	0.17	0.13	0.16
Morning / Noon	0.27	0.29	0.28	0.25	0.28
Afternoon / Evening	0.28	0.33	0.35	0.26	0.31
Light-duty vehicles (1000 veh/hr)	1.82	1.57	1.61	1.46	1.61
Night	2.20	1.83	1.95	1.73	1.92
Late-night / Dawn	0.80	0.68	0.66	0.62	0.69
Morning / Noon	2.22	1.92	1.88	1.80	1.96
Afternoon / Evening	2.52	2.20	2.38	2.05	2.28
Percentage of heavy-duty vehicles (%)	18.91	22.50	22.27	21.16	21.51
Night	15.85	19.52	19.45	17.83	18.49
Late-night / Dawn	26.49	31.36	32.15	29.50	30.23
Morning / Noon	18.76	21.60	21.31	20.99	20.86
Afternoon / Evening	16.81	20.40	19.89	18.90	19.33
Tractor-trailers (1000 veh/hr)	0.19	0.16	0.16	0.17	0.17
Night	0.16	0.15	0.15	0.15	0.15
Late-night / Dawn	0.14	0.12	0.12	0.12	0.12
Morning / Noon	0.24	0.20	0.18	0.21	0.20
Afternoon / Evening	0.21	0.20	0.19	0.20	0.20

Note: The maximum mean is underlined and the minimum mean is in bold type.

Figure 3 further depicts the crash percentages under four time periods of segments in four clusters by Box-plotting. Fig. 3 shows a markedly different time-of-day distribution in the four clusters. Segments in metropolitan area (Cluster 1) have crashes frequently occur in the morning and noon (07-13) and afternoon and evening (14-19). There is a rising trend of crashes in rural areas (Cluster 2) from night (20-23) to afternoon and evening (14-19). Segments far from system interchanges (Cluster3) have crashes concentrated in the afternoon and evening (14-19) while the crash percentage of Cluster 4 concentrates in the morning and noon (07-13).



(a) Cluster 1 (b) Cluster 2

(c) Cluster 3 (d) Cluster 4

Fig. 3. Box-plots of crash percentages of segments in four clusters

The results obtained by the better performing crash count model are combined with the above clustering results. In the process of crash count model estimation, only the significantly tested explanatory variables (with $|t| > 1.645$) are retained. Table 4 gives the estimation results for the PO or NB models. The over-dispersion coefficient of the NB model is significantly different from zero, suggesting that the over-dispersion nature of the crash data and the NB model should be adopted. Meanwhile, the t values of explanatory variables in PO and NB models markedly differ, because the standard errors of the PO model are underestimated. In terms of goodness-of–fit indices, ρ^2 value and BIC value, the NB model outperforms the PO model. The log-likelihood ratio test ($\chi^2 = -2 \times (LL(\beta_{PO}) - LL(\beta_{NB}) = 352.50 > \chi^2_{(0.05,6)} = 12.59$) also confirms the better performance of the NB model.

Table 4. Estimated PO and NB models

Explanatory variables	Model			
	PO		NB	
	Para.	t-value	Para.	t-value
Constant	0.73	1.15	3.27	7.62
Over-dispersion	-	-	0.28	6.49
Geometric factors				
Maximum downward slope	-0.19	-7.61	-0.19	-3.96
Curvature	0.16	6.60	0.11	2.08
Clothoid parameter	-0.18	-4.99	-0.12	-1.72
Posted speed limit	1.34	3.43	-	-
Environmental factors				
Number of speeding cameras	-0.05	-4.32	-0.04	-1.87
Annual rainfall	0.27	4.50	-	-
Presence of rest area	0.41	4.32	-	-

Presence of toll station	-0.34	-4.06	-	-
Adjacent to urban area	0.36	7.65	0.37	3.21
Traffic factors				
Small vehicles	0.24	2.05	0.24	2.73
Bus and trucks	1.20	2.38	-	-
exp (Tractors-trailer)	0.48	3.04	-	-
Percentage of large vehicles	-0.04	-3.45	-0.04	-2.50
Goodness of fit				
Sample size		124		124
$LL(C)$		-1053.58		-1053.58
$LL(\beta)$		-603.49		-427.23
ρ^2		0.43		0.59
BIC		1274.454		897.85
Number of covariates		13		7

For the multivariate modeling approach, the MPO model directly uses the number of crashes in the four time periods in a day as the target variables. Table 5 shows the estimation results for the MPO model. Notably, the overall goodness-of-fit value of the MPO model is 0.39. The significantly tested explanatory variables in four time periods differ, suggesting the necessity to identify the key factors affecting crash counts of various time periods in a day. The estimated correlation coefficient indicates that the crash counts in the four time periods have a strong positive correlation (all of them are larger than 0.9), which strongly suggests that separately modeling the crash counts in various time periods is inappropriate.

According to the estimation results of the MPO model, an increased number of speeding cameras and increased volumes of buses decrease the number of crashes during night (20-23). In segments adjacent to urban areas, the number of crashes increases. During late-night/dawn (00-06), the presence of toll station and higher percentage of heavy-duty vehicles remarkably increase number of crashes. During morning and noon (07-13), the number of crashes increases in segments with larger difference in the slope and Clothoid parameter, in segments with a high number of speeding cameras, and in segments with toll stations. However, the number of crashes is even larger in the segments that have high annual rainfall, high total traffic and that are adjacent to an urban area. During afternoon/evening (14-19), higher maximum downward slope, Clothoid parameter, and percentage of heavy-duty vehicles obviously decrease number of crashes, while the segments adjacent to an urban, the number of crashes obviously increases.

In sum, geometric and environmental factors highly affect number of crashes occurring in the afternoon and evening, while traffic and environmental factors affect greater on the crashes in the nighttime. However, geometric, environmental, and traffic factors all contribute to crashes in the morning and noon.

Table 5. Estimated MPO model

Variables	Night		Late night / Dawn		Morning / Noon		Afternoon / Evening	
	Para.	t-Stat	Para.	t-Stat	Para.	t-Stat	Para.	t-Stat
Constant	0.71	2.30	0.57	1.49	0.23	0.60	2.39	7.91
Geometric factors								
Maximum downward slope	-	-	-	-	-	-	-0.10	-2.22
Maximum upward slope	-	-	-	-	-0.05	-2.17	-	-
Clothoid parameter	-	-	-	-	-0.16	-2.32	-0.16	-2.32
Environmental factors								
Number of speeding cameras	-0.07	-2.59	-	-	-0.09	-4.40	-	-
Annual rainfall	-	-	-	-	0.39	4.21	-	-
Presence of toll station	-	-	-0.29	-2.13	-0.29	-2.13	-	-
Presence of rest area	-	-	-	-	-	-	-0.43	-2.12
Adjacent to urban area	0.34	2.09	-	-	0.17	1.72	0.17	1.72

Traffic factors								
Total traffic	-	-	-	-	0.14	3.64	-	-
Buses and trucks	-4.50	-1.73	-	-	-	-	-	-
(Buses and trucks)^2	8.21	1.84	-	-	-	-	-	-
Small vehicles	-	-	0.56	2.36	-	-	-	-
Percentage of large vehicles	-	-	-0.02	-1.90	-	-	-0.04	-2.73
Error components								
δ_1、δ_2、δ_4、δ_7	0.61	7.47	0.21	2.71	0.47	7.70	0.73	11.44
δ_3、δ_5、δ_8	-	-	-0.02	-0.58	-0.03	-0.95	-0.05	-1.83
δ_6、δ_9	-	-	-	-	0.16	8.41	0.09	4.24
δ_{10}	-	-	-	-	-	-	-0.04	-2.08
Correlation coefficient								
Night		1.000						
Late-night / Dawn		0.995		1.000				
Morning / Noon		0.945		0.946		1.000		
Afternoon / Evening		0.989		0.991		0.978		1.000
Goodness of fit								
Sample size								124
$LL(C)$								-1646.67
$LL(\beta)$								-1001.80
ρ^2								0.39
BIC								2148.22
Number of covariates								16

To compare the estimation and validation performance of the proposed two approaches, Table 7 compares the predicted crash counts based on crash datasets for 2005 and 2006. As expected, the estimation performance of two approaches is slightly better than the validation performance in terms of two performance indices, *adj*-MAPE and RMSE. Additionally, the clustering approach performs better than the multivariate modeling approach. However, the difference in the performance of two approaches does not substantially differ, which suggests that both approaches can be used for modeling time-of-day crash frequencies.

Table 7. Performance indices of two approaches

Performance indices	Time periods				Total crashes
	Night	Late-night/Dawn	Morning /Noon	Afternoon /Evening	
Year of 2005 (estimation)					
Actual crashes	8.56	8.81	25.39	32.33	75.09
Predicted crashes					
Clustering approach	9.53	10.86	28.52	33.38	82.29
Multivariate approach	8.48	8.64	23.33	32.32	72.76
Adj-MAPE (%)					
Clustering approach	67.66	54.93	52.36	53.90	57.21
Multivariate	80.83	56.00	67.89	85.13	72.46
RMSE					
Clustering approach	7.36	10.00	20.37	24.80	17.20
Multivariate approach	11.91	5.42	23.77	50.72	28.76
Year of 2006 (validation)					
Actual crashes	7.54	7.36	25.98	31.33	72.21
Predicted crashes					
Clustering approach	8.58	8.91	29.62	34.27	81.38
Multivariate approach	8.43	8.57	18.55	32.54	68.09

Adj-MAPE (%)					
Clustering approach	68.39	57.50	56.51	59.80	60.55
Multivariate approach	88.05	58.97	72.38	81.35	75.19
RMSE					
Clustering approach	6.75	7.76	21.65	24.03	16.97
Multivariate approach	11.92	6.09	23.89	53.25	29.94

5. Conclusions

This study developed two novel approaches for modeling crash frequency under a time-of-day distribution and for identifying key factors that affect crash frequency during various time periods so as to propose more effective and time-specific countermeasures. The first approach is the clustering approach, which uses univariate count models to model total crash frequency and uses a clustering method to determine the time-of-day crash percentage distribution. The second approach models time-of-day crash frequencies by using multivariate count models. A case study on the crash data of Taiwan Freeway No. 1 in 2005 and 2006 is conducted. The results show that the clustering approach performs slightly better than the multivariate modeling approach, implying the way to separate time periods of day may have more significant impact than modeling methods.

According to the estimation results, crash counts are higher in the segments adjacent to metropolitan areas in the time periods of morning/noon (07-13) and afternoon/evening (14-19), while those in rural areas concentrate in the time periods of night (20-23) and afternoon/evening (14-19). Crash counts of segments far from system interchanges concentrate in the time period of the morning/noon. Additionally, geometric and environmental factors have higher effect on crashes in the afternoon and evening, while traffic factors have higher effect in the time period of the morning/noon. Additionally, geometric and environmental factors have the largest effect on crashes in the afternoon and evening while traffic factors have the largest effect in the morning/noon. Additionally, geometric and environmental factors have the largest effect on crashes in the afternoon and evening while traffic factors have the largest effect on crashes in the night and morning/noon. Geometric, environmental and traffic factors all affect crashes in the time periods of morning/noon. These findings show that the "peak and off-peak" of crash counts in various segments markedly differ. Corresponding countermeasures (such as freeway police routine patrol and set up the speeding camera) can be designed for the high risk road segment at specific time. Additionally, from the estimation of MPO model, the correlation among four time periods do exist, suggesting the necessity of simultaneous modeling.

Due to the limited availability of data and the high model complexity, four time periods are assumed according to the crash distribution to estimate the MPO model. Future studies can enhance the model performance and application by using detailed time periods such as 1 hour. However, under finer time period segmentation, many segments may have zero crash counts in many time periods; therefore, the zero-inflated MPO model should be adopted instead. This study used K-means method to classify segments into clusters under given number of clusters. More flexible clustering method with self-determining number of clusters, such as Genetic Clustering Algorithm proposed by Chiou and Lan (2001), can be adopted. Finally, most of the explanatory variables considered in this study (e.g. geometric and environmental factors) are not time variant. Finally, the traffic factors that could be the major sources depicting the time variation in four time periods, the following study can further incorporate other time variant explanatory factors such as rainfall and work zone, so as to improve the model applicability on crash modeling.

References

Abdel-Aty, M.A., Radwan, A.E., 2000. Modeling traffic accident occurrence and involvement. Accident Analysis and Prevention 32, 633-642.

Aguero-Valverde J., 2013. Multivariate spatial models of excess crash frequency at area level: Case of Costa Rica. Accident Analysis and Prevention 59, 365-373.

Anderson, T.K., 2009. Kernel density estimation and K-means clustering to profile road accident hotspots. Accident Analysis and Prevention 41, 359-364.

Chiou, Y.C., Fu, C., 2013. Modeling crash frequency and severity using multinomial- generalized Poisson model with error components. Accident Analysis and Prevention 50, 73-82.

Chiou, Y.C., Fu, C., Hsieh C.W., 2014. Incorporating spatial dependence in simultaneously modeling crash frequency and severity. Analytic Methods in Accident Research 2, 1-11.

Chiou, Y.C., Lan, L.W., 2001. Genetic clustering algorithms. European Journal of Operational Research 135, 413-427.
Clarke, D.D., P., Bartle W.C., Truman, W., 2006. Young driver accidents in the UK: The influence of age, experience, and time of day. Accident Analysis and Prevention 38, 871-878.
Doherty, S.T., Andrey J.C., MacGregor C., 1998. The situational risks of young drivers: The influence of passengers, time of day and day of week on accident rates. Accident Analysis and Prevention 30, 45-52.
Fridstrom, L., Ifver, J., Ingebrigtsen, S., Kulmala, R., Thomsen, L.K., 1995. Measuring the contribution of randomness, exposure, weather, and daylight to the variation in road accident counts. Accident Analysis and Prevention 27, 1-20.
Ivan, J.N., Pasupathy, R.K., Ossenbruggen, P.J., 1999. Differences in causality factors for single and multi-vehicle crashes on two-lane roads. Accident Analysis and Prevention 31, 695-704.
Ivan, J.N., Wang, C., Bernardo, N.R., 2000. Explaining two-lane highway crash rates using land use and hourly exposure. Accident Analysis and Prevention 32, 787-795.
Jones, B., Janssen, L., Mannering, F., 1991. Analysis of the frequency and duration of freeway accidents in Seattle. Accident Analysis and Prevention 23, 239-255.
Karlaftis, M.G., Tarko, A.P., 1998. Heterogeneity considerations in accident modeling. Accident Analysis & Prevention 30, 425-433.
Khattak, A., Pawlovich, M., Souleyrette, R.S., Hallmark, S., 2002. Factors related to more severe older driver traffic crash injuries. Journal of Transportation Engineering 128, 243-249.
Lord, D., 2006. Modeling motor vehicle crashes using Poisson-gamma models: Examining the effects of low sample mean values and small sample size on the estimation of the fixed dispersion parameter. Accident Analysis and Prevention 38, 751-766.
Malyshkina, N., Mannering, F., 2010. Empirical assessment of the impact of highway design exceptions on the frequency and severity of vehicle accidents. Accident Analysis and Prevention 42, 131-139.
Marquis, R., 2014. Investigating Temporal Effects on Truck Accident Occurrences in Manhattan. Department of Civil and Environmental Engineering, Rensselaer Polytechnic Institute, New York, USA.
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, 471-482.
Milton, J., Mannering, F., 1998. The relationship among highway geometrics, traffic related elements and motor vehicle accident frequencies. Transportation 25, 395-413.
Mohamed, M.G., Saunier, N., Miranda-Moreno, L.F., Ukkusuri, S.V., 2013. A clustering regression approach: A comprehensive injury severity analysis of pedestrian–vehicle crashes in New York, US and Montreal, Canada. Safety Science 54, 27-37.
Naderan, A., Shahi, J., 2010. Aggregate crash prediction models: Introducing crash generation concept. Accident Analysis and Prevention 42, 339-346.
Poch, M., Mannering, F., 1996. Negative binomial analysis of intersection. Journal of Transportation Engineering 12, 105-113.
Qin, X., Ivan, J.N., Ravishanker, N., Liu J., Tepas D., 2006. Bayesian estimation of hourly exposure functions by crash type and time of day. Accident Analysis and Prevention 38, 1071-1080.
Shankar, V., Mannering, F., Barfield, W. (1995) Effect of roadway geometrics and environmental factors on rural accident frequencies. Accident Analysis and Prevention 27(3), 371–389.
Shankar, V., Milton, J., Mannering, F. (1997) Modeling accident frequencies as zero-altered probability processes: an empirical study. Accident Analysis and Prevention 29(6), 829–837.
Venkataraman N., Ulfarsson G., Shankar V., 2014. Random parameter models of interstate crash frequencies by severity, number of vehicles involved, collision and location type. Accident Analysis and Prevention 59, 309-318.
Wang, Y., Nihan, N., 2004. Estimating the risk of collisions between bicycles and motor vehicles at signalized intersections. Accident Analysis and Prevention 36, 313-321.
Wong, S., Sze, N., Li, Y., 2007. Contributory factors to traffic crashes at signalized intersections in Hong Kong. Accident Analysis and Prevention 39, 1107-1113.
Ye, X., Pendyala, R.M., Washington, S.P., Konduri, K., Oh, J., 2009. A simultaneous equations model of crash frequency by collision type for rural intersections. Safety Science 47, 443-452.
Ye, X., Pendyala, R.M., Shankar, V., Konduri, K.C., 2013. A simultaneous equations model of crash frequency by severity level for freeway sections. Accident Analysis and Prevention 57, 140-149.