TRB 15-0521

Multi-level Hot Zone Identification for Pedestrian Safety

Jaeyoung Lee*
Mohamed Abdel-Aty

* Corresponding Author

[1]Department of Civil, Environmental and Construction Engineering
University of Central Florida
Orlando, Florida 32816-2450
(407) 823-0300
jaeyoung@knights.ucf.edu

Keechoo Choi

Department of Transportation Systems Engineering
Ajou University
Suwon, 443-749, South Korea

Helai Huang

Urban Transport Research Center
School of Traffic and Transportation Engineering
Central South University
Changsha, Hunan, 410075 P.R. China

August 2014

Prepared for Possible Presentation at the 94th Transportation Research Board Annual Meeting

Word count: 4,735 words + 5 tables + 6 figures = 7,485 equivalent words

ABSTRACT

According to the National Highway Traffic Safety Administration (NHTSA), while fatalities from traffic crashes have decreased, the proportion of pedestrian fatalities has steadily increased from 11% to 14% over the past decade. This study aims at identifying two zonal levels factors. The first is to identify hot zones at which pedestrian crashes occurs, while the second are zones where crash-involved pedestrians came from. Bayesian Poisson Lognormal Simultaneous Equation Spatial Error Model (BPLSESEM) was estimated and revealed significant factors for the two target variables. Then, PSIs (Potential for Safety Improvements) were computed using the model. Subsequently, a novel hot zone identification method was suggested to combine both hot zones from where vulnerable pedestrians originated with hot zones where many pedestrian crashes occur. For the former zones, targeted safety education and awareness campaigns can be provided as countermeasures whereas area-wide engineering treatments and enforcement may be effective safety treatments for the latter ones. Thus, it is expected that practitioners are able to suggest appropriate safety treatments for pedestrian crashes using the method and results from this study.

Key words: pedestrian safety, Bayesian approach, spatial error modeling, CAR, screening, macroscopic analysis, Poisson lognormal model, simultaneous equations modeling

INTRODUCTION

According to the National Highway Traffic Safety Administration (NHTSA), both fatalities and fatality rates from road traffic crashes in the United States have steadily declined from 2006 to 2011. Conversely, fatalities resulting from traffic crashes slightly increased in 2012 (*1*). Totally 33,561 lives were lost due to road traffic crashes in 2012. Among these fatalities, the proportion of pedestrian has steadily increased from 11% to 14% over the past decade (*2*). It shows the reason why we must keep focusing on the pedestrian crash issues. There are two perspectives to analyze traffic safety. The first perspective, microscopic safety analysis focuses on specific roadway entities including segments, intersections, corridors and so forth. The microscopic safety analysis aims to find out factors affecting traffic safety risk from geometric design and/or traffic characteristics of the roadway entities, and suggest specific engineering solutions to alleviate this risk. On the other hand, macroscopic safety analysis concentrates on zonal-level traffic safety with zonal characteristics. The macroscopic safety analysis can provide a broad spectrum perspective, and it suggests policy-based countermeasures including enactments of traffic rules, police enforcements, education/safety campaigns, and area-wide engineering treatments. In this study, the multi-level pedestrian safety was explored at the macroscopic level with the objective of providing guidance to policy decision makers to effectively improve pedestrian safety.

Pedestrian crashes have been considered a serious issue and many researchers have conducted pedestrian crash analysis at the macroscopic level (*3-14*). LaScala et al. (*3*) examined pedestrian injury rates across 149 census tracts in the city of San Francisco. Authors found out that the pedestrian injury rates were associated with traffic flow, population density, age composition of the local population, unemployment, gender and education. Ng et al. (*4*) revealed that the number of cinema seats, commercial area, flatted factory area, market stall, and MTR catchment area were positively related to the pedestrian crashes. Meanwhile, the greenbelt area, specialized factory area, and school places had negative relationships with pedestrian crashes in Hong Kong.

Noland and Quddus (*5*) developed two pedestrian crash models for severe crashes and minor injury crashes. The authors figured out that the percentage of local roads, income, and the percentage of people aged 45-64 decreased the severe pedestrian crashes, whereas the total population was negatively related with the severe pedestrian crashes. In regards to the minor injury crashes of pedestrians, more persons waiting for hospital treatment, higher percentage of trunk road, higher income, and the percentage of population aged 45-64 had positive associations with minor injury pedestrian crashes. On the other hand, the percentage of motorways, and the trunk road density were negatively associated with pedestrian related minor injury crashes. Loukaitou-sideris et al. (*6*) explored the pedestrian collisions based on census tracts in the city of Los Angeles. They found out that pedestrian collisions are more likely to occur in neighborhoods with high population and employment density, high traffic volumes, and a large concentration of commercial/retail and multifamily residential land uses. Moreover, zones with high concentration of Latino population had a higher chance to have more pedestrian crashes per capita. Wier et al. (*7*) investigated pedestrian crashes using 176 census tracts of San Francisco. The authors showed that the traffic volume, arterials without transit, the proportion of land area zoned for commercial and residential uses, employee and resident populations, and the proportion of people living in poverty, were found significant and positively affecting pedestrian crashes. In contrast, total land area (mi^2) and the proportion of population aged 65 or over had negative signs in the bicycle crash model.

Furthermore, Cotrill and Thakuriah (*8*) analyzed the pedestrian crashes in deprived areas with many low-income and minority populations. The authors corrected the underreporting problem using a Poisson model, and found that the exposure including the suitability of the area for walking and transit accessibility, crime rates, transit availability, income, and presence of children to be significant for pedestrian crashes. Ukkusuri et al. (*9*) used census tracts of New York City and discovered several socioeconomic and environmental factors for the frequency of pedestrian crashes using the NB (Negative Binomial) with random parameters model. Siddiqui et al. (*10*) found that the roadway length with 35 mph, intersections, dwelling units, population density, the percentage of households with 0 or 1 vehicle, long term parking cost, and total employment had positive relationship with the number of pedestrian crashes, whereas income reduced pedestrian crashes, from their Bayesian Poisson-lognormal model with a spatial error component. Moreover, Siddiqui and Abdel-Aty (*11*) estimated pedestrian crash models for zonal interior and boundary crashes, separately. They pointed out that the models could capture several unique explanatory variables explicitly related to interior and boundary crashes. For instance, total roadway length with 35 mph speed limit and long term parking cost were not significant in the interior pedestrian crash model but they were significant in the boundary model. It was also found that hotel units were positively associated with interior crashes whereas it had a negative sign in the boundary crash model.

Recently, Wang and Kockelman (*12*) studied the relationship between pedestrian crash frequency and land use, network and demographic attributes at the census tract level. They revealed that the higher shares of residences near transit stops are associated with pedestrian crash risks. In addition, the provision of sidewalk is associated with lower pedestrian crash rates. Abdel-Aty et al. (*13*) compared the pedestrian crash models based on different spatial units as census tracts, block groups and traffic analysis zones. It was found that VMT (Vehicle-Miles-Traveled) and the number of intersections, the number of workers commuting by public transportation, the workers commuting by walking and the proportion of minority population were significant in all models Moreover, roadways with relatively lower speed limits were positively associated with the pedestrian crashes in block group/traffic analysis zone based models. Furthermore, the roadways with high speed limit (65mph) variable was significant and negatively associated with pedestrian crashes solely in the census tract based model. In addition, the population of children aged from 0 to 15 was negatively related to pedestrian crashes in the block group/traffic analysis zone based models whereas the density of children (K to 12th grade) was positively associated with pedestrian crashes only in the traffic analysis zone based model. Workers with commute time 15 to 19 minutes was significant only in the traffic analysis zone based model. Furthermore, the number of home workers had a negative relationship with pedestrian crashes for the block group/traffic analysis zone based models. Lee et al. (*14*) investigated the residential characteristics of pedestrians who are involved in the traffic crash using the NB model formulation. The authors discovered that people living in the ZIP code area with lower median age, larger number of Hispanic population, more workers commuting by public transportation, shorter travel time to work, lower income, older buildings, smaller number of workers in the primary industry field were more likely to be involved in pedestrian crashes.

Although many pedestrian crashes have been analyzed at the macroscopic level, there are still unanswered research questions: (1) What are the contributing factors for the number of pedestrians per crash location zone and factors per residence zone? (2) Do the two targets have common unobserved factors and spatial autocorrelations? (3) How are the hot zones for both targets spatially distributed? (4) Is there a way to analyze the hot zones for the two targets simultaneously?

Therefore, the main objectives of this study are to identify the contributing factors both for ‘Pedestrian crashes per crash location ZIP code area’ and ‘Crash-involved pedestrians per residence ZIP’ and compare them. It was hypothesized that these two targets have commonly shared factors but have unique factors as shown in Figure 1. The common factors may include unobserved shared factors and spatial autocorrelation across the two targets. Subsequently, PSI (Potential for Safety Improvement) was selected for the screening performance measure and calculated for the two targets. Then, hot zones were identified for both targets separately and then integrated. It is expected that this integrated screening method can provide a more comprehensive perspective by location and residence zonal factors for pedestrian safety.

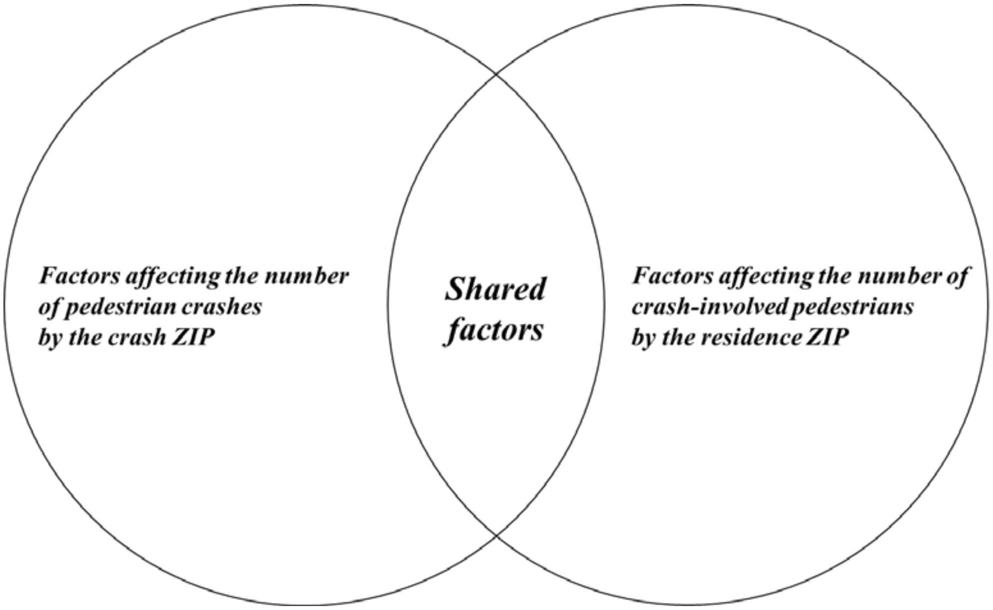

Factors affecting the number of pedestrian crashes by the crash ZIP
Shared factors
Factors affecting the number of crash-involved pedestrians by the residence ZIP

Figure 1 Factors affecting the two targets

DATA PREPARATION

Data from 983 ZIP areas in Florida were used for the analysis. Pedestrian crashes occurring between 2009 and 2011 were collected from Florida Department of Transportation (FDOT). Demographic, commute pattern, and socio-economic data were obtained from the U.S. Census Bureau and the roadway/traffic data were acquired from FDOT Roadway Characteristics Inventory. Lastly, the facility/attraction data were obtained from FDOT Unified Basemap Repository. Overall 40 candidate explanatory variables and 2 target variables were processed. The prepared data are summarized in Table 1.

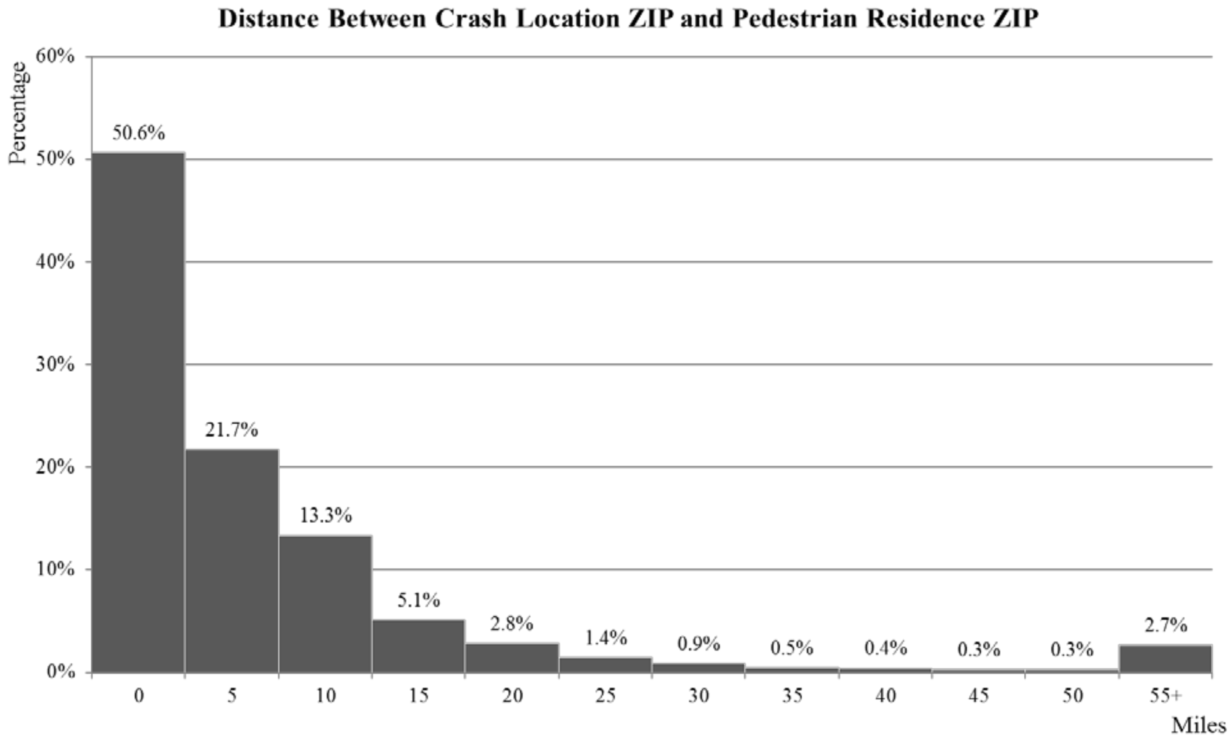
Table 1 Descriptive statistics of the prepared data

Category	Variable	Mean	Stdev	Min	Max
Target	Pedestrian crashes per crash location ZIP	17.411	22.849	0	227
	Crash-involved pedestrians per residence ZIP	23.779	26.843	0	224
Demographic	Population	19126.5	14561.8	0	72257
	Proportion of children (5-14 years)	0.112	0.037	0	0.222
	Proportion of adolescents (15-19 years)	0.065	0.054	0	0.974
	Proportion of young people (20-24 years)	0.063	0.049	0	0.643
	Proportion of elderly people (65-74 years)	0.189	0.116	0	1
	Proportion of very elderly people (75 years or older)	0.085	0.063	0	0.788
	Proportion of African Americans	0.138	0.163	0	0.970
	Proportion of Hispanics	0.169	0.189	0	1.000
Socioeconomic	Proportion of workers in the tertiary sector	0.780	0.138	0	1
	Proportion of households without available vehicle	0.027	0.034	0	0.462
	Proportion of households below poverty level	0.150	0.105	0	1
	Proportion of unemployed people	0.102	0.053	0	0.545
	Proportion of households below poverty level	0.169	0.189	0	1
	Median household income (in $1,000)	50.023	18.796	9.979	250
	Whether median year of structure built is before 1984 (yes=1, no=0)	0.507	0.500	0	1
Commute	Proportion of commuters using public transportation	0.017	0.046	0	1
	Proportion of commuter using non-motorized modes	0.026	0.052	0	1
	Proportion of people working at home	0.054	0.070	0	1
	Proportion of workers whose commute time is 15 min or shorter	0.263	0.143	0	1
	Proportion of workers whose commute time is 45 min or longer	0.157	0.108	0	1
Roadway/traffic	VMT	391498	334027	0	2426838
	Proportion of trucks	0.080	0.051	0	0.405
	Proportion of low-speed roads (speed limit: 35 mph or lower)	0.227	0.253	0	1
	Proportion of medium-speed roads (speed limit: 40-45 mph)	0.402	0.268	0	1
	Proportion of high-speed roads (speed limit: 55 mph or higher)	0.350	0.311	0	1
	Proportion of roads with poor pavement condition	0.003	0.016	0	0.220
	Number of traffic signals per miles	0.542	0.801	0	8.903
	Number of intersections per miles	10.732	29.699	0	908.265
Facility/attraction	Number of retail stores (grocery, home improvement, pharmacy, etc.) per mi^2	4.936	9.448	0	165.540
	Number of restaurants per mi^2	4.270	11.830	0	284.136
	Number of banks per mi^2	0.840	4.297	0	128.479
	Number of hotels, motels, and guest houses per mi^2	0.791	3.061	0	54.2714
	Number of K-12 schools per mi^2	0.610	1.078	0	20.342
	Number of gas stations per mi^2	0.553	0.726	0	4.726
	Number of parks and recreation areas per mi^2	0.375	0.777	0	6.510
	Number of department stores and shopping malls per mi^2	0.325	0.738	0	9.883
	Number of tourist attractions per mi^2	0.310	0.950	0	19.766
	Number of colleges and universities per mi^2	0.072	0.362	0	7.412
	Number of marinas/ferry terminals per mi^2	0.066	0.256	0	4.536
	Number of hospitals per mi^2	0.046	0.176	0	3.506

DISTANCE ANALYSIS

In this study, both ‘Pedestrian crashes per crash location ZIP’ and ‘Crash-involved pedestrians per residence ZIP’ were analyzed simultaneously. Some may argue that most of pedestrians are involved in traffic crashes at the same or very close to their residence ZIP area, and it is not necessary to separate these two targets.

In order to justify the objectives of this study, the distance between pedestrian crash location and pedestrians’ residence were explored. Both crash location ZIP and pedestrians’ residence ZIP information were collected from each pedestrian crash and the coordinate information of the ZIP centroids were obtained using GIS. After that, the distance between crash location ZIP and crash-involved pedestrians’ residence ZIP was calculated for each pedestrian crash. Figure 2 exhibits the distribution of these distances. It was shown that actually the distance between pedestrian crashes and their residence zones are quite close. About 90% of pedestrian crashes occur within 14 miles within the pedestrians’ residence zone and approximately 50% of pedestrian crashes happened in the pedestrians’ residence zones; however still other 50% occur in zones other than their residence. Thus, it can be justified to separately explore ‘Pedestrian crashes per crash location ZIP’ and ‘Crash-involved pedestrians per residence ZIP’.



Basic statistics			Quantiles					
Mean	Stdev	zero distance %	100% (max)	95%	90%	75%	50% (median)	50-0%
7.610	26.813	50.6%	533.799	24.798	13.841	5.796	0	0

Figure 2 Descriptive statistics of distance between crash location ZIP and pedestrian residence ZIP

STATISTICAL MODELING

Bayesian Poisson Lognormal Simultaneous Equations Spatial Error Model (BPLSESEM) was adopted in this study. Different from the classical models, Bayesian models do not depend on the assumption of asymptotic normality. Sampling based methods of Bayesian estimation focus on estimating the entire density of parameters as compared to the traditional classical estimation methods which are intended for finding a single point estimate using the maximum likelihood approach (*15*). Of course, sometimes point estimates may be more convenient for the practical application since it clearly suggest a single point. However, the Bayesian approach has a significant advantage over the maximum likelihood estimation. The Bayesian estimation determines posterior density for each parameter under consideration. This density estimation is the outcome of a process where a long run or a series of long runs of samples are taken from the posterior density based on the prior information about the parameter and data. Accordingly a Bayesian approach provides a considerable interpretive advantage since posterior estimates reflect the probabilities that the analyst is primarily interested in, the probability of the null hypothesis being true called a Bayesian Credible Interval (BCI). On the other hand, classical confidence intervals on parameter estimates provide the probability of observing data given that a parameter takes on a specific value. This distinction of the Bayesian approach provides a substantial philosophical and practical advantage (*16*). Therefore, the Bayesian approach is thought to be more suitable compared to the classical likelihood based inference methods and thus have been popular in recent traffic safety research.

The Poisson regression models have played a key role in analyzing crash frequency data. The Poisson regression model has been broadly used by many researchers since it can cope with non-negative integers. This study also adopted the Poisson regression based model because Poisson distribution approximates rare event frequency data such as the number of pedestrian crashes or the number of crash-involved pedestrians, which were used as the response variables in this study.

The probability of ZIP *i* having y_i the number of pedestrian crashes aggregated based on their crash location zone (or the number of crash-involved pedestrians aggregated based on their residence zone) per time period is given by:

$$\mathrm{P}(y_i) = {exp(-\lambda_i){\lambda_i}^{y_i}}\Big/{y_i!} \qquad (1)$$

where $\mathrm{P}(y_i)$ is the probability of entity *i* having y_i pedestrian crashes (or crash-involved pedestrians) per given time period, and λ_i is the Poisson parameter, which shows the expected number of pedestrian crashes (or crash-involved pedestrians) per period,

Poisson regression models are estimated by specifying λ_i (Poisson parameter) as a function of explanatory variables, the most widely used functional form is:

$$\lambda_i = exp(\beta X_i) \qquad (2)$$

where X_i is a row vector of explanatory variables of entity i, and β is a coefficient estimate of model covariates βX_i.

Nevertheless, the Poisson models cannot manage both over- and under-dispersion in the data since it assumes mean and variance are equal. Hence, the Poisson lognormal model was suggested as one of the alternative models to Poisson models to account for the over-dispersion of crash data (*17*). Furthermore, simultaneous equations and random parameters shared by two equations for the two targets were used to account for unobserved factors between the two targets (*18*). The expected number of pedestrian crashes (or crash-involved pedestrians) is formulated as follows:

$$\lambda_{i1} = exp(\beta_1 X_{i1} + \theta_{i1} + \varphi_i) = exp(\beta_1 X_{i1} + \delta_1 u_{i1} + \varphi_i) \quad (3)$$

$$\lambda_{i2} = exp(\beta_2 X_{i2} + \theta_{i2} + \varphi_i) = exp(\beta_2 X_{i2} + \delta_2 u_{i1} + \delta_3 u_{i2} + \varphi_i) \quad (4)$$

where, λ_{ik} is the expected number of pedestrian crashes per crash location ZIP i (*k*=1) or the expected number of crash-involved crashes per residence ZIP i (*k*=2), X_{ik} is a row vector of explanatory variables showing characteristics of ZIP *i*, for target k, β_k is a coefficient estimate of model covariates X_{ik} , θ_{ik} is a random error term representing normal heterogeneity of ZIP *i*, for target k, u_{ik} follows standard normal distribution (0, τ_θ) for ZIP *i* and target k, τ_θ is the precision parameter that is the inverse of the variance; it follows prior gamma (0.5, 0.005), δ_1 is the coefficient for u_{i1}in Equation (3), while δ_2 and δ_3 are the coefficients for u_{i1} and u_{i2} in Equation (4), respectively, and φ_i is a shared spatial autocorrelation error term.

The model was run considering a non-informative normal (0,10000) prior for both δ_m and β_k. In the case of the univariate model structure, δ_2, δ_4 and δ_5 are set to zero because the univariate model do not account for correlations between heterogeneities of crashes by different modes.

Spatial autocorrelation is a technical term for the fact that spatial data from near sites are more likely to be similar than data from distant sites (*19*). The existence of the spatial autocorrelation in the crash data may invalidate the assumption of the random distribution (*20*). In order to control for the spatial autocorrelation, a spatial error term (φ_i) was included in the model specification. Spatial distribution was implemented by specifying intrinsic Gaussian Conditional Autoregressive (CAR) prior with normal ($\bar{\varphi}_i$,τ_i^2) distribution as recommended by Besag (*21*).

Mean of φ_i is calculated as follows:

$$\bar{\varphi}_i = {\left(\sum_{i\neq j} \varphi_j \times w_{ij}\right)}\Big/{\left(\sum_{i\neq j} w_{ij}\right)} \quad (5)$$

where, w_{ij} is the element of adjacency matrix with a value of 1 if *i* and *j* are adjacent or 0 otherwise.

Among the benefits of the Bayesian approach are a more natural interpretation of parameter intervals, termed Bayesian Credible Interval (BCI) and the freedom of obtaining true parameter density (*15*). On the other hand, likelihood based estimates depend on normality approximations based on large sample asymptotics (*15*). In this study, only variables whose 90% BCI of the posterior parameter estimates showing the same sign were included in the final model.

Furthermore, DIC (Deviance Information Criterion) was computed. The following equation is used to calculate DIC (*22*). Models with smaller DIC are preferred to models with larger DIC. Therefore, the model with the smallest DIC was chosen as a final model.

$$DIC = 2 \times \bar{D} - \hat{D} \quad (6)$$

where $\overline{D}$: posterior mean of deviance, D, $\widehat{D} = 2 \times (p(y|\theta))$, and $\bar{\theta}$: posterior mean of θ, respectively.

Regarding the exposure variable, it was thought that both population and traffic volume can be used as exposure variables for the two targets. Initially, both ‘Log of population’ and ‘Log of VMT’ were attempted at the same time in the model; however, these two variables are highly correlated with each other (r = 0.720) and thus cannot be used simultaneously. We calculated the product of ‘Log of population’ and ‘Log of VMT’ and tried it as an exposure variable, since it was believed that this new variable can reflect both population and traffic volume at the same time. ‘Log of population’, ‘Log of VMT’, and the product of ‘Log of population’ and ‘Log of VMT’ were attempted one by one, and all of these variables were found to be significant at the 5% level. Nevertheless, it was uncovered that ‘the product of ‘Log of population’ and ‘Log of VMT’ was the best exposure variable for ‘Pedestrian crashes per crash location ZIP’, whereas ‘Log of population’ was the best exposure variable for ‘Crash-involved pedestrians per residence ZIP’ (Table 2). It implies that the number of pedestrian crashes is largely affected both by population and traffic volume because a pedestrian crash is a collision between pedestrian and motor vehicle. When it comes to the number of crash-involved pedestrian per their residence, the populations plays a more important role because traffic volume is not directly related to crash-involved pedestrians aggregated based on their residence whereas the population is a direct exposure measure for the pedestrian crashes. For example, consider a zone with many pedestrians who were involved in crashes but with low car-ownership. More than likely, the zone have lower traffic volume but they are likely to have large population exposed to traffic crashes.

Table 2 Selection of exposure variables for each target variable

Target	**Exposure Variable**	$\boldsymbol{\beta_0}$ **(intercept)**				$\boldsymbol{\beta_1}$ **(exposure variable)**				**DIC**
		mean	**s.d.**	**BCI**		**mean**	**s.d.**	**BCI**		
				2.5%	**97.5%**			**2.5%**	**97.5%**	
Pedestrian crash per crash location ZIP	Log of population	-7.374	0.254	-7.739	-6.833	0.999	0.026	0.943	1.036	5437.23
	Log of VMT	-3.671	0.192	-4.029	-4.003	0.049	0.002	0.046	0.051	5424.98
	(Log of population) × (Log of VMT)	**-7.693**	**0.373**	**-8.378**	**-7.063**	**0.786**	**0.029**	**0.736**	**0.840**	**5360.89**
Crash-involved pedestrians per residence ZIP	**Log of population**	**-6.836**	**0.242**	**-7.283**	**-6.387**	**0.989**	**0.025**	**0.943**	**1.035**	**5877.91**
	Log of VMT	-4.240	0.161	-4.572	-3.940	0.547	0.013	0.523	0.573	6046.88
	(Log of population) × (Log of VMT)	-2.895	0.151	-3.151	-2.553	0.046	0.001	0.043	0.048	5883.75

MODELING RESULTS

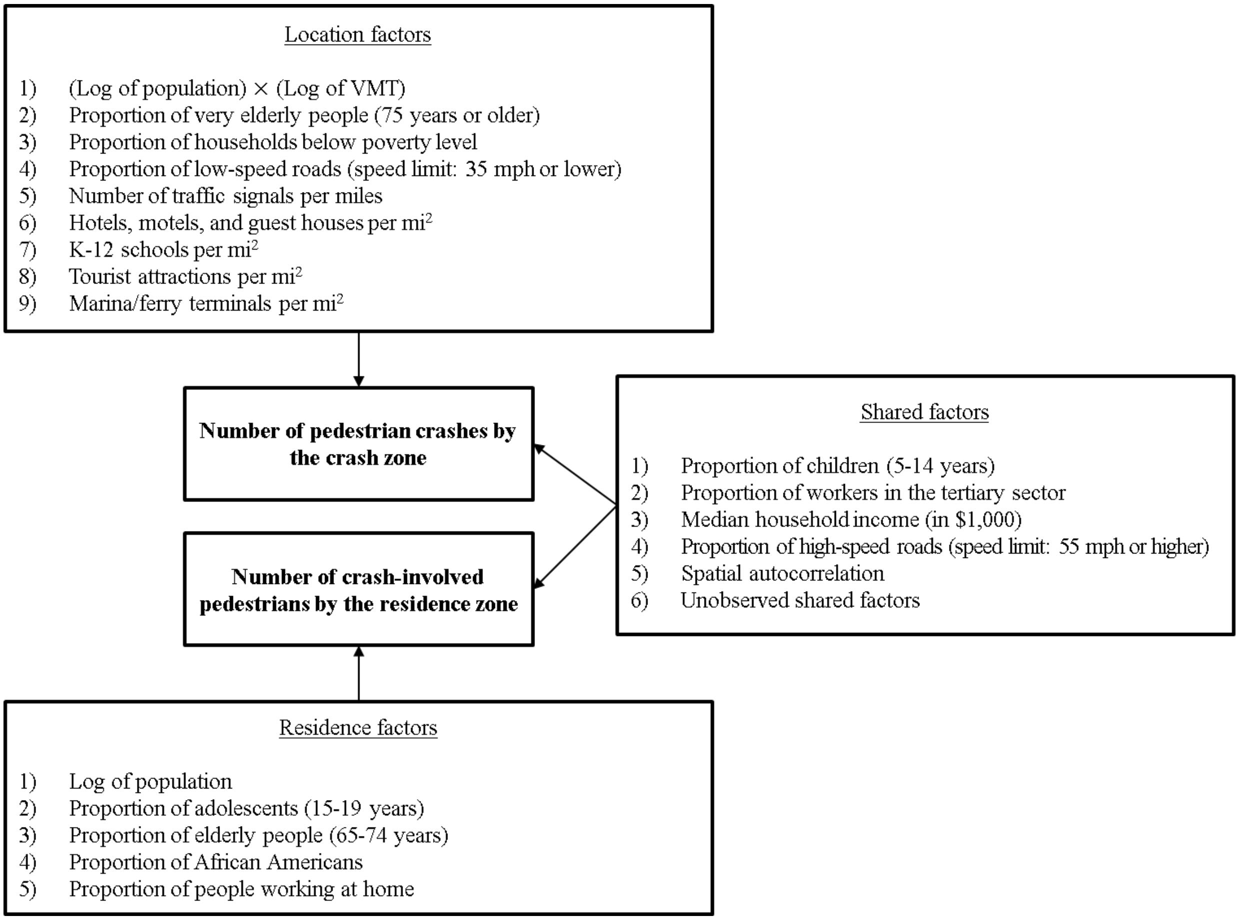
The modeling results are summarized in Table 3. It is shown that two target variables: ‘Pedestrian crashes per crash location ZIP’ and ‘Crash-involved pedestrians per residence ZIP’ have different significant variable sets. For the first target variable, ‘Pedestrian crashes per crash location ZIP’, overall 17 explanatory variables were significant at 5%, except for ‘Proportion of households below poverty level. Only this variable was significant at the 10% level. As stated earlier, the product of ‘Log of population’ and ‘Log of VMT’ were used as an exposure variable for the first target variable. As expected, the exposure variable is positively associated with the first target variable. There are two significant demographic variables. Both ‘Proportion of children (5-14 years)’ and ‘Proportion of elderly people (75 years or older)’ were negatively related to the first target variable. Also, it was revealed that 3 socioeconomic variables are significant. Both ‘Proportion of workers in the tertiary sector’ and ‘Median household income (in $1,000)’ have negative relationships whereas ‘Proportion of households below poverty level’ has a positive relationship with the first target variable. Moreover, 3 roadway variables were significant. ‘Proportion of low-speed roads (speed limits: 35 mph or lower)’ and ‘Number of traffic signals per mile’ are negatively related while ‘Proportion of high-speed road (speed limit: 55 mph or higher)’ is positively related to the first target variable. Furthermore, the first target has 4 facilities/attractions explanatory variables including ‘Number of hotels, motels, and guest houses per mi^2’, ‘Number of K-12 schools per mi^2’, ‘Number of tourist attractions per mi^2’, and ‘Number of marina/ferry terminals per mi^2’. All these facilities/attractions have positive effects on the first target variable.

Concerning the second target variable, ‘Crash-involved pedestrian per residence ZIP’, totally 9 explanatory variables were significant at the 5% level. The exposure variable of the second target is ‘Log of population’ as mentioned earlier and it positively influences the second target variable. It has 4 significant demographic variables. Age related factors such as ‘Proportion of children (5-14 years)’, ‘Proportion of adolescents (15-19 years)’, and ‘Proportion of elderly (65-74 years)’ are negatively related to the second target. On the other hand, a race related factor, ‘Proportion of African Americans’ has a positive association with the second target variable. In addition, there is a significant commute variable, ‘Proportion of people working at home’ which has a negative effect on the second target. Moreover, two socioeconomic factors were found to be significant. Both ‘Proportion of workers in the tertiary sector’ and ‘Median household income (in $1,000)’ lowers the probability to have crash-involved pedestrians. Lastly, it has only one significant roadway variable, ‘Proportion of high-speed roads (speed limit: 55 mph or higher)’ which is negatively associated with the second target variable.

With regards to the random parameters (δ_1, δ_2, and δ_3) that reflect the unobserved common components between ‘Pedestrian crashes per crash location ZIP’ and ‘Crash-involved pedestrians per residence ZIP’, δ_1 and δ_3 are significant at the 5% level and δ_2 is significant at 10%. It implies the existence of common factors between the two targets although they are unobserved. Furthermore, the standard deviation of the shared spatial error term, ‘s.d. of φ_i’ is statistically significant at 5%. It suggests that both targets are spatially correlated among adjacent zones and the spatial autocorrelation is controlled by the spatial error term included in the model. In summary, ‘Pedestrian crashes per crash location ZIP’ has more roadway and facility/attraction variables while ‘Crash-involved pedestrians per residence ZIP’ has more demographic variables. Nevertheless, the two targets have 6 common significant variables including ‘Proportion of children (5-14 years)’, ‘Proportion of workers in the tertiary sector’, ‘Median household income (in $1,000)’, ‘Proportion of high-speed roads (speed limit: 55 mph or higher)’, ‘Spatial autocorrelation’, and ‘Unobserved shared factors’ as shown in Figure 3.

Table 3 Bayesian Poisson Lognormal Simultaneous Equations Spatial Error Modeling Result

Variable	Pedestrian crashes per crash location ZIP						Crash-involved pedestrians per residence ZIP					
	mean	s.d.	Bayesian Credible Interval				mean	s.d.	Bayesian Credible Interval			
			2.5%	5.0%	95.0%	97.5%			2.5%	5.0%	95.0%	97.5%
Intercept	2.711	0.490	1.931	2.038	3.434	3.456	2.578	1.211	0.174	0.290	3.821	3.844
(Log of population) × (Log of VMT)	0.027	0.008	0.011	0.012	0.038	0.038						
Log of population							0.481	0.189	0.177	0.187	0.744	0.751
Proportion of children (5-14 years)	-8.129	1.660	-10.710	-10.610	-5.369	-4.904	-7.472	1.422	-9.562	-9.430	-4.550	-4.407
Proportion of adolescents (15-19 years)							-3.727	0.867	-5.363	-5.177	-2.236	-2.067
Proportion of elderly people (65-74 years)							-5.395	0.905	-7.172	-6.978	-4.051	-3.964
Proportion of very elderly people (75 years or older)	-3.675	0.588	-4.776	-4.655	-2.624	-2.438						
Proportion of African Americans							0.214	0.096	0.026	0.060	0.382	0.408
Proportion of people working at home							-1.583	0.618	-2.687	-2.559	-0.508	-0.369
Proportion of workers in the tertiary sector	-1.966	0.891	-3.273	-3.249	-0.358	-0.311	-2.399	1.120	-3.603	-3.574	-0.420	-0.395
Proportion of households below poverty level	0.471	0.260	-0.021	0.036	0.908	0.976						
Median household income (in $1,000)	-0.018	0.002	-0.022	-0.021	-0.016	-0.016	-0.015	0.002	-0.019	-0.019	-0.012	-0.011
Proportion of low-speed roads (speed limit: 35 mph or lower)	0.351	0.081	0.186	0.208	0.485	0.501						
Proportion of high-speed roads (speed limit: 55 mph or higher)	-0.981	0.127	-1.202	-1.171	-0.751	-0.713	-0.727	0.085	-0.906	-0.873	-0.591	-0.566
Number of traffic signals per miles	0.042	0.022	0.001	0.008	0.080	0.086						
Number of hotels, motels, and guest houses per mi^2	0.007	0.004	0.000	0.001	0.013	0.014						
Number of K-12 schools per mi^2	0.048	0.016	0.016	0.021	0.075	0.078						
Number of tourist attractions per mi^2	0.087	0.017	0.053	0.059	0.115	0.119						
Number of marina/ferry terminals per mi^2	0.156	0.067	0.012	0.038	0.262	0.281						
$\delta_1,\ \delta_2$	7.933	5.326	2.729	3.030	19.470	20.730	5.497	5.099	-0.016	0.176	14.970	17.940
δ_3							3.289	1.119	0.123	0.783	4.716	4.934
s.d. of φ_i	0.982	0.133	0.757	0.775	1.192	1.214	same					
DIC	10798.3											



Location factors

1) (Log of population) × (Log of VMT)
2) Proportion of very elderly people (75 years or older)
3) Proportion of households below poverty level
4) Proportion of low-speed roads (speed limit: 35 mph or lower)
5) Number of traffic signals per miles
6) Hotels, motels, and guest houses per mi^2
7) K-12 schools per mi^2
8) Tourist attractions per mi^2
9) Marina/ferry terminals per mi^2

Number of pedestrian crashes by the crash zone

Number of crash-involved pedestrians by the residence zone

Shared factors

1) Proportion of children (5-14 years)
2) Proportion of workers in the tertiary sector
3) Median household income (in $1,000)
4) Proportion of high-speed roads (speed limit: 55 mph or higher)
5) Spatial autocorrelation
6) Unobserved shared factors

Residence factors

1) Log of population
2) Proportion of adolescents (15-19 years)
3) Proportion of elderly people (65-74 years)
4) Proportion of African Americans
5) Proportion of people working at home

Figure 3 Summary of significant factors for the two targets

HOT ZONE IDENTIFICATION ANALYSIS

In order to identify hot zones, the performance measure should be determined. A variety of performance measures have been used in previous screening studies. They include crash frequency, EPDO (Equivalent Property Damage Only) crash frequency, crash rate, proportion by crash types, Empirical Bayes (EB), PSI, and so forth. In this study PSI was selected as the performance measure. PSI, or excess crash frequency, is defined as a performance measure indicating the number of pedestrian crashes aggregated based on their crash location zone (or the number of crash-involved pedestrians aggregated based on their residence zone) that could effectively be reduced for a particular zone in this study. The PSI for each zone is a difference between the expected number of pedestrian crashes (or crash-involved pedestrians) and the predicted number of pedestrian crashes (or crash-involved pedestrians). Therefore, this performance measure can effectively identify those zones experiencing more pedestrian crashes or having more crash-involved pedestrians than other zones with similar characteristics. Therefore, if a zone has PSI greater than zero, the zone is considered hazardous whereas a zone is regarded as safe if its PSI is smaller than zero.

The predicted number of pedestrian crashes (or crash-involved pedestrians) for each zone was calculated from the model (Table 3). PSIs were calculated by the following equations (*23*):

$$PSI = N_{expected} - N_{prdeicted} \tag{7}$$
$$= \exp(\beta_0 + \beta X_i + \theta_i + \varphi_i) - \exp(\beta_0 + \beta X_i) \tag{8}$$
$$= \exp(\beta_0 + \beta X_i)(\exp(\theta_i + \varphi_i) - 1) \tag{9}$$

All zones in the study area were classified into three categories based on their PSIs: Hot ('H'), Warm ('W'), and Cold ('C') zones. Hot zones are defined as zones with a top 10% PSI, warm zones refer to zones with a PSI between 0 and top 10%, and cold zones are those with PSI less than 0, as shown in Figure 4. Thus, 'H' have much more pedestrian crashes (or crash-involved pedestrians) compared to other zones with similar characteristics. 'W' zones also have some room for pedestrian crash reduction (or crash-involved pedestrians); however the pedestrian safety is not much risky as in 'H' zones. In case of cold zones, it has less pedestrian crashes (or crash-involved pedestrians) compared to other similar zones.

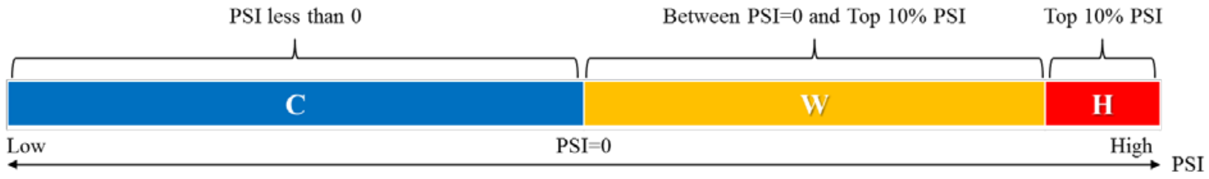


Figure 4 Definition of zonal screening categories

Table 4 exhibits a part of the screening results of 'Pedestrian crashes per crash location ZIP'. In case of ID 1, its PSI is 0.638 and ranked number 513 based on the PSI. Because it is in the top 52.2% PSI, this zone was categorized as 'W', which has a pedestrian safety problem in the location but it is not severe as much as in 'H'. ID 3 zone has a negative PSI, -9.369, and thus it is classified as 'C', which is relatively safe for pedestrian crashes. On the other hand, the PSI of ID 983 is 26.460, which is the top 6.9% PSI.

Therefore, it is classified as ‘H’, which shows that the zone has serious pedestrian safety problems, compared to other similar zones.

Table 4 Example of screening result: pedestrian crashes per crash location ZIP

ID	ZIP	PSI	Rank	Percentage	Category
1	32606	0.638	513	52.2%	W
2	32609	7.850	263	26.8%	W
3	32612	-9.639	975	99.2%	C
4	32234	-1.595	735	74.8%	C
5	32438	-2.445	823	83.7%	C
:	:	:	:	:	:
983	34668	26.460	68	6.9%	H

Figure 5 displays the screening result based on the PSI of pedestrian crashes per crash location ZIP (left figure) and that based on PSI of crash-involved pedestrians per residence ZIP (right figure). As shown in the figures, the locations of ‘H’ zones of the two targets are quite comparable but not exactly the same. The general trend of ‘H’ zones of the targets shows that most of them are concentrated in the urban area. In case of ‘W’ and ‘C’ zones, they also have similar spatial distributions between the two target variables.

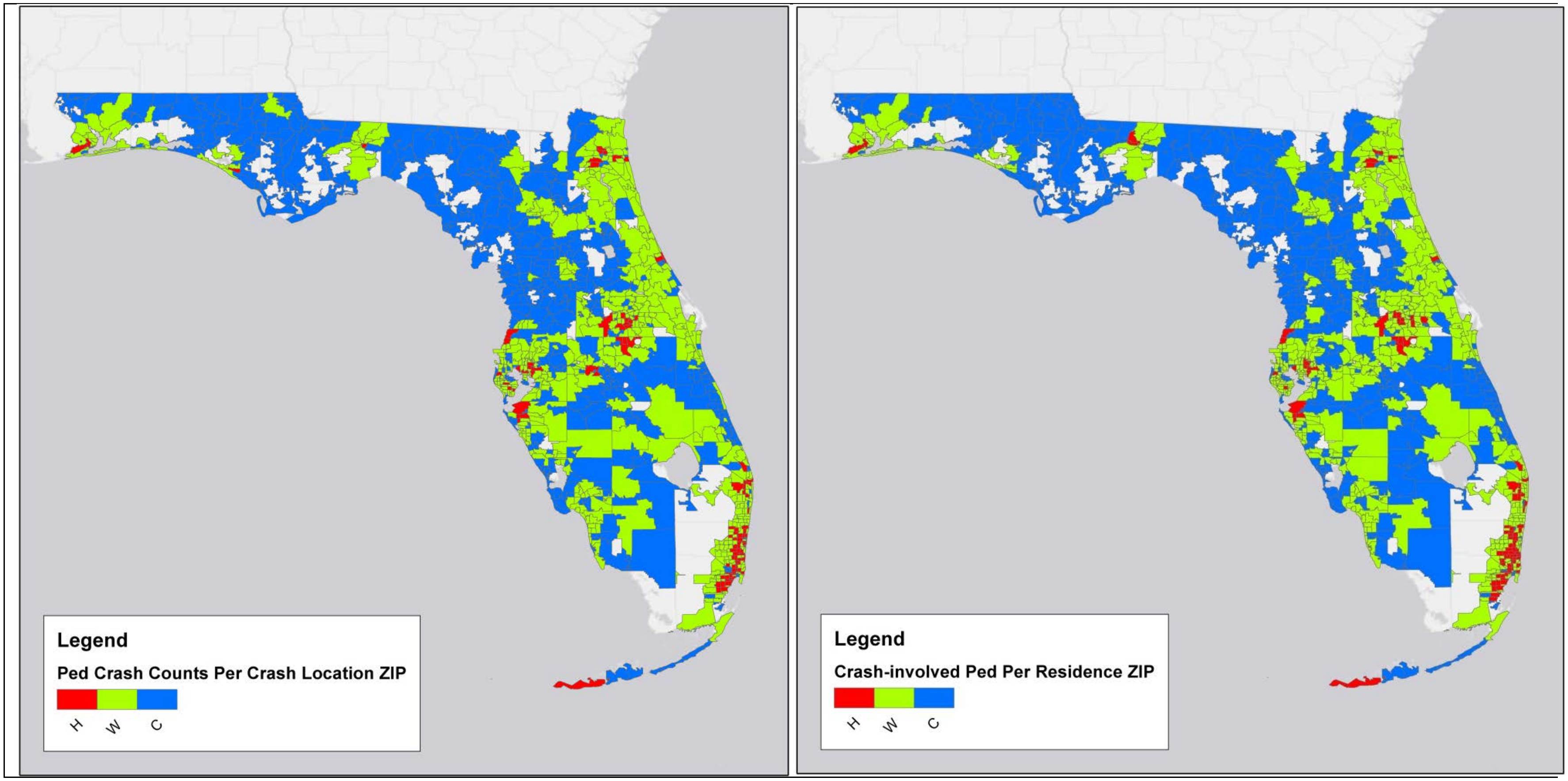

Legend
Ped Crash Counts Per Crash Location ZIP
H
W
C
Legend
Crash-involved Ped Per Residence ZIP
H
W
C

Figure 5 Screening results based on PSI of pedestrian crashes per crash location ZIP (left) and that based on PSI of crash-involved pedestrians per residence ZIP (right)

INTEGRATED SCREENING

In the preceding section, hot zones for two targets: ‘Pedestrian crashes per crash location ZIP’ and ‘Crash-involved pedestrians per residence ZIP’ are identified individually. In this section, the hot zone identification results of the two targets are combined to provide a broad spectrum perspective for both locations with higher risk for pedestrians and residences with many pedestrians vulnerable to crashes. All zones were again categorized according to the two scopes: location and residence, and 3 traffic safety levels: ‘H’, ‘W’, and ‘C’. Therefore, there are overall 9 combination classifications: ‘HH’, ‘HW’, ‘HC’, ‘WH’, ‘WW’, ‘WC’, ‘CH’, ‘CW’, and ‘CC’. The initial letter of the classifications represents the location-based pedestrian safety risk, and the latter character symbolizes the residence-based pedestrian safety risk. However, ‘HC’ and ‘CH’ cases, which have extremely different pedestrian safety levels between location and residence, are not observed in the screening result. The integrated screening result is summarized in Figure 6 and Table 5 with seven screening categories.

Overall, 76 ‘HH’ zones (7.7%) were identified, which is top priority for pedestrian safety treatments because they have serious pedestrian problems in their locations and also many pedestrians vulnerable to traffic crashes. Furthermore, there are 22 ‘HW’ (2.2%) zones and 22 ‘WH’ (2.2%) zones, the next highest priority for pedestrian treatments. ‘HW’ zones have very high risk for pedestrians in their zones; but pedestrians from the zones are not particularly exposed to crashes at other zones. Pedestrians in ‘WH’ zones are very vulnerable to traffic crashes whereas their locations are not exceptionally dangerous for pedestrians. Nearly half of zones (43.1%) are classified as ‘WW’ zones. ‘WW’ zones have moderate risks in their physical locations and their pedestrians vulnerable to crashes. There are 20 ‘WC’ zones (2.0%), which are with intermediate risk in their locations but their pedestrians are less likely to be involved in crashes. Twenty ‘CW’ zones (2.0%) have little pedestrian problems in their locations but their pedestrians are a little vulnerable to crashes. Lastly, in case of 399 ‘CC’ zones (40.6%), pedestrians are relatively safe for the two targets.

Engineering, Education, and Enforcement (3E) are traditional but still valid treatments to reduce traffic crashes effectively. We can implement different 3E treatment strategies for different screening categories. If zones have higher pedestrian crash risks (i.e. HH, HW, etc.), both area-wide engineering treatments and enforcement can effectively reduce pedestrian crashes in these zones. Education and/or safety campaigns for pedestrians may be a good way to reduce the number of crash-involved pedestrians aggregated based on their residence zone, if the residence has more crash-involved pedestrians (i.e. HH, WH, etc.). All 3E general countermeasures need to be implemented for ‘HH’ zones. Of course specific treatments would need to be tailored to the specific problem and location.

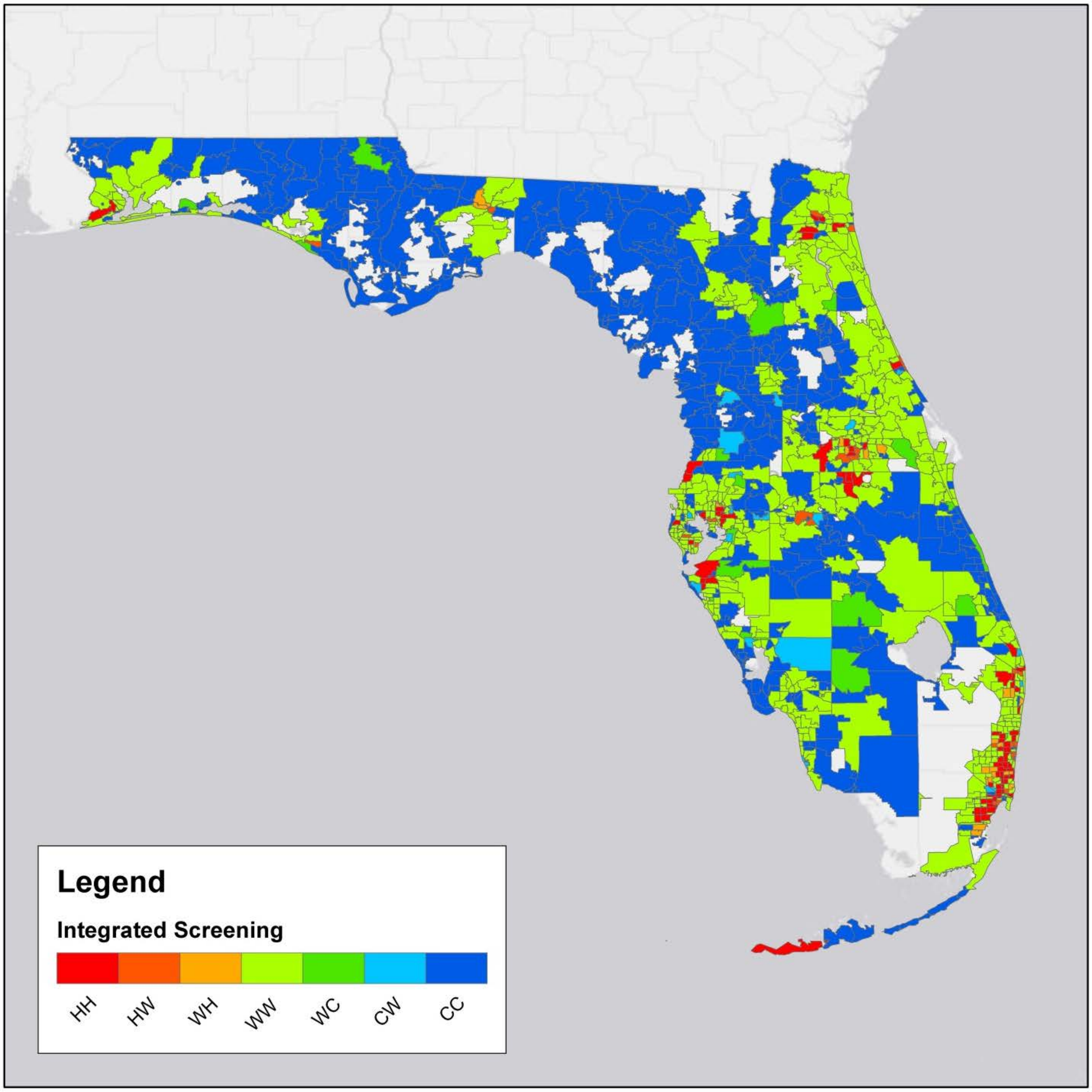


Figure 6 Integrated screening result

Table 5 Zones by integrated screening categories

Category	**HH**	**HW**	**WH**	**WW**	**WC**	**CW**	**CC**	**Sum**
Counts	76	22	22	424	20	20	399	983
Percentage	7.7%	2.2%	2.2%	43.1%	2.0%	2.0%	40.6%	100.0%

SUMMARY AND CONCLUSIONS

In this study, two targets: 'Pedestrian crashes per crash location ZIP' and 'Crash-involved pedestrians per residence ZIP' were comprehensively analyzed. In the preliminary analysis, it was shown that pedestrian crashes do not necessarily occur in the pedestrians' residence zone. Of course, the distance between pedestrian crashes and their residence zones are close. About 90% of pedestrian crashes occur within 14 miles within the pedestrians' residence zone. Approximately 50% of pedestrian crashes happened in the pedestrians' residence zones; however the other 50% occurred in zones other than their residence. Thus, it can be justified to separately explore 'Pedestrian crashes per crash location ZIP' and 'Crash-involved pedestrians per residence ZIP'.

Different exposure variables were applied for the two targets. 'Log of population', 'Log of VMT', and the product of 'Log of population' and 'Log of VMT' were attempted one by one, and all of these variables were found to be significant at the 5% level. Nevertheless, it was uncovered that 'the product of 'Log of population' and 'Log of VMT' was the best exposure variable for 'Pedestrian crashes per crash location ZIP', whereas 'Log of population' was the best exposure variable for 'Crash-involved pedestrians per residence ZIP'. BPLSESEM (Bayesian Poisson Lognormal Simultaneous Equations Spatial Error Model) was adopted in this study to account for unobserved common factors between the two targets and spatial autocorrelation among adjacent zones. The BPLSESEM revealed that two targets have different contributing factors. The first target, 'Pedestrian crashes per crash location ZIP', has 17 significant factors whereas the second target, 'Crash-involved pedestrians per residence ZIP' has 9 significant factors. It was shown that the first target has more variables related to location factors such as roadway and facility factors. In contrast, the second target is associated with more demographic factors. Also, It was found that 4 significant factors are commonly significant for the two targets. It is probable that there are common factors between two target variables although the shared factors are unobserved. Moreover, both the spatial autocorrelations among adjacent zones are detected in the both targets but they are controlled by the spatial error term included in the model.

Subsequently, hot zones for 'Pedestrian crashes per crash location ZIP' and 'Crash-involved pedestrians per residence ZIP' were identified, separately, using PSI measures. It was shown that the screening results for the two targets are similar but not exactly the same. After that, the screening results for the two targets were integrated to provide a more comprehensive perspective for pedestrian safety problems. A novel hot zone identification method was suggested to combine both hot zones with many pedestrian crash occurrences and hot zones with many crash-involved pedestrians in the residence. For the former zones, area-wide engineering treatments and enforcement can be provided as general countermeasures whereas targeted safety education and campaigns may be effective safety treatments for the latter ones. In conclusion, it is expected that practitioners are able to suggest appropriate safety treatments for pedestrian crashes using the screening method and results from this study.

ACKNOWLEDGEMENTS

The authors wish to thank the Florida Department of Transportation for funding this study. Some of the research for this paper was conducted as part of the efforts of the Southeastern Transportation Center (STC) at the University of Tennessee. The primary sponsor for the STC is the United States Department of Transportation through grant number DTRT13-G-UTC34.

REFERENCES

1. NHTSA, 2012 Motor Vehicle Crashes: Overview, U.S. Department of Transportation, National Highway Traffic Safety Administration, 2013.
2. NHTSA, Traffic Safety Facts 2012 Data: Pedestrians, U.S. Department of Transportation, National Highway Traffic Safety Administration, 2014.
3. LaScala, E. A., D. Gerber and P. J. Gruenewald. Demographic and Environmental Correlates of Pedestrian Injury Collisions: A Spatial Analysis. Accident Analysis and Prevention, Vol. 32, No. 5, 2000, pp. 651-658.
4. Ng, K., W. Hung, and W. Wong. An Algorithm for Assessing the Risk of Traffic Accident. Journal Of Safety Research, Vol. 33, No. 3, 2002, pp. 387–410.
5. Noland, R. B., and M. A. Quddus. Analysis of pedestrian and Bicycle Casualties with Regional Panel Data. Transportation Research Record, No. 1897, 2004, pp. 28-33.
6. Loukaitou-Sideris, A., R. Liggett, and H-G Sung. Death on the Crosswalk: A Study of Pedestrian-Automobile Collisions in Los Angeles, Journal of Education and Research, Vol. 26, No. 3, 2007, pp. 338-351.
7. Wier, M., J. Weintraub, E. H. Humphreys, E. Seto, and R. Bhatia. An Area-Level Model of Vehicle-Pedestrian Injury Collisions with Implications for Land Use and Transportation Planning. Accident Analysis and Prevention, Vo. 41, 2009, pp. 137-145.
8. Cottrill, C. D., and P. Thakuriah. Evaluating Pedestrian Crashes in Areas with High Low-income or Minority Populations. Accident Analysis and Prevention, Vol. 42, No. 6, 2010, pp. 1718-1728.
9. Ukkusuri, S., S. Hasan, and H. M. A. Aziz. Random Parameter Model Used to Explain Effects of Built-Environment Characteristics on Pedestrian Crash Frequency. Transportation Research Record: Journal of the Transportation Research Board, No. 2237, 2011, pp. 98-106.
10. Siddiqui, C., M. Abdel-Aty, and K. Choi. Macroscopic Spatial Analysis of Pedestrian and Bicycle Crashes. Accident Analysis and Prevention, Vol. 45, 2012. pp. 382–391.
11. Siddiqui, C., and M. Abdel-Aty. On the Nature of Modeling Boundary Pedestrian Crashes at Zones. In Transportation Research Board 91th Annual Meeting, Transportation Research Board of the National Academics, Washington, D.C., 2012.
12. Wang, Y., and K. M. Kockelman. A Poisson-Lognormal Conditional-Autoregressive Model for Multivariate Spatial Analysis of Pedestrian Crash Counts Across Neighborhoods. Accident Analysis and Prevention, Vol. 60, 2013, pp. 71-84.
13. Abdel-Aty, M., J. Lee, C. Siddiqui, and K. Choi. Geographical Unit Based Analysis in the Context of Transportation Safety Planning. Transportation Research Part A: Policy and Practice, 49, 2013, pp. 62-75.
14. Lee, J., M. Abdel-Aty, K. Choi, and C. Siddiqui. Analysis of Residence Characteristics of Drivers, Pedestrians, and Bicyclists Involved in Traffic Crashes. In Transportation Research Board 92nd Annual Meeting, Transportation Research Board of the National Academics, Washington, D.C., 2013.
15. Congdon, P. Applied Bayesian Modelling (2nd Edition). John Wiley and Sons, 2006.
16. Mitra, S., and S. Washington. On the Nature of Over-dispersion in Motor Vehicle Crash Prediction Models. Accident Analysis and Prevention, Vol. 39, No. 3, 2007, pp. 459-468.
17. Lord, D. and F. Mannering. The Statistical Analysis of Crash-Frequency Data: A Review and Assessment of Methodological Alternatives. Transportation Research Part A: Policy and Practice, Vol. 44, No. 5, 2010, pp. 291-305.

18. Ye, X., R. M. Pendyala, S. P. Washington, K. Konduri, and J. Oh. A Simultaneous Equations Model of Crash Frequency by Collision Type for Rural Intersections. Safety Science, Vol. 47, No. 3, 2009, pp. 443-452.
19. O'Sullivan, D. and D. Unwin. Geographic Information Analysis. John Wiley and Sons, 2002.
20. LeSage, J. P. and R. K. Pace. Models for Spatially Dependent Missing Data. The Journal of Real Estate Finance and Economics, Vol. 29, No. 2, 2004, pp. 233-254.
21. Besag, J. Spatial Interaction and the Statistical Analysis of Lattice Systems. Journal of the Royal Statistical Society B, Vol. 36, No. 2, 1974, pp. 192-236.
22. Spiegelhalter, D. J., N. G. Best, B. P. Carlin, and A. Van Der Linde. Bayesian Measures of Model Complexity and Fit. Journal of the Royal Statistical Society: Series B (Statistical Methodology), Vol. 64, No. 4, 2002, pp. 583-639.
23. Aguero-Valverde, J., and P. P. Jovanis. Bayesian Multivariate Poisson Lognormal Models for Crash Severity Modeling and Site Ranking. Transportation Research Record: Journal of the Transportation Research Board, No. 2136, 2009, pp. 82-91.

Comparing Hotspot Identification Methods at the Macroscopic Safety Analysis Level

Pei-Fen Kuo[*]
Assistant Professor
Department of Crime Prevention and Correction
Central Police University, Taiwan
Tel. (886) -033282321
Email: kpf@mail.cpu.edu.tw

Jaeyoung Lee
Postdoctoral Research Associate
Dept. of Civil, Environmental & Construction Engineering
University of Central Florida
Orlando, Florida 32816-2450
Tel. (407) 823-0300
jaeyoung@knights.ucf.edu

Mohamed Abdel-Aty
Professor and Chair
Dept. of Civil, Environmental & Construction Engineering
University of Central Florida
Orlando, FL 32816
Tel. (407) 823-4535
Email: m.aty@ucf.edu

Submitted for potential publication in Transportation Research Part A: Policy and Practice

February 15, 2015

[*]Corresponding Author

ABSTRACT

Compared to micro scale safety studies, macroscopic-focused research is more efficient at integrating zone-level features into crash prediction models and identifying hot zones in large study areas. However, few studies have focused on the limitations of current hotspot/hot-zone[1] identification methods (HSID) applied at the macro level. This study applied six common HSID methods and compared their consistency in identifying hot-zones. The crash data was based on five years of crash records from Central Florida (Orange, Seminole, and Osceola Counties).

The results showed that the hot-zones identified by the crash frequency, Empirical Bayesian, and Potential for Safety Improvement methods all had high consistency and stability over time, followed by the crash rate and Equivalent Property Damage Only methods. The Proportion method had the lowest consistency. Other possible factors related to the methods' performance were also examined, which included the time length of the before period, the time length of the after period, the time gap, hot-zone threshold (α), and different crash types. However, these factors affected the performance of the methods only slightly. Also, the main problem of the crash frequency method, regression-to-the-mean, was not found to affect the performance of the method at the macro level because the consistency stayed high even in cases where the time length of the before period was as low as one year. The detail proof is given in Appendix A.

Keywords: traffic safety, macroscopic screening, hot zone, microscopic screening, hotspot/hot-zone identification, regression-to-the-mean

1. INTRODUCTION

Network screening, or hotspot identification, is the process for reviewing a highway network to identify and rank sites with respect to traffic safety (*1*). There is a growing body of literature on the development of traffic crash network screening methods. The majority of these studies are performed at the microscopic level, which deals with the safety screening of road segments or intersections. In contrast, there have not been many studies focusing on the macroscopic-level screening analysis. The macroscopic level analyses concentrate on area-wide traffic safety, and they aim to incorporate traffic safety considerations into long-term transportation plans (*2*). In the screening studies, various hotspot identification performance measures have been used at the micro and macro levels. They include the crash frequency, Equivalent Property Damage Only (EPDO) crash frequency, crash rate, proportion, Empirical Bayes (EB), Potential for Safety Improvement (PSI), and excess crash frequency methods.

[1] We use the term hot-zone instead of hotspots at the macro level.

Traffic engineers adopt different HSID methods based on the specific research goals and data limitations. For example, both the crash frequency and crash rate measures are the most straightforward to implement, but they are not popular performance measures for screening since they have many disadvantages. For instance, both of them do not account for regression-to-mean bias, and do not estimate a threshold to indicate sites experiencing more crashes than predicted for sites with similar characteristics. Especially, in the case of crash frequency, it also does not account for traffic volume (*1*). Only a few researchers, who have attempted to compare various performance measures, used the crash frequency and crash rate measures (*3*-*6*). For a more advanced HSID method, the EPDO crash frequency measure assigns weighting factors to crashes by three severity levels (i.e., fatal, injury, and property damage only) to develop a combined EPDO frequency. The EPDO crash frequency measure has been adopted by Montella (*6*), Aguero-Valverde (*7*) and Young and Park (*8*).

Even though the proportion measure is often used to identify crash patterns of sites, it is not that widely used for hotspot identification. As its name indicates, sites are ranked based on the probability that the proportion of a specific crash type is larger compared to the threshold proportion (*6*). The proportion measure was adopted by Lyon et al. (*9*) and Montella (*6*) for screening analysis.

The EB method started with its application in the traffic safety field by Abbess et al. (*10*). It is a preferred method in the HSM (*1*). The EB estimate of expected crash frequency for a location is a weighted combination of the prediction obtained from an Safety Performance Function (SPF) and the observed crash frequency for the given location. Many researchers have adopted the EB measure for hotspot identification (*3*-*6*, *11*-*14*).

PSI, or excess crash frequency, is a performance measure of how many crashes can be reduced effectively for a particular site. The PSI for each site is the difference between the expected number of crashes and predicted number of crashes of the site. The expected crash count is generally calculated using EB or FB (Full Bayes/Hierarchical Bayes) methods, and the predicted crash count is estimated from SPF. Some researchers have applied PSI as a performance measure for screening analysis (*3*, *6*-*7*, *15*-*16*).

As mentioned previously, relatively few researchers have conducted screening analyses at the macroscopic scale. Aguero-Valverde (*7*) analyzed PDO, injury, and fatal crashes based on the data from Cantons in Costa Rica. Aguero-Valverde (*7*) estimated EPDO using a multivariate spatial model and ranked Cantons based on excess EPDO. Jiang et al. (*17*) adopted the random forest technique for hotspot identification at the macroscopic level. The authors screened hot zones based on Traffic Analysis Zones (TAZs) in Central Florida and used performance measures such as crash rates (i.e., crashes per mile and crashes per million-vehicle-miles-traveled) and crash density (i.e., crashes per square mile). The authors found that the crash density model performed best and recommended the use of crash density for the macroscopic level screening. Lee (*16*) and Abdel-Aty et al. (*18*) analyzed hot zones for total and fatal-and-injury crashes by integrating macroscopic and microscopic screening results. The authors employed the PSI as a screening performance measure in the study. Also, it should be noted that a new macro study unit: Traffic Safety Analysis Zones (TSAZs) systems were developed by using regionalization (*19*). In other words, this regionalization can alleviate limitations of the TAZ system by aggregating TAZs into a sufficiently large and homogenous zonal system. Regionalization refers to the process of combining a number of areal units into a smaller number of areas, while simultaneously optimizing an objective function (*20*).

Some effort has been made for comparing multiple hotspot identification performance measures. Persaud et al. (*3*) conducted a comparative analysis for different hotspot identification methods using signalized intersections and two-lane rural highway data and concluded that the refined EB method is relatively efficient. Also, Cheng and Washington (*4*) used simulated data and argued that the EB method outperforms other methods. Cheng and Washington (*5*) suggested novel evaluation tests for comparing different screening measures. These tests assess the reliability of the results, ranking consistency, false identification consistency, and reliability of the screening measures. The authors compared the four most common screening measures (i.e., crash frequency, crash rate, PSI, and EB) and concluded that EB is the superior method in most of the evaluation tests. Elvik (*14*) compared the crash frequency, crash rate, combining a critical crash frequency and crash rate, EB, and local risk factors to the EB using Norwegian data. The author found that the EB is the most reliable based on the epidemiological criteria in the study. Montella (*6*) assessed various screening performance measures using several quantitative evaluation tests. The authors compared the crash frequency, crash rate, EPDO, proportion, EB of total crash counts, EB of severe crash counts, and PSI methods. The authors argued that the EB measure performs better than other measures, which is quite consistent with previous studies. These comparative analysis studies commonly suggested that EB method is more desirable for the network screening analysis.

Nevertheless, the above findings were based on the results from comparative analysis studies at the microscopic level. No studies have attempted to compare the performance of HSID method in the macroscopic screening studies. We believe that some characteristics of microscopic studies differ from macroscopic crash analyses. For example, the effects of the regression-to-the-mean at the macro level should be smaller than that at the microscopic level. It is because entities in the macroscopic studies are aggregated by nearby geographic units, which account for the regression-to-mean bias. Justification for this assumption and a more detailed discussion are given in Appendix A. Also, the before time period at the macro level can be shorter than that at the micro level because the mean crash frequency of each unit is increased after regionalization at the macro level. Therefore, there is a need to compare various screening performance measures at the macroscopic level, which will contribute to the traffic safety field to guide the best measures for macroscopic screening analysis.

The rest of this paper is divided into four sections. Data collection and preparation are presented in section 2. Section 3 lists formulas of six common HSID methods and the criteria to evaluate different methods' performance. Section 4 provides the comparison results. Finally, conclusions and recommendations are provided in section 5.

2. STUDY DATA

The study area includes Orange, Seminole, and Osceola Counties (the largest metropolitan areas in Central Florida). The study period is from 2005 to 2010. Crashes reported on long-forms (from FDOT Crash Analysis Reporting (CAR)) and short-form crash reports (from the Signal Four Analytics) data were combined as one comprehensive crash dataset for analysis. Also, demographic, socioeconomic, planning, and safety data were collected from the US Census Bureau website, MPOs and FDOT district offices. The inventory file of the intersections was based on the Roadway Characteristics Inventory (RCI) dataset.

For conducting macro level safety analyses, TSAZs systems and a nested structure were used for minimizing boundary crashes. TSAZ is a new study unit aggregating current Traffic Analysis Zones (TAZs). Readers are referred to Lee et al. (*19*) for detailed information regarding TSAZs. GIS techniques were used to update crash and other characteristics data, when the study scale changes from TAZ or census block to TSAZ. In addition, a Bayesian Poisson Lognormal Spatial Error Model (BPLSEM) was adopted for the PSI analysis. The Poisson Lognormal models have been proposed as an alternative for the negative binomial (or Poisson Gamma) models for frequency data in traffic safety modeling. The Poisson Lognormal model is comparable to the negative binomial model; however, the Poisson lognormal model provides more flexibility compared to the negative binomial model. A spatial effect term was included in the equation to account for the spatial autocorrelation in the data. The BPLSEM is specified as follows:

$$y_i \sim Poisson(\mu_i) \tag{1}$$
$$\lambda_i = \exp(\beta_0 + \beta X_i + \theta_i + \varphi_i) \tag{2}$$
$$\theta_i = Normal\ (0, \tau_\theta) \tag{3}$$

where,

y_i is the number of aggregated total crashes of the *i*^th^ TSAZ,

β_0 is the intercept,

β's are the coefficient estimates of covariates (X_i),

θ_i is the random effect term,
φ_i is the spatial effect term, and
τ_θ is the precision parameter, which is the inverse of the variance and a given prior gamma distribution (0.5, 0.005).

The model was fitted with non-informative prior distributions, *Normal* (0, 10^{-6}) for β. Furthermore, the spatial distribution was implemented by specifying an intrinsic Gaussian Conditional Autoregressive prior with a $Normal\ (0, \tau_\varphi)$ distribution. The mean of φ_i is defined by

$$\bar{\varphi}_i = {\sum_{i\neq j} \varphi_j \times w_{ij}} \Big/ {\sum_{i\neq j} w_{ij}} \qquad (4)$$

where,
$w_{ij} = 1$, if zone *i* and *j* are adjacent, and
$w_{ij} = 0$, otherwise.

3. METHODOLOGY

As mentioned above, there are six common methods to identify hot-spots: crash frequency, crash rate, EPDO crash frequency, proportion, EB and PSI methods. The following section describes the details of these methods. As mentioned earlier, we use the term hot-zones instead of hotspots at the macro level.

The formulae and notations are based on Montella (*6*) and Cheng and Washington (*5*) for comparison purposes, and some settings were adjusted to account for the screening scale being changed from the micro to the macro level.

1. Hot-zone identification method

- Crash frequency

The crash frequency method is the most straightforward. Each study unit (e.g., TSAZ in this study) is ranked by its total crash frequency, and the hot-zones are the areas that have crash frequencies over defined thresholds. For example, the hot-zones are where the TSAZs have top 5 % of crash frequencies. It should be noted that the crash frequency method at the macro level is not normalized. One example of normalization is that crash frequency is divided by the segment length at the micro level. There are two reasons for not normalizing the data at the macro level. The first one is that the TSAZ has been regionalized, which accounts for crash heterogeneity and deals with boundary crash problems. The second reason is that existing variables, such as the area or population, do not entirely represent the exposure.

- Crash Rate

The crash rate method represents the average crash risk for individual drivers, while the crash frequency method reflects the crash risk for each TSAZ. The crash rate is the total crash frequency divided by the overall exposure, such as AADT or VMT for each TSAZ. This study used VMT from the US census website as traffic volume data. Compared to the crash frequency method, the hot-zone results of the crash rate method tend to shift toward the area that has lower traffic flow, like rural areas.

- Equivalent Property Damage Only (EPDO) Crash Frequency

This method accounts for crash costs for different injury levels. Different weights were developed to combine frequency and severity based on the approach of willing to pay. To maintain comparability, this study followed the same weighting as Montella (*6*), which gave different injury level, crashes weights of fatal: injury: PDO = 771:35:1.

- Proportion method: injury severe crash

The fourth method is a different type than the preceding three. Instead of focusing on crash frequency data, this method studies the probability of crashes in each zone. The first step is to define parameters regarding one target crash type, such as collision type, weather condition, or injury level. Then, an estimate of the probability of this specific crash type occurring among all crashes (e.g., p=N$_{\text{rear-end}}$/ N$_{\text{all}}$) is made. This study defines the probability as the ratio of injury and fatal crash frequency compared to the total crash frequency. Hot-zones are the sites having high probability for this specific crash type over the defined threshold.

$$P(X_{ij} \le x-1, n; p_j) = \sum_{i=0}^{x-1} \frac{(n)!}{(n-1)!(i)!} p_j{}^{i} (1-p_j)^{n-i} \qquad (5)$$

Where

P: the probability of crash type j occurring among all crashes

x: The crash frequency of crash type j for zone *i*

n: The crash frequency of all crashes for zone *i*

- Empirical Bayesian method : EB

The EB method is a preferred method in the Highway Safety Manual to estimate the long-term expected crash frequency. The EB estimate is a weighted combination of the predictions obtained from an SPF and the observed crash frequency for the given location (See Equation (2)). The weights are calculated based on the over-dispersion parameter, and the crash safety prediction models are based on Abdel-Aty et al. (*18*). The predicted crash counts were estimated using six sub-models in a nested structure for different roadway types.

$$EB = w \times E(Y) + (1 - W)N_i \qquad (6)$$

Where

W: The weight for zone *i* in the EB method $W_i = \dfrac{1}{1 + \hat{\Lambda}_{i1}^{T} \times \alpha}$

E(Y): The estimator for the average crash frequency of zone *i* in the study period,

N: The observed response for zone *i*

- Potential for Safety Improvement (PSI)

The PSI method is a measure of how many crashes can be reduced by implementing countermeasures. In other words, the PSI for each zone is the difference between the expected crash count and the predicted crash count

$$PSI = EB - E(Y) \qquad (7)$$

The PSIs were calculated and the TSAZs were ranked separately for urban and rural areas.

2. Criteria to evaluate method performance

Consistency and accuracy are two common criteria used to evaluate different HSID methods. The former means a high-risk hot-zone repeated during a study period, and the latter is whether the above HSID methods identify correct hot-zones (which is rarely known in the real world). Hence, this study will focus on the site consistency test (SCT). The reader may refer to Montella (*6*) and Cheng and Washington (*5*) for more details about other consistency tests.

$$SCT_i = \sum_{k=n-n\alpha+1}^{n} C_{k,j,i+1} \Big/ \left(\sum_{k=n-n\alpha+1}^{n} L_{k,j} \times y_{i+1} \right) \qquad (8)$$

Where

$C_{k,j,i+1}$: Crash frequency for zone j ranked k in the time period i;

$L_{k,j}$: The total road length within zone j;

y_{i+1}: the length of time period i+1

From the perspective of the engineer, a hot-zone with a high consistency represents a site that has high long-term crash rates and safety problems over years. In this case, infrastructure and engineering solutions and persistent enforcement and education might be needed, based on the type of safety problem over a short-term management plan.

4. RESULTS

After identifying hot-zones by using the six commonly HSID methods described above, we used the SCT test to evaluate the performance of different methods. Based on the limitations of HSID method at the micro level and previous study results (*5*-*6*), six scenarios were used to examine possible factors related to method consistency. They include a direct comparison of the methods, the length of the before period, the length of the after period, the time gap, the hot-zone threshold (α), and the different crash types. Scenario 1 compared the six most common HSID methods: crash frequency, crash rate, equivalent property damage crash frequency, proportion method, EB method, and the potential safety improvement method. Recall that Equations (1) to (3) are used to rank the hot-zone and Equation (4) is used for estimating the consistency. Scenario 2 varied the length of the before period from 1 year (short before period) to 4 years (long before period). It should be pointed out that HSM suggested that using 2 to 3 years for the length of the before period crash data in order to account for regression-to-mean bias at the micro level. For Scenario 3, the consistencies when the length of after period was increased from 1 to 2 years were examined. Scenario 4 varied the time gap (the length of time between the before and after periods) from 1 to 4 years. For Scenario 5, the consistencies of the different methods when the hot-zone threshold (α) was varied from the top 1 %, 5 %, and 10 % were examined. Finally, Scenario 6 investigated the consistency using different crash types: fatal-injury crash, and pedestrian crashes. For all scenarios, the default before period is 2008 and the after period is 2009. The hot-zone threshold (α) is 5 %. Note that scenarios 1 and 2 were analyzed simultaneously.

Scenario 1: Different HSID Methods

Figure 1 shows the order of consistency in Scenario 1 in crash frequency, EB, PSI, EPDO, and the proportion method. Overall, crash frequency, EB, and PSI method all have high consistency, followed by the crash rate and EPDO method. The proportion method has the lowest consistency.

Scenario 2: The Length of the Before Period

Similar results were observed in this scenario compared with Scenario 1. As shown in Figure 1, the crash frequency, EB, and PSI methods gave the highest consistency while the proportion method had the lowest consistency when the length of the before period increased from one year (2008) to four years (2005, 2006, 2007, and 2008). However, when the length of before period increases, the consistency of EPDO method increased while that of the crash rate method decreased. These findings are not consistent with Montella (*6*), because even the easiest-to-implement method (crash frequency) and PSI work as well as the EB method. However, poor performance of the proportion method is in line with Montella (*6*).

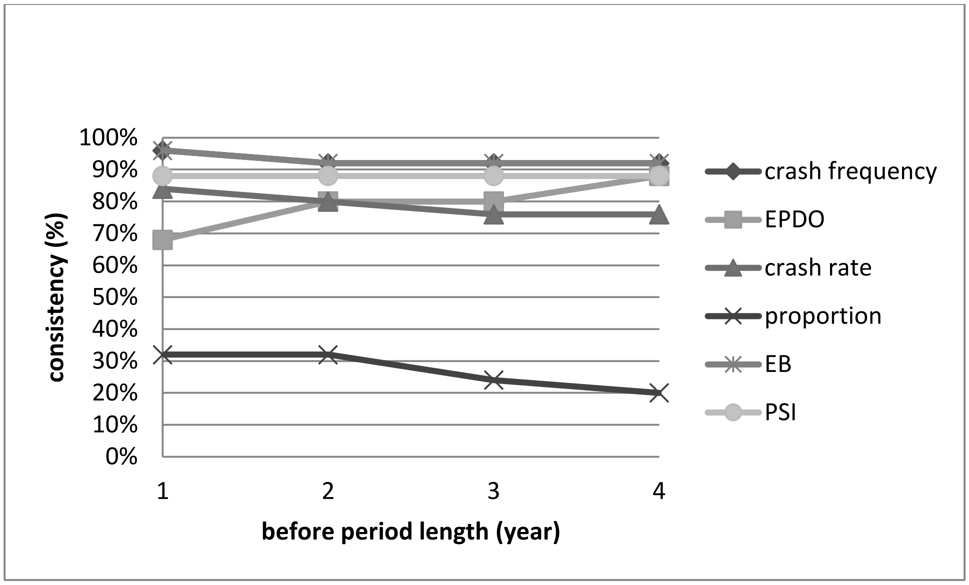


Figure 1. Consistency for different HSID methods in Scenario 2.

Scenario 3: The Length of the After Period

The length of the after period was extended from 1 year to 2 years (2009 and 2010). In other words, we want to examine the effect of adding one extra year of crash data on the HSID method performance. Crash frequency and PSI methods still have high consistencies, although their consistencies were reduced slightly when the length of after period was increased to 2 years. However, the consistencies of EPDO method, crash rate, and the proportion method increased when the length of after period increased. This finding is not significant and the trend may change when the length of the after period is extended even further (2010 is the latest year available).

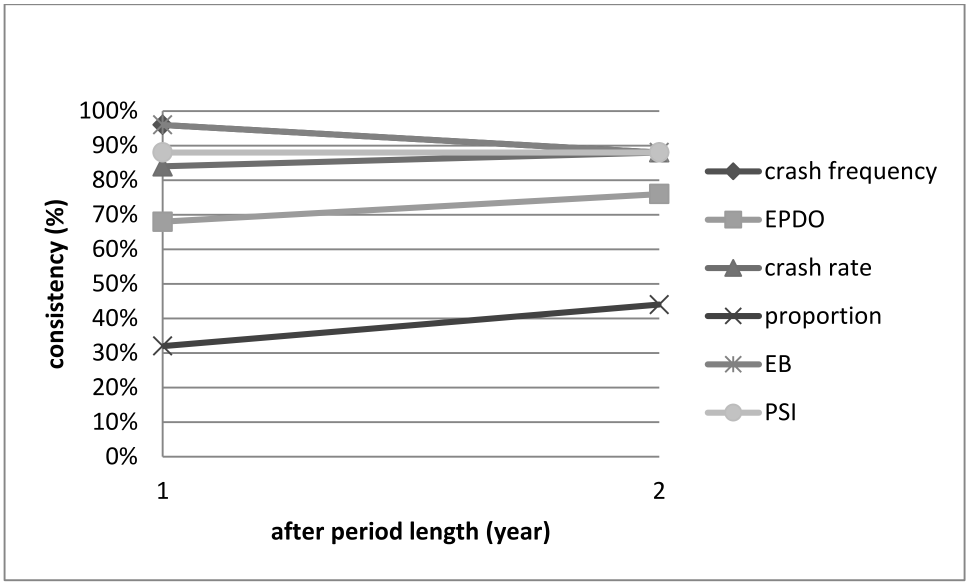


Figure 2. Consistency for different HSID methods in Scenario 3.

Scenario 4: Time Gap

Figure 3 shows how the time gap changes the consistency of the methods. The time gap length was increased from 1 year to 4 years. This figure shows that the crash frequency and PSI methods still have high consistency, although their consistencies reduced slightly when the length of the time gap increased. In addition, there was no clear trend of the consistency of the EPDO method and the crash rate method when the length of the time gap was increased. The proportion method still had the lowest consistency out of all the methods, and its consistency decreased when the length of the after period was increased, but recovers by the four-year point. The results from Scenario 4 indicate that the use of historical crash data to identify hot-zones does not change the consistency of the method in use.

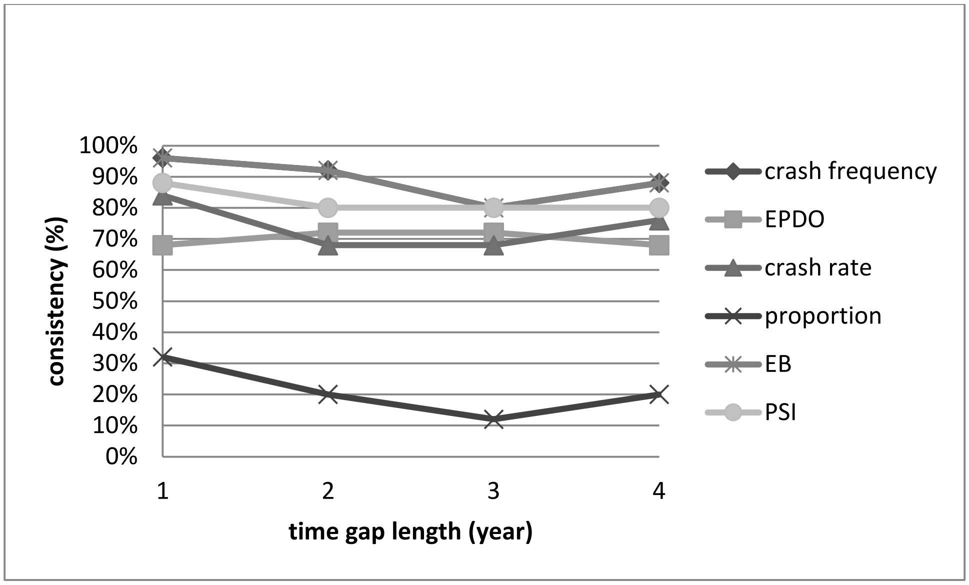


Figure 3. Consistency for different HSID methods in Scenario 4.

Scenario 5: Hotspot Threshold

Figure 5 shows that there is no clear trend of the consistency when the hotspot threshold changes. When the hotspot threshold was reduced from 95 % to 90 %, the consistencies of crash frequency, PSI, and crash rate methods were reduced. The consistency of the EPDO method was maintained, and the consistency of the proportion method increased, but was still very low (40 %). When the hotspot threshold was increased from 95% to 99%, the consistencies of the crash frequency, crash rate, and proportion methods increased as well. The consistencies of the EPDO and PSI methods, however, were reduced significantly.

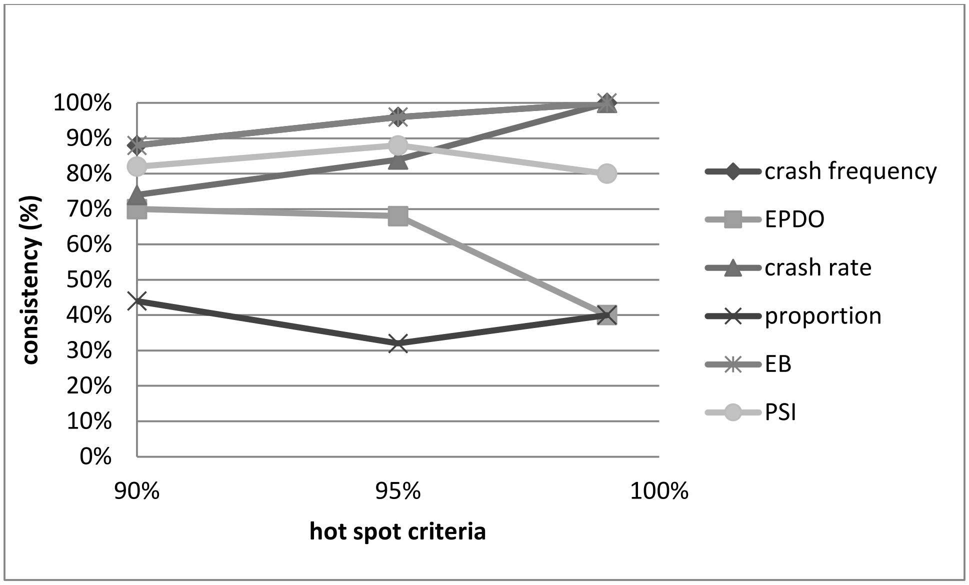


Figure 4. Consistency for different hotspot threshold in Scenario 5.

Scenario 6: Different Crash Types FI, Pedestrian crashes

In this Scenario, we focused on the consistency by examining specific crash types. The first type we looked at was fatal and injury crash data only. Figure 5 shows that the crash frequency and PSI methods have high consistency; even though their values in this case were slightly lower than when total crashed were used (reduced from 90 % to 80 %). When the length of before period was increased, the consistencies of all methods remained the same. Note that there are no results given for the EPDO, proportion, and PSI methods because PDO crash data was removed from the all crash data set.

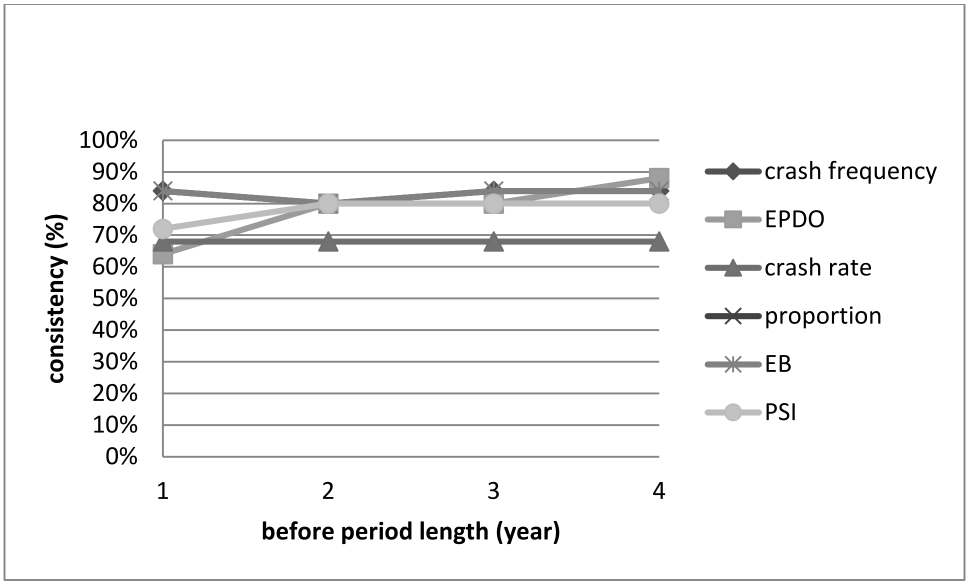


Figure 5. Consistency for fatal and injury crashes in Scenario 6.

Then, the same procedure was conducted by using pedestrian crash data. The results show that crash frequency and EB method still have high consistency, but the values were much lower than when total crash data was used (reduced from 90 % to 70 %). When the length of before period was increased, the consistency of all methods stayed the same. As mention above no results were given for the EPDO, proportion, and PSI methods because non-pedestrian crash data was removed from the all crash data set and we don’t have specific weighting factor and SPFs for the pedestrian crashes..

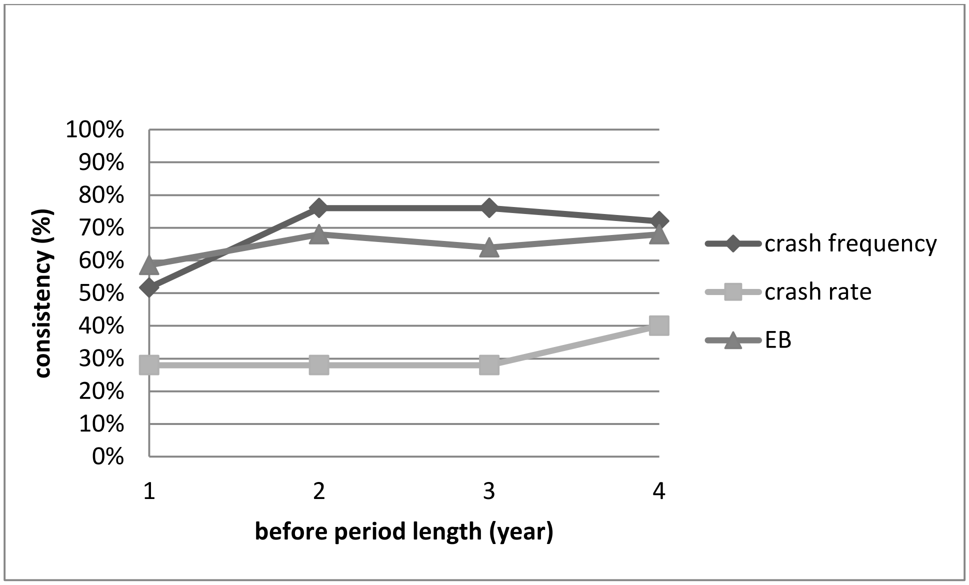


Figure 6. Consistency for pedestrian crashes in Scenario 7.

5. CONCLUSIONS

Previous studies have compared the performance of HSID methods at the micro level. However, none of the previous studies have examined whether the limitations of current HSID methods exist at the macroscopic safety analysis level. Hence, this paper compared the performance of six common HSID methods by the site consistency test at the macro level. Also, the limitations of current HSID methods were examined at the macro level.

The results showed that there is no significant regression-to-the-mean bias effect at the macro level even in case of crash frequency or crash rate, because the consistency of crash frequency and crash rate methods are relatively high. It may imply that regression-to-the-mean is not frequently observed since crashes were already highly aggregated by zones at the macro level. For more details, readers can refer to Appendix A. This is still true even when the length of the before period is just one year. We also showed that the limitations of HSID methods at the micro level were not an issue during the macro scale analysis. First, the length of the before period when applied to hotspot screening methods at the macro level can be short - even one year of crash data provided very high consistency. For different HSID methods, crash frequency, EB, and PSI methods all showed high consistency. The EPDO and crash rate methods showed the next highest consistency. The proportion method showed the lowest consistency, and its consistency was further reduced when the length of the before period was increased. When the length of the after period increased, we showed that the relative consistency of the different methods stayed the same. The crash frequency and PSI methods showed high consistency, although their consistencies were slightly reduced in value when the length of the time gap was increased.

There was no clear trend seen in the consistencies of the EPDO method and the crash rate method. When the hotspot criterion was increased from 95 % to 99 %, the consistencies of the crash frequency, crash rate, and proportion methods increased as well. The consistencies of the EPDO and PSI methods, on the other hand, were lowered significantly. For fatal and injury crashes, the crash frequency and PSI method still showed high consistency, although at slightly lower values than for total crash data (90 % → 80 %). For more rare crash types such as pedestrian crashes, the crash frequency and EB methods showed high consistency, with only slightly lower values than for total crash data (90 % → 70 %).

There are a few limitations to this study. First, the consistency results may change at different macroscopic levels. Although TSAZs were developed and suggested for the macro-level screening here, larger geographic units such as traffic analysis districts (TADs) and counties may generate different results. Also, we have used Orange, Seminole, and Osceola Counties as the study area. This area is a part of the FDOT

District 5 and Metro Plan Orlando. Our current results are area-specific at this stage, particularly because the research team has integrated TAZs to develop a new zonal system: TSAZs. Thus, in order to make other districts or MPOs use this method, there is a need to collect data from at least one or more districts (e.g., Tampa) and validate and refine our results.

ACKNOWLEDGEMENTS

The authors wish to thank the Florida Department of Transportation for funding this study. The authors would also like to thank the Southeast Transportation Center (STC) for partially funding this research. STC is funded by a regional University Transportation Center lead by the University of Tennessee.

REFERENCE

1. Highway Safety Manual, AASHTO, 2010.
2. Washington, S., I. V. Schalwyk, S. Mitra, M. D. Meyer, E. Dumbaugh., and M. Zoll. NCHRP Report 546: Incorporating Safety into Long-range Transportation Planning. Transportation Research Board, Washington, D.C., 2006.
3. Persaud, B., C. Lyon, and T. Nguyen. Empirical Bayes Procedure for Ranking Sites for Safety Investigation by Potential for Safety Improvement. Transportation Research Record: Journal of the Transportation Research Board, Vol. 1665, 1999, pp. 7-12.
4. Cheng, W., and S. P. Washington. Experimental Evaluation of Hotspot Identification Methods. Accident Analysis and Prevention, Vol. 37, No. 5, 2005, pp. 870-881.
5. Cheng, W., and S. Washington. New Criteria for Evaluating Methods of Identifying Hot spots. Transportation Research Record: Journal of the Transportation Research Board, No. 2083, 2008, pp. 76-85.
6. Montella, A. A Comparative Analysis of Hotspot Identification Methods. Accident Analysis and Prevention, Vol. 42, No. 2, 2010, pp. 571-581.
7. Aguero-Valverde, J. Multivariate Spatial Models of Excess Crash Frequency at Area Level: Case of Costa Rica. Accident Analysis and Prevention, Vol. 59, 2013, pp. 365-373.
8. Young, J., and P. Y. Park. Hotzone Identification with GIS-based Post-network Screening Analysis. Journal of Transport Geography, Vol. 34, 2014, pp. 106-120.
9. Lyon, C., B. Gotts, W. K. F. Wong, and B. Persaud. Comparison of Alternative Methods for Identifying Sites with High Proportion of Specific Accident Types. Transportation Research Record: Journal of the Transportation Research Board, Vol. 2019, 2007, pp. 212-218.
10. Abbess, C., D. F. Jarrett, and C. C. Wright. Accidents at Blackspots: Estimating the Effectiveness of Remedial Treatment, With Special Reference to the" Regression-to-Mean" Effect. Traffic Engineering and Control, Vol. 22, No. 10, 1981, pp. 535-542.
11. Higle, J. L., and M. B. Hecht. A Comparison of Techniques For The Identification of Hazardous Locations. Transportation Research Record, No. 1238, 1989, pp. 10-19.
12. Miaou, S. P., D. and Lord. Modeling Traffic Crash-Flow Relationships for Intersections: Dispersion Parameter, Functional Form, and Bayes versus Empirical Bayes Methods. Transportation Research Record: Journal of the Transportation Research Board, Vol. 1840, 2003, pp. 31-40.
13. Lord, D. 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, Vol. 38, No. 4, 2006, pp. 751-766.

14. Elvik, R. Comparative analysis of techniques for identifying locations of hazardous roads. Transportation Research Record: Journal of the Transportation Research Board, 2083, 2008, pp. 72-75.
15. Jiang, X., M. Abdel-Aty, and S. Alamili. Application of Poisson Random Effect Models for Highway Network Screening. Accident Analysis and Prevention, Vol. 63, 2014, pp. 74-82.
16. Lee, J. Development of Traffic Safety Zones and Integrating Macroscopic and Microscopic Safety Data Analytics for Novel Hot Zone Identification (Doctoral Dissertation). University of Central Florida, Orlando, Florida, 2014.
17. Jiang, X., M. Abdel-Aty, and J. Lee. Investigating Macrolevel Hotzone Identification and Variable Importance Using Random Forest Models. Presented at the 93rd Annual Meeting of the Transportation Research Board, Washington, D.C., 2014.
18. Abdel-Aty, M., P. F. Kuo, X. Jiang, J. Lee, and S. Al Amili. Two Level Approach to Safety Planning Incorporating the Highway Safety Manual (HSM) Network Screening (BDK78-977-13), Retrieved from Florida Department of Transportation (http://www.dot.state.fl.us/research-center/Completed_Proj/Summary_SF/FDOT-BDK78-977-13-rpt.pdf).
19. Lee, J., M. Abdel-Aty, and X. Jiang. Development of Zone System for Macro-level Traffic Safety Analysis. Journal of Transport Geography, Vol. 38, 2014, pp. 13-21.
20. Guo, D., and H. Wang. Automatic Region Building for Spatial Analysis. Transactions in GIS, Vol. 15, No. s1, 2011, pp. 29-45.
21. Moeltner, K. Addressing Aggregation Bias in Zonal Recreation Models. Journal of Environmental Economics and Management Vol. 45, No. 1, 2003, pp. 128-144.
22. Shaw, D. On-Site Sample's Regression: Problems of Non-Negative Integers, Truncation, and Endogenous Selection, Journal of Econometrics, Vol. 37, 1988, pp. 211-223.
23. Barnett, AG, Pols van der JC, and Dobson AJ. Regression to the mean: what it is and how to deal with it. International Journal Epidemiology Vol. 34, 2005, pp. 215-220.

APPENDIX A

Two main reasons describe why the regression-to-the-mean phenomenon is not frequently observed in the macro level safety analysis.

(1) In the environmental economic field, aggregated Negative Binomial models usually do not need to be adjusted for truncation, because we can get missing information of non-participant from census data (21) and (22).

(2) For the micro level safety analysis, we assume that the crash frequency of site i (an intersection or a segment), Y_i, follows the negative binomial distribution. As for the macro level safety analysis, the crash frequency of subarea j is the sum of independent negative-binomially distributed random variables Y1, ..., Yn (n: the number of sites located within subarea j). Because the negative binomial distribution is infinitely divisible, which means that the sum of independent negative-binomially distributed random variables with shape parameter, r1 and r2, and the same value for parameter p is also negative-binomially distributed with the same p but with new shape parameter, r'=r1 + r2. Moreover, when this new r' is sufficiently large, Σ Y_i is therefore approximately normal as a result of the central limit theorem. Then, we can define the regression-to-the-mean effect by equation (A. 1)

$$E\left(Y_{i1} \middle| Y_{i1} > C\right) = \mu_1 + \sqrt{\sigma^2 + \phi} \times \frac{f(d)}{(1 - F(d))} \qquad \text{(A.1)}$$

where $d = \frac{(c - \Lambda_1)}{\sqrt{\sigma^2 + \phi}}$ and f(d) and F(d) are the PDF and CDF of standard normal distribution respectively. According to Barnett et al. (23), the effect of RTM will decrease because of the smaller measurement variability ($\sigma^2 \rightarrow \sigma^2/m$). In other words, the effect by aggregating m adjacent intersections is similar as selecting subjects based on their multiple measurements within m year.

How Traffic Crashes Affect Congestion on Urban Expressway

Qi Shi *

Department of Civil, Environmental and Construction Engineering
University of Central Florida
Orlando, Florida 32816-2450
(407) 823-0300
shiqi@knights.ucf.edu

Mohamed Abdel-Aty

Department of Civil, Environmental and Construction Engineering
University of Central Florida
Orlando, Florida 32816-2450
(407) 823-1374
Fax: (407) 823-3315
m.aty@ucf.edu

Word Count: 5098 words + 4 Figures + 5 Tables = 7248 equivalent words

July 23, 2014

Submitted for possible presentation and publication at the TRB 2015 Annual Meeting

How Traffic Crashes Affect Congestion on Urban Expressway

By
Qi Shi, Mohamed Abdel-Aty

ABSTRACT

Provision of efficient and safe services to motorists has long been the major tasks for traffic professionals. Researchers have made considerable effort to explore the crash contributing factors and factors determining the incident durations. However, the issue of how crashes lead to congestion hasn’t yet been addressed. This study aims at clarifying this question by evaluating three urban expressways in Central Florida area. Both real-time traffic detection data and individual crash reports were employed. According to the real-time traffic data, it was found that a proportion of crashes led to congestion while other didn’t. For a comprehensive interpretation of the distinct effects, four classes of crashes based on their impact on congestion were generated and potential contributing factors were extracted. According to the structure of crash classification, one multinomial and two separate binomial logit models were developed under Bayesian framework to identify the effects of the candidate variables. Conclusion and model performance of the multinomial and binomial logit models generally agree with each other while binomial model offer more straight forward interpretation. Peak hour, number of lanes, weather condition and crash severity significantly affect the probability of the occurrence of the four types of crashes. However, the effects and the significance of some variables differ based on pre-crash congestion status. The findings of this paper suggested the necessity to include real-time traffic data in emergency response strategies. Moreover, the response procedure could also be assisted by the temporal, spatial, weather and severity related information about the crashes.

Keywords: Urban Expressway; Crashes; Congestion; Real-time Traffic Data; Bayesian Logit Model

INTRODUCTION

How to provide motorists with efficient and safe services is the principal concern for traffic engineers. Past decades have seen the development of high speed facilities and introduction of advanced Intelligent Transportation System (ITS) technologies to improve highway operation. In the meantime, safety campaigns including regulation, education, and scientific research have been carried out to bring down the losses associated with crashes. Although great efforts were made, issues regarding traffic safety and operation still remain hot topics for researchers. Extensive studies have been conducted to explore crash contributing factors and corresponding countermeasures to reduce crash occurrence. Conclusions from most existing literature have confirmed the relationship between traffic flow parameters and safety conditions. In the face of incidents, incident duration has also been examined by many researchers to reduce its impact on traffic operation.

In this paper, one issue that is overlooked by both types of research discussed above is investigated. Different from analyses identifying factors leading to crash occurrence or factors affecting the incident duration, the objective of this study tries to answer the following questions: 1) do all the crashes cause congestion? 2) if not, what factors make the crashes' impact on congestion diverse? To achieve the goal, three urban expressways operated by Central Florida Expressway Authority (CFX) were evaluated. The expressways are toll roads connecting downtown Orlando and neighboring area, carrying both commuting and tourist traffic. For more accurate and effective traffic monitoring, the authority have installed Microwave Vehicle Detection System (MVDS) on the expressways. On the 75-mile network of interest, 275 MVDS detectors are deployed. These detectors monitor traffic flow continuously and archive the data at one-minute interval. Operational performance of the expressways then can be evaluated through the MVDS traffic data. In this study, real-time traffic data and the detailed information from crash reports were extracted for each crash case to identify the effects of crashes on traffic congestion. Both Bayesian binomial and multinomial logit models were utilized to identify the factors leading to those potential diverse effects.

BACKGROUND

Existing studies exploring the relationship between operation and safety emphasize on whether congestion leads to crash occurrence. Both crash frequency models and real-time prediction models are developed. The common objective of these analyses is that by identifying the effects of congestion on traffic safety, traffic professionals can come up with countermeasures to reduce crashes. Unfortunately, no conclusive statement has yet been reached. Research with results confirming that congestion can increase crash occurrence (*1-2*); with results that the onset of congestion could increase crash occurrence, but would have low crash rates under highest traffic volume (*3*); with results declaring that congestion has no significant impact on safety or specific types of crashes (*4; 5*); and with results that congestion decreases crash occurrence (*6*) were all found in literature. In most of these studies, the effects of congestion on safety are interpreted from the density, speed and speed variation points of view. This type of work only evaluated operation-safety relationship in a unidirectional way. By pinpointing the contributing traffic parameters to crashes, we could reduce crash occurrence by adopting more proactive traffic management strategies. Nevertheless, crashes are still highly random in nature and could hardly be eliminated. Once a crash occurs on roadway, the most urgent task is to restore the traffic by quick response of the traffic authority and police patrol.

It is widely acknowledged that traffic crashes as unexpected events would temporarily reduce the road capacity and result in non-recurrent congestion. Based on this argument, many researchers have explored the incident duration under the traffic incidents (*7-14*). By modeling the time duration of incidents, it is hoped that the factors affecting the incident duration could be identified to help avoid secondary crashes and minimize its impact on traffic flow. In these studies, crashes as one type of traffic incidents are studied together with vehicle breakdowns, debris on road and other unplanned events (*12*). Police response time (*9*), the locations of the incidents (*10*), incident types (*13*), degree of incidents (*12*), etc. are suggested as strongly related to incident duration. More effective response strategies are expected based on the results from the above studies.

These two types of studies aim at preventing traffic crash occurrence and limiting traffic incident duration. Yet there is still a need to provide a more comprehensive understanding of the safety-operation relationship. Traffic crashes pose much more hazard for motorists on the roadways and cause huge social-economic losses compared with other types of traffic incidents. The effects of crashes then should be examined in more details. Moreover, the effects of incidents on traffic flow could be distinct. In some cases, only traveling lanes or shoulders are blocked due to the incidents; in some cases both traveling lanes and shoulders are blocked during different phase of clearance; and in other cases neither is blocked. Therefore it is possible that the incident duration of a crash is different from the time duration it affects operation especially congestion. These issues serve as the motivation of this study. In this current study, real-time traffic information was introduced to illustrate the effects of crashes on traffic congestion. Individual crash reports were utilized to identify significant factors leading to these effects. Expected contributions from this paper are deeper insights into the operation-safety relationship and practical suggestions for allocation of the rescue resources. Although for freeways without the ITS traffic detection facilities the direct measurement of crashes' effects is not available, part of the conclusions of this study are still applicable.

DATA PREPARATION

The three expressways under evaluation in this study are the segments managed by CFX. The network consists mainly of SR 408, SR 417 and SR 528, reaching 75 miles. SR 408 is the spine of the system which travels through downtown Orlando. Compared with the other two expressways, SR 408 has both the highest overall and commuting traffic. SR 417 is located in the outer Orlando area, providing a fast passage to suburban areas. SR 528 connects the Orlando international airport and coastal and attraction areas, serving for the convenience of both residents and tourists. CFX has converted the toll plazas to open tolling and installed Electronic Toll Collection (ETC) systems on the expressways. Recently they also introduced MVDS for more active traffic management. Table 1 implies that the MVDS covers the expressway network with an average distance between adjacent detectors of less than 1 mile. The deployment density of the devices ensures to reflect the traffic conditions on the whole network with convincible precision. Each detector monitors the traffic flow at the installed location and returns the data containing volume, speed, occupancy, volume by vehicle types on each traveling lane at 1-minute interval. Based on the information, real-time congestion intensity can be calculated. In this study, the rate of reduction in speed caused by congestion from the free flow speed condition is adopted as congestion index (*3; 15*). It is defined as

$$\begin{aligned} \mathrm{CI} &= \frac{\text{free flow speed} - \text{actual speed}}{\text{free flow speed}} \text{ when } \mathrm{CI} > 0; \\ &= 0 \text{ when } \mathrm{CI} \leq 0 \end{aligned} \tag{1}$$

where CI is a continuous congestion measure from 0 to 1. The free flow speed in this study is the 85th percentile speed at each detection location. The higher the CI, the more severe the congestion is. It is defined when CI is above 0.2, congestion occurs. The MVDS data have been collected since July, 2013.

Except for April, 2014 during which month the authority upgraded their system and did not archive the MVDS data, eleven months traffic data till June, 2014 were collected.

TABLE 1 MVDS Deployment on CFX System

Route	Length (mi)	Direction	Mainline Detectors	Distance between adjacent detectors			
				Mean	Std Dev	Min	Max
SR 408	21.4	EB	55	0.38	0.18	0.10	1.00
		WB	55	0.39	0.18	0.10	1.00
SR 417	31.5	NB	55	0.58	0.28	0.20	1.30
		SB	55	0.58	0.28	0.20	1.20
SR 528	22.4	EB	26	0.84	0.79	0.10	3.00
		WB	29	0.84	0.82	0.10	3.10

The crash data were downloaded from Signal Four Analytics database. For each crash case, the basic information (crash time, geocoded location, crash type, severity, vehicles involved, weather conditions, etc.) is incorporated in the crash report. During the studied time period, 838 crashes occurred on the mainline of the three expressways. The geocoded locations of crashes were used to match MVDS detectors to the crashes. As illustrated in Figure 1, to study the effects of crashes on congestion, the detector upstream to the crash location can reflect the traffic condition after crash occurrence. Consequently, the nearest upstream MVDS detector (U1) was assigned to each crash case. The traffic conditions 10 to 5 minutes prior to the reported time of crash and 0 to 5 minutes after the reported time were extracted. The 10 to 5 minutes instead of 5 to 0 minutes prior to crashes were selected to account for the possible delay between the real crash time and the time it is reported and recorded. Among the total 838 crashes, the real-time traffic data were successfully matched for 809 crashes and missing for the other 29 crashes. In the following analysis of the effects of crashes on congestion, the 809 crashes were used.

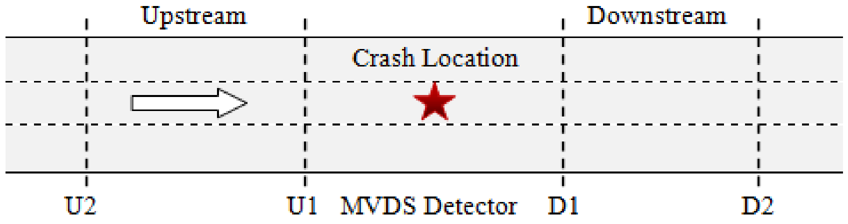
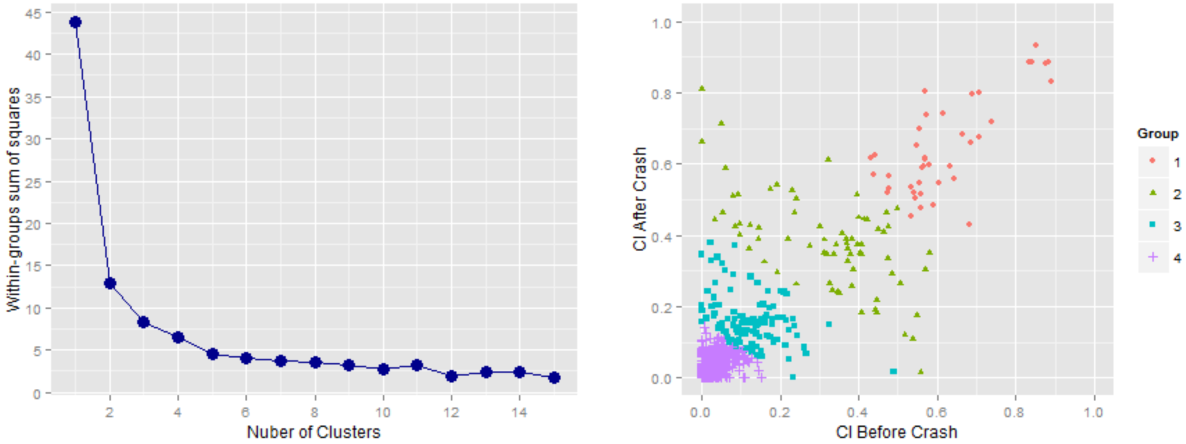


FIGURE 1 Crash location and MVDS detector assignment.

CRASH CLASSIFICATION

How the crash alters the congestion status on the mainline can be identified by comparing the CIs before and after crashes. With the total 809 crashes, the patterns of before-after congestion conditions were evaluated first using clustering method. To partition the crashes into different clusters within which they share higher similarities, K-means clustering method was tested. K-means clustering method is a popular method for unsupervised classification. Several traffic safety studies have implemented this technique to group the crash data (*16-18*). In the K-means clustering analysis, the number of clusters has to be specified in advance. Therefore the selection of appropriate number of clusters will be crucial for the interpretation of the clustering results. The theoretical foundation of the method lies in that the sum of squares of the observations to their assigned cluster centers is a minimum (*19*). Figure 2 (a) was generated to show the total within-groups sum of squares under different number of clusters. The sharp decrease from 1 to 4 clusters and the relatively flat curve after 4 clusters suggest a 4-cluster solution. Figure 2 (b) shows that K-means clustering classified the crashes based on the congestion status before and after crashes. However, the clustering results do not differentiate the crash effects clearly. Within the same group, part of the crashes exhibit significant changes in before-after congestion intensity while others not. As a result, machine learning might be inappropriate for the purpose of this research and manual classification was applied.



(a) **(b)**

FIGURE 2 (a) K-means cluster determination; (b) K-means cluster results.

In the scatter plot of Figure 2 (b), most of the dots representing the 809 crashes concentrate in the lower left corner. These dots indicate that the crashes occurred under non-congested conditions and did not lead to congestion afterwards. A proportion of the crashes were located along the 45-degree line in the higher part. Therefore they stand for those crashes happening under congestion. Nevertheless, their effects on traffic were very limited and did not worsen the congestion conditions. Another significant portion of crashes are in the upper left side of the figure. For these crashes, they occurred either under congested or non-congested conditions. The CIs after the crashes are much higher than CIs before crashes, which imply them as crashes that deteriorate the congestion on the mainline. One will also notice few crashes are located at the lower right side of the figure. Plain interpretation for these crashes should be that these crashes were observed under congestion conditions. Yet after the crash occurrences, the congestion intensity was relieved significantly. These cases will rarely exist in reality. Based on the above analysis, five clusters were manually created. Figure 3 illustrates how the Type 1 to Type 4 crashes were classified. To differentiate whether significant delays were caused by the crashes, 5 mph reduction in speed was selected as the cutoff point. One concern may rise regarding this classification that whether the same speed reduction would represent much different impact under different congestion states. This is a reasonable suggestion and perhaps could be considered together with the Level of Service in future analysis. As a pioneering evaluation, this study aims to draw a general conclusion first and perhaps go into detailed analysis in the future.

Crashes in the 5th cluster are defined as abnormal data are defined based on the assumption that crashes could either significantly reduce or have minor impact on the traffic speed at upstream locations but not increase the speed significantly. In the crash data, 46 crashes (5.6% out of the total 809 crashes) had speed after crash occurrence higher than speed before crash for 5 mph or more, therefore the 5th cluster was eliminated from further analysis. Whether the 5 mph is the best cutoff value is worth further exploration in the future. Figure 4 shows the manual clustering results. From the figure, we can answer our first question raised previously: crashes do not necessarily cause traffic congestions on the urban expressways. The contributing factors that make their impact diverse will be investigated using the information from individual crash reports in the following section.

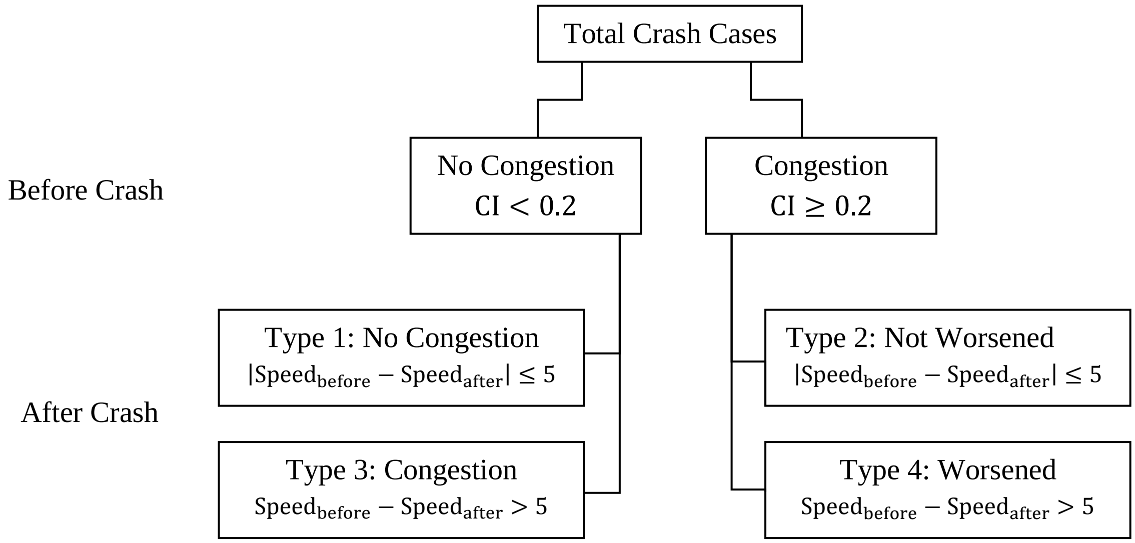


FIGURE 3 Crash classification procedure.

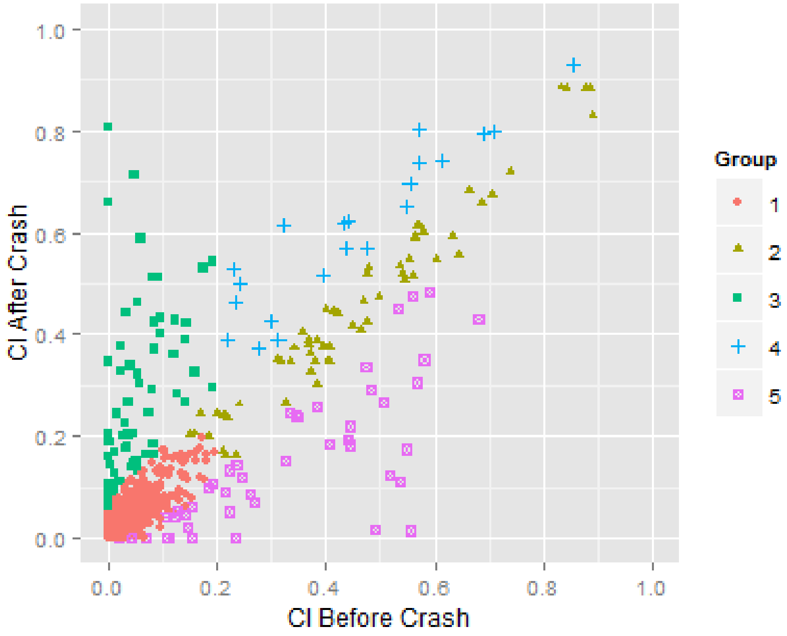


FIGURE 4 Crash classification based on the effects of crashes.

BAYESIAN LOGIT MODEL

As discussed above, four types of crashes were classified. To statistically study their patterns and the contributing factors to each class, logit models were applied. Given that four types of crashes were involved, multinomial logit (MNL) models were considered. The MNL model was constructed in Bayesian framework. However, a concern raised regarding the independence from irrelevant alternatives (IIA) assumption in multinomial logit regression. IIA assumption means that the choice between two alternatives is unaffected by introduction of additional choices, which might not hold true in reality. To overcome the issue, nested logit models are suggested by researchers to relax the IIA assumption (*20*). In this study, all the explanatory variables are crash related characteristics. The odds ratio between two clusters won't be affected by attributes of the additional clusters. As a result, the IIA assumption is not violated in this study and MNL model is valid. Regarding the current work, Figure 3 shows the nested structure of the crash classification. However, traffic congestion conditions prior to crashes can be determined by the MVDS data and two separate binomial logit models conditional on prior crash conditions instead of nested logit model can be developed. The first binary model compares the crash effects under non-congested before crash condition (Type 1 vs. Type 3) while the second binary model compares their effects under congestion before condition (Type 2 vs. Type 4). The specifications of the models are generalized below:

Binary logit model

$$\pi(\mathrm{x}) = P(Y = 1|\mathbf{x}) = 1 - P(Y = 0|\mathbf{x}) \tag{2}$$

$$\log\frac{\pi(\mathrm{x})}{1-\pi(\mathrm{x})} = logit[\pi(\mathrm{x})] = \beta_0 + \boldsymbol{X\beta} \tag{3}$$

where Y is the binary response and $\boldsymbol{X}$ stands for the matrix of explanatory variables. β_0 and $\boldsymbol{\beta}$ are intercept and vector for parameter coefficient.

Multinomial logit model

$$\pi_j(\mathrm{x}) = P(Y = j|\mathbf{x}) \text{ with } \textstyle\sum_j \pi_j(\mathbf{x}) = 1 \tag{4}$$

$$\log\frac{\pi_j(\mathrm{x})}{\pi_1(\mathrm{x})} = \beta_0 + \boldsymbol{X\beta},\ j = 2,3, \dots J \tag{5}$$

where Y is a categorical response with J categories ($J = 4$ in this study). The probability of each category sums to one. $\boldsymbol{X}$, β_0, and $\boldsymbol{\beta}$ bear the same meaning as in the binary model.

Bayesian inference is becoming popular in the transportation research arena, largely due to the use of Markov Chain Monte Carlo (MCMC) methods. Compared with traditional statistical inference methods such as Maximum Likelihood or Ordinary Least Squares, Bayesian inference allows prior knowledge, includes uncertainty in the model, is less sensitive to small sample size and accommodates more complicated models. It is adopted in this research for more realistic parameter estimates and prediction. In Bayesian inference, prior distributions for the parameters are required. In both the binary and multinomial logit models, non-informative priors $Normal(0, 10^3)$ are assigned to β_0 and $\boldsymbol{\beta}$. The models were built in WinBUGS software, 15000 iterations were run and the first 5000 were discarded as burn-in period. To ensure parameter convergence, three chains were simulated and the trace plots overlapped one another. The Deviance Information Criterion (DIC) was used as a Bayesian measure of model complexity and fit (2). Bayesian Credible Interval (BCI) was used for parameter estimation. If the 95% BCI does not contain 0, then the effect of the variable is significant. Classifier performances of the logit models were evaluated using receiver operating characteristics (ROC).

MODELING RESULTS AND DISCUSSION

Variable Description

To identify how the crashes affected the traffic congestion on the expressways. Information potentially pertinent to crash effects was extracted from the crash reports. The information could be broken down into four categories: spatial related factors, temporal related factors, crash related factors and weather related factors. Spatial related factors are expressway identifiers and number of lanes at the crash locations. Temporal related factors include peak-hour indicator, weekend indicator. Peak hours are defined as 7:00 to 9:00 and 17:00 to 19:00. Crash related factors are the number of vehicles involved in a crash and crash severity. Crashes involving four or more vehicles were rare and therefore they were combined with crashes involving three vehicles. In Highway Safety Manual (2010), crash severity is divided into five levels denoted as "KABCO" considering possible injury and differentiating incapacitating injury and non-incapacitating injury. Nevertheless, the crash database used in this research only recorded the crash severity as property damage only (PDO), injury and fatal. On the three expressways, only two fatal crashes occurred during the study time period. In response, injury and fatal crashes were aggregated to become severe crashes against the PDO crashes. Weather related factors are the weather conditions recorded at the time of crash occurrences. Fog cases were few and they were combined with rainy conditions. All of the seven candidate variables are categorical variables as shown in Table 2. In Table 2, summary statistics of the crash frequencies for each category of the variables and the percentage of each type of crashes are provided.

Whether to incorporate all of the variables in Table 2 in the Bayesian logit models depends on the correlation between the variables. For categorical variables, the Pearson's Chi-square test was implemented. Of the seven candidate variables as shown in Table 3, peak hour indicator, crash severity, number of lanes and weather condition were identified to be independent from each other with $\mathrm{P}-$ value ≥ 0.05. They were included in the final logit models. Other three variables, namely expressway, weekend and vehicles involved were significantly correlated with some of the four variables above and did not outperform the four variables in the logit models. Therefore they were excluded.

Multinomial Logit Model Results

To evaluate the impact of crashes on congestion, a preliminary attempt has been made to explore the change in speed before and after the crash occurrence and their relationship with the crash-related characteristics using linear regression. Nevertheless, directly setting the change in speed before and after crashes as the dependent variable and adoption of linear regression did not provide satisfying goodness of fit (low R^2s) and most of the involved variables were insignificant. Consequently, classification of crashes and logit model might prove more efficient in identifying crashes' impact on congestion.

Multinomial Bayesian logit model was evaluated first (as shown in Table 4). Type 1 crashes were set up as the baseline. Peak hours were found to significantly increase the logarithmic odds ratio of other types of crashes against Type 1 crashes. The effects of peak hours can be understood as during peak hours, recurrent congestion tends to occur. Therefore crashes under congested traffic flow are more likely to be observed during peak hours. In addition, even if no congestion exists prior to crashes, the higher traffic volume in peak hours will cause the impact of crashes to be more prominent.

TABLE 2 Statistics Summary of Variables

Crash Clusters	Type 1		Type 2		Type 3		Type 4		Total
Expressway									
SR 408	290	(76.7%)	49	(13.0%)	31	(8.2%)	8	(2.1%)	378
SR 417	169	(92.9%)	4	(2.2%)	7	(3.8%)	2	(1.1%)	182
SR 528	153	(75.4%)	9	(4.4%)	30	(14.8%)	11	(5.4%)	203
Peak Hour									
No	447	(88.9%)	16	(3.2%)	36	(7.2%)	4	(0.8%)	503
Yes	165	(63.5%)	46	(17.7%)	32	(12.3%)	17	(6.5%)	260
Weekend									
No	465	(77.8%)	58	(9.7%)	55	(9.2%)	20	(3.3%)	598
Yes	147	(89.1%)	4	(2.4%)	13	(7.9%)	1	(0.6%)	165
Number of Vehicles Involved									
1	178	(89.9%)	2	(1.0%)	17	(8.6%)	1	(0.5%)	198
2	367	(77.6%)	51	(10.8%)	41	(8.7%)	14	(3.0%)	473
3+	67	(72.8%)	9	(9.8%)	10	(10.9%)	6	(6.5%)	92
Number of Lanes									
2	333	(89.8%)	5	(1.3%)	29	(7.8%)	4	(1.1%)	371
3	174	(72.2%)	33	(13.7%)	25	(10.4%)	9	(3.7%)	241
4	65	(67.7%)	17	(17.7%)	7	(7.3%)	7	(7.3%)	96
5	40	(72.7%)	7	(12.7%)	7	(12.7%)	1	(1.8%)	55
Crash Severity									
PDO	421	(81.3%)	51	(9.8%)	35	(6.8%)	11	(2.1%)	518
Severe	191	(78.0%)	11	(4.5%)	33	(13.5%)	10	(4.1%)	245
Weather Condition									
Clear	398	(81.1%)	45	(9.2%)	37	(7.5%)	11	(2.2%)	491
Cloudy	111	(79.9%)	9	(6.5%)	13	(9.4%)	6	(4.3%)	139
Rain/Fog	103	(77.4%)	8	(6.0%)	18	(13.5%)	4	(3.0%)	133
Total	612	(80.2%)	62	(8.1%)	68	(8.9%)	21	(2.8%)	763

TABLE 3 Pearson's Chi-square Correlation Test for Variables

Chi-square (P-value)	Expressway	Weekend	Peak Hour	Vehicles Involved	Crash Severity	Number of Lanes	Weather Condition
Expressway	--	4.041 (0.1326)	11.508 (0.0032)	42.238 (<.0001)	3.763 (0.1524)	148.933 (<.0001)	9.407 (0.0091)
Weekend		--	15.508 (<.0001)	14.556 (0.0001)	0.323 (0.5697)	0.123 (0.7262)	0.086 (0.7691)
Peak Hour			--	13.695 (0.0002)	1.927 (0.1650)	1.582 (0.2084)	2.820 (0.0931)
Vehicles Involved				--	0.018 (0.8936)	15.340 (<.0001)	0.110 (0.7397)
Crash Severity					--	0.1746 (0.6761)	1.686 (0.1942)
Number of Lanes						--	3.787 (0.0517)
Weather Condition							--

TABLE 4 Parameter Estimates and Model Fitting for Multinomial Logit Model

	$\log(P_{Type2}/P_{Type1})$			$\log(P_{Type3}/P_{Type1})$			$\log(P_{Type4}/P_{Type1})$		
	Mean	Std. Errors	95% BCI	Mean	Std. Errors	95% BCI	Mean	Std. Errors	95% BCI
Intercept	-5.379	0.564	(-6.497, -4.358)	-3.452	0.329	(-4.094, -2.823)	-7.026	0.854	(-8.932, -5.458)
Peak: Yes vs. No	2.252	0.332	(1.629, 2.956)	1.057	0.276	(0.484, 1.555)	2.755	0.618	(1.635, 4.069)
Lanes: 3 vs. 2	2.896	0.518	(1.971, 4.022)	0.631	0.308	(0.016, 1.241)	1.884	0.670	(0.735, 3.414)
Lanes: 4 vs. 2	2.970	0.559	(1.945, 4.134)	0.226	0.477	(-0.772, 1.106)	2.364	0.716	(1.081, 3.877)
Lanes: 5 vs. 2	2.609	0.673	(1.327, 3.977)	0.742	0.494	(-0.215, 1.672)	0.540	1.450	(-2.744, 3.027)
Weather: Cloudy vs Clear	-0.143	0.453	(-1.04, 0.700)	0.402	0.355	(-0.341, 1.067)	1.049	0.601	(-0.120, 2.234)*
Weather: Rain vs Clear	-0.246	0.445	(-1.145, 0.613)	0.664	0.327	(0.023, 1.282)	0.619	0.655	(-0.746, 1.791)
Severity: Severe vs PDO	-0.720	0.377	(-1.479, -0.004)	0.777	0.270	(0.269, 1.304)	0.839	0.500	(-0.193, 1.881)
Model Performance									
$\overline{D}$					907.893				
p_D					24.016				
DIC					931.909				
AUC					0.715				

*significant at 90% BCI

The number of lanes at the crash location was also found positively related with the log odds ratio. From a demand-capacity point of view, number of lanes on expressway reflects the traffic demand on the segment. More lanes indicate higher traffic load and potential congestion. Thus the effects of number of lanes on Type 2 and Type 4 crashes are understandable. Type 3 crashes occurred under non-congested before crash conditions. If the cross section has only two lanes, single vehicle crashes are more likely to occur. The vehicles can move to shoulder or median after crashes and reduce their impact on the upstream traffic. In contrast, if the crashes occurred in the middle of a cross section with more than 2 lanes, the probability of multi-vehicle crashes would be higher and cause more severe delay before they could be moved out of the roadway.

Weather conditions in the crash report have three categories. Compared with clear weather condition, cloudy weather won't significantly alter the log odds ratio of crashes. Rainy/fog conditions, nevertheless, will greatly increase the probability of congestion after crashes under non-congested conditions. Rain and fog can significantly impair drivers' visibility and the friction between pavement and tires. Once a crash occurs under these weather conditions, the severity level might be high and the adverse weather can extend the time needed to clear the scene. Under congestion, the adverse weather's effect is not significant since the speed of vehicles is expected to be low. Less severe crashes are expected under congested conditions. Thus the impact of adverse weather on congestion might be limited in this situation.

The crash severity also exhibits distinct effects on different crash types. If the traffic conditions before crashes are congested and the crashes do not worsen congestion, their severities would be much lower compared with Type 1 crashes. However, under non-congested conditions, if the crashes are severe, they have significant higher chance to result in congestion. One should understand that the speed under congested or non-congested traffic prior to crashes would mean quite different severity levels; and the severity levels will partially determine congestion status after crashes together with other factors.

Binomial Logit Model Results

The multinomial model in the above section provided relatively comprehensive and sound conclusions about the effects of crashes on congestion. However, the interpretation of the factors often involves differentiating congestion conditions prior to crashes first. To gain more clear understanding and relaxing the IIA assumption of the multinomial logit model, two separate binomial logit models based on the congestion status before crashes were constructed. Table 5 displays the modeling results.

Both Type 1 and Type 3 crashes had non-congested before crash conditions. The effects of the variables were the same as those found in the multinomial model. For Type 2 and Type 4 crashes which had the congested before crash conditions, the results shed some lights not revealed by the multinomial model. First of all, only the number of lanes and severity were found to significantly influence the probability of these two types of crashes. Since the congested traffic was mostly due to peak hour traffic before Type 2 and Type 4 crashes, the peak hour indicator would not play a crucial rule classifying these two crash types. For vehicles moving in congestion, speed has already been reduced. The effects of weather conditions on traffic flow parameters would be limited. The effects of number of lanes are worth elaboration. In contrast to the findings regarding Type 1 and Type 3 crashes, more lanes can efficiently reduce the impact of crashes under congested conditions. Crashes under congestion are more likely to involve multiple vehicles and block the traveling lanes. If the crash spot has more lanes, other vehicles can use adjacent lanes and avoid total shutdown of the mainline. The effects of crash severity do not differ for non-congested and congested conditions. The severe crashes will worsen the congestion conditions significantly in both cases.

TABLE 5 Parameter Estimates and Model Fitting for Separate Binomial Logit Model

	$\log(P_{Type3}/P_{Type1})$			$\log(P_{Type4}/P_{Type2})$		
	Mean	Std. Errors	95% BCI	Mean	Std. Errors	95% BCI
Intercept	-3.465	0.311	(-4.110, -2.901)	--	--	--
Peak: Yes vs. No	1.054	0.260	(0.558, 1.580)	--	--	--
Lanes: 3 vs. 2	0.634	0.315	(0.039, 1.261)	-2.013	0.517	(-3.087, -1.094)
Lanes: 4 vs. 2	0.160	0.470	(-0.864, 1.011)	-1.332	0.508	(-2.335, -0.403)
Lanes: 5 vs. 2	0.838	0.479	(-0.168, 1.755)*	-2.704	1.322	(-5.663, -0.645)
Weather: Cloudy vs Clear	0.340	0.363	(-0.394, 1.066)	--	--	--
Weather: Rain vs Clear	0.691	0.340	(0.037, 1.317)	--	--	--
Severity: Severe vs PDO	0.795	0.272	(0.272, 1.323)	1.587	0.601	(0.431, 2.781)
Model Performance						
$\bar{D}$	419.943			88.288		
p_D	8.101			3.938		
DIC	428.044			92.227		
AUC	0.686			0.728		

*significant at 90% BCI

Since the multinomial and binomial logit models employed different data, direct comparison via DIC is not appropriate. Area under the ROC Curve (AUC) was calculated to evaluate the performances of the models. The AUC values were all about 0.7, meaning the overall performances highly comparable. Both the multinomial and separate binomial models answered our second question regarding how the crashes could have distinct effects on congestion. According to the structure of crash classification, the separate binomial logit models generate results slightly easier for understanding and unveil distinct effects of the contributing factors on different types of crashes.

CONCLUSIONS

Traffic safety and operation are major indicators of highway performance. A large body of literature has investigated the operation-safety relationship. The current study focuses on one issue that was overlooked by previous studies: how crashes lay their impact on traffic congestion. To answer the question, three expressways managed by CFX in Central Florida area were investigated. Detailed information from crash reports and real-time traffic data from MVDS system on the expressways were extracted.

The crashes were first clustered according to their effects on congestion. Traffic congestion status before and after each crash case were matched with the crash data. Machine Learning (K-means) method was tested to partition the data. However, the clustering results didn't offer reasonable insights into the crashes effects. As a solution, the crashes were manually classified. Four types of the effects of crashes were identified for further analysis. Based on the real-time traffic data, it was found that not all of the crashes would lead to congestion on the urban expressways. Both crashes occurring under congested or non-congested traffic flow could either increase the congestion intensity afterwards or exert insignificant influence on upstream traffic. To understand the distinct effects, information from crash reports were applied in statistical analysis.

Since the target of the statistical analysis is classification, logit models under Bayesian framework were constructed. Considering the structure used to cluster these crashes, both multinomial logit model and two separate binomial logit models were tested. Seven candidate variables that were possibly pertinent to crash effects were prepared. All of the candidate variables were categorical and Pearson' Chi-square test was conducted. Peak hour indicator, crash severity, number of lanes, and weather conditions were retained in the logit models. All of the four variables were found significant in the multinomial model and the binomial model for uncongested conditions. The separate binomial models generated results easier for explanation. Under non-congested before crash conditions, the peak hours suggest higher traffic load. The crashes during peak hours would also pose more significant impact on the traffic congestion. If the roadway experiences no congestion prior to crashes, then the more lanes the segment has, the higher probability that crashes occurring on it would lead to congestion. It could be explained as when a crash occurs in the middle lanes of a cross section, it is possible to involve multiple vehicles and hence cause traffic congestion. With non-congested traffic flow before crashes, adverse weather could significantly increase the probability of congestion after crashes. On the contrary, in the binomial model for congested before crash conditions, only the number of lanes and crash severity were significant. The effects of crash severity are the same as that in uncongested conditions. However, the number of lanes shows different impact. Under congestion, in the face of crash occurrence, more lanes suggest less probability of congestion. Traveling speed will be greatly reduced by congestion, and therefore the crash manner. It is expected that single-vehicle crashes due to driving error or distraction would reduce under the congested conditions while the probability of multi-vehicle crashes greatly increases. In this case, more traveling lanes imply that motorists can use alternative lanes and avoid total shutdown of the mainline. As a result, traffic authorities should be careful when they interpret the crash effects on safety. The traffic state prior to crashes should be taken into account.

Potential improvement for future emergency response strategies can be raised based on the findings of this research. First of all, the real-time ITS traffic data should be incorporated in the response procedure. To estimate if the reported crashes would deteriorate congestion, current traffic condition at the crash site should be referred to. For less-instrumented freeways, the real-time traffic information might not be available. In such cases, general congestion conditions and congestion time based on historical data may be leveraged to decide the most likely traffic conditions at the crash sites. Second, the time of the crash, weather conditions at that time and the geometric characteristics of the crash location should all be considered. More police patrols might be helpful during peak hours or under adverse weather conditions. As for the segments with multiple lanes, their effects should be evaluated based on congestion levels. Last but not least, the necessity to report the potential severity of a crash has been confirmed. The quicker response for severe crashes can effectively diminish their effects on congestion.

ACKNOWLEDGMENT

The authors wish to thank the Central Florida Expressway Authority (CFX) for funding this research and CFX for providing the data. Partial funding was also provided by Southeastern Transportation Center (STC), the regional UTC Lead by the University of Tennessee. The authors also wish to thank Charles R. Lattimer, PE, PMP Sr. Project Manager, Intelligent Transportation Systems at ATKINS for his suggestions.

REFERENCE

[1] Baruya, A. Speed-accident relationships on European roads.In *9th International Conference on Road Safety in Europe*, 1998.

[2] Hossain, M., and Y. Muromachi. A Bayesian network based framework for real-time crash prediction on the basic freeway segments of urban expressways. *Accident Analysis & Prevention,* Vol. 45, 2012, pp. 373-381.

[3] Kononov, J., B. Bailey, and B. K. Allery. Relationships between safety and both congestion and number of lanes on urban freeways. *Transportation Research Record: Journal of the Transportation Research Board,* Vol. 2083, No. 1, 2008, pp. 26-39.

[4] Wang, C., M. A. Quddus, and S. G. Ison. Impact of traffic congestion on road accidents: A spatial analysis of the M25 motorway in England. *Accident Analysis & Prevention,* Vol. 41, No. 4, 2009, pp. 798-808.

[5] Wang, C., M. Quddus, and S. Ison. A spatio-temporal analysis of the impact of congestion on traffic safety on major roads in the UK. *Transportmetrica A: Transport Science,* Vol. 9, No. 2, 2013, pp. 124-148.

[6] Shefer, D., and P. Rietveld. Congestion and safety on highways: towards an analytical model. *Urban Studies,* Vol. 34, No. 4, 1997, pp. 679-692.

[7] Golob, T. F., W. W. Recker, and J. D. Leonard. An analysis of the severity and incident duration of truck-involved freeway accidents. *Accident Analysis & Prevention,* Vol. 19, No. 5, 1987, pp. 375-395.

[8] Sullivan, E. C. New model for predicting freeway incidents and incident delays. *Journal of Transportation Engineering,* Vol. 123, No. 4, 1997, pp. 267-275.

[9] Garib, A., A. Radwan, and H. Al-Deek. Estimating magnitude and duration of incident delays. *Journal of Transportation Engineering,* Vol. 123, No. 6, 1997, pp. 459-466.

[10] Jones, B., L. Janssen, and F. Mannering. Analysis of the frequency and duration of freeway accidents in Seattle. *Accident Analysis & Prevention,* Vol. 23, No. 4, 1991, pp. 239-255.

[11] Smith, K. W., and B. Smith. Forecasting the clearance time of freeway accidents. PhD diss., University of Virginia, 2001.

[12] Junhua, W., C. Haozhe, and Q. Shi. Estimating freeway incident duration using accelerated failure time modeling. *Safety Science,* Vol. 54, 2013, pp. 43-50.

[13] Sethi, V. Duration and travel time impacts of incidents. The Transportation Center, Northwestern University, 1994.

[14] Xiaoqiang, Z., L. Ruimin, and Y. Xinxin. Incident duration model on urban freeways based on classification and regression tree.In *Intelligent Computation Technology and Automation, 2009. ICICTA'09. Second International Conference on, No. 3,* IEEE, 2009. pp. 625-628.

[15] Hamad, K., and S. Kikuchi. Developing a measure of traffic congestion: fuzzy inference approach. *Transportation Research Record: Journal of the Transportation Research Board,* Vol. 1802, No. 1, 2002, pp. 77-85.

[16] Xu, C., P. Liu, W. Wang, and Z. Li. Evaluation of the impacts of traffic states on crash risks on freeways. *Accident Analysis & Prevention,* Vol. 47, 2012, pp. 162-171.

[17] Oltedal, S., and T. Rundmo. Using cluster analysis to test the cultural theory of risk perception. *Transportation research part F: traffic psychology and behaviour,* Vol. 10, No. 3, 2007, pp. 254-262.

[18] Anderson, T. K. Kernel density estimation and K-means clustering to profile road accident hotspots. *Accident Analysis & Prevention,* Vol. 41, No. 3, 2009, pp. 359-364.

[19] Hartigan, J. A., and M. A. Wong. Algorithm AS 136: A k-means clustering algorithm. *Applied statistics*, 1979, pp. 100-108.

[20] Anderson, S. P., A. De Palma, and J. F. Thisse. *Discrete choice theory of product differentiation*. MIT press, 1992.

[21] Spiegelhalter, D., Thomas, A., Best, N., Lunn, D. Winbugs user manual. Cambridge: MRC Biostatistics Unit, 2003.

PREDICTING CRASHES ON EXPRESSWAY RAMPS WITH REAL-TIME TRAFFIC AND WEATHER DATA

Ling Wang*
Department of Civil, Environmental and Construction Engineering
University of Central Florida
Orlando, Florida 32816-2450
(407)823-0300
lingwang@knights.ucf.edu

Qi Shi
Department of Civil, Environmental and Construction Engineering
University of Central Florida
Orlando, Florida 32816-2450
(407)823-0300
shiqi@knights.ucf.edu

Mohamed Abdel-Aty
Department of Civil, Environmental and Construction Engineering
University of Central Florida
Orlando, Florida 32816-2450
(407)823-4535
M.Aty@ucf.edu

Peifen Kuo
Department of Civil, Environmental and Construction Engineering
University of Central Florida
Orlando, Florida 32816-2450
(407)823-0300
peifenkuo@gmail.com

* Corresponding Author

October 2014

Revised manuscript for presentation at the *Transportation Research Board* Annual Meeting and possible publication in a *Transportation Research Record*

Word count: 5,789 words text + 6 tables/figures x 250 words (each) = 7,289 words

ABSTRACT

Very limited research have been conducted on real-time crash analysis of expressway ramps, although there have been many studies in recent years on estimating real-time crash prediction models for mainlines. This study presents Bayesian logistic regression models for single-vehicle (SV) and multi-vehicle (MV) crashes on expressway ramps using real-time Microwave Vehicle Detection System (MVDS) data, real-time weather data, and ramp geometric information. The results find that the Logarithm of vehicle count, average speed in a 5-minute interval, and visibility are significant factors for the occurrence of SV and MV crashes. The Bayesian logistic regression models show that curved ramps and wet road surfaces would increase the possibility of an SV crash, and off-ramps would result in high MV crash risk. The high standard deviation of speed in a 5-minute interval would significantly increase MV crash likelihood. Random forest is applied in variable importance analysis, and the result reveals that the most important factors influencing crashes on ramps are traffic variables, the second most important factors are weather variables, and the least important but still significant factors are the ramp geometry.

Keywords: real-time crash prediction, expressway ramps, Bayesian logistic regression, MVDS data, real-time weather data, random forest, variable importance analysis

1 INTRODUCTION

Very limited research has been conducted on real-time crash prediction for expressway ramps. Ramp and mainline characteristics are not the same, e.g., ramps may have smaller radii or are steeper than mainlines. These differences result in different crash mechanisms (*1*). Meanwhile, it is important to analyze crashes by type, particularly in real-time risk assessment (*2; 3*), and two most important sub-groups are single-vehicle (SV) and multi-vehicle (MV) crashes. Hence, there is a need to build separate crash risk models for SV and MV crashes on ramps.

In general, primary crash factors are environmental, traffic, vehicle, and driver (*4*). The former two factors are more important and can be collected easier compared with the latter two. Environmental factors include geometric design and weather. Traffic factors include volume, speed, lane occupancy, etc. Among these environmental and traffic variables, weather has not been universally studied in real-time crash prediction, though it is an important factor. On average, from 2002 to 2012 in the United States, twenty-three percent (23%) of crashes were weather-related, and seventy-four percent (74%) of weather-related crashes happened on wet pavement (*5*). Meanwhile, weather-related crashes caused 94 million to 272 million hours of delay each year (*6*). As a result, in addition to traffic factors, weather also should be addressed in crash risk prediction.

The study presented in this paper has two basic objectives: 1) to explore factors contributing to crashes on expressway ramps; 2) to develop Bayesian logistic regression models for real-time ramp crash likelihood. The studied ramp area in this study is the area between the painted gore point and ramp terminal intersections at crossroads, and the ramp crashes do not include the crashes at ramp terminal intersections.

This paper is organized into six sections. The second section reviews the previous studies and findings on real-time crash prediction. The third section describes the research methodology. The fourth section discusses the data used in this paper and the descriptive and exploratory analysis of data. The fifth section shows the results of the model estimation and the variable importance analysis. Finally, the sixth section summarizes the findings and the applications of this study.

2 BACKGROUND

Since 1995, there have been numerous studies on real-time crash prediction models which have linked real-time crash likelihood with traffic flow characteristics, e.g., volume, speed, lane occupancy, and weather.

Oh et al. (*4*) applied the non-parametric Bayesian method to determine whether the speed variation in 5-minute intervals was a good indicator of crashes. They compared the normal condition and the disruptive traffic condition and found the standard deviation of speed was the most significant variable. Abdel-Aty et al. (*7*) developed a matched case-control logistic regression model to link crash risk with traffic turbulence while controlling for road geometry, day of week and time of day. The results showed that crash risk was associated with the upstream average lane occupancy and downstream variation of speed, and both variables were 5-10 minutes ahead of crashes. A Bayesian matched case control logistic regression model was used by Abdel-Aty et al. (*8*) to compare the accuracy of visibility-related crash prediction models which used AVI data and the loop detector data separately. The results showed that loop detector data were better than the AVI data, and average speed and speed variation which were 5-10 minutes ahead of crashes were significant. Hossain and Muromachi (*9*) built a Bayesian belief net (BBN) model to predict the crashes that occurred at ramp vicinities. It was concluded that 5-

minute upstream volume and congestion index, and 5-minute downstream volume and speed, along with ramp volume, were significant variables affecting crashes. Hossain and Muromachi (*10*) also built a model for basic freeway segment. Later, Hossain and Muromachi (*11*) estimated real-time crash prediction models for basic segment and ramp vicinities in one paper. Xu et al. (*12*) predicted the crash likelihood at different levels of crash severity with a sequential logit model and elasticity analysis. The finding showed that different crashes had different precursors, e.g. congested traffic flow with a high speed variance and frequent lane changes would significantly increase the property-damage-only (PDO) crash rates. Zheng et al. (*13*) studied the impact of traffic oscillations, which is also known as stop-and-go driving, on freeway crashes in real-time. The matched case control model showed that the deviation of speed was a significant variable, which had positive impact on crash occurrence.

In addition to traffic variables, weather variables were also studied in real-time crash prediction research. Madanat and Liu (*14*) first used traffic and weather data to develop a binary logit prediction model to predict crashes in real-time. However, the traffic parameters were not significant in their model, and they found that visibility and rain would affect crash occurrence. Lee et al. (*15*) built an aggregate log-linear model to estimate the frequency of crashes in 5 minutes. They concluded that the significant predictors of crashes were: weather condition, the speed variation along the section, the speed difference across lanes, and traffic density. The weather condition was a binary variable indicating whether it was severe or not in their paper. Abdel-Aty and Pemmanaboina (*16*) added the hourly rainfall information in a matched case-control logit model for crash prediction. Ahmed et al. (*17*) first used airport weather data in real-time crash risk assessment based on Bayesian logistic regression. The results indicated that airport weather information was valid. However, the traffic variable used in model building was AADT which could not sufficiently represent the real-time traffic turbulence, and visibility was the only weather factor in the model. Zoi et al. (*18*) also entered the binary weather condition and lighting condition in the model and found that the crash type can be predicted by the traffic and other conditions shortly before its occurrence on freeways, e.g., multi-vehicle sideswipes crashes are related to high speeds, daytime and flat freeways. Almost all former research didn't include visibility and rainfall at the same time in a model.

The previous studies on real-time crash prediction are valuable, the results demonstrated that the crash risk on mainlines was affected by the average or variance of speed, and/or volume in a 5-10 minute period at the upstream and/or downstream stations, and/or the visibility, etc. Nonetheless, it cannot be directly applied to or transferred to the prediction of real-time crashes on ramps. First, few of the former research incorporated geometric variables into the models. They excluded it since the geometry did not change significantly on the mainline, but the ramp geometry was more site-specific, i.e., different ramp types (e.g., on or off-ramps) and configurations (e.g., diamond, loop, etc.). Second, weather didn't play an important role in real-time crash prediction and usually only contained visibility or rain information.

From the discussion above, it is not difficult to conclude that it may be advantageous to work on ramp crashes. Several papers focused on ramp crash frequency, e.g., Lord and Bonneson (*19*) and Garnowski and Manner (*20*). There have been very few studies on real-time ramp crash prediction. Lee and Abdel-Aty (*21*) estimated the risk of crashes on freeway ramps and at ramp intersections using log-linear models. They found that higher volumes and lower speeds would result in higher crash risk. They also found that crash rates on loop and outer connection ramps were higher than on diamond ramps. Despite their innovative approach, their study had some limitations. First, the ramp traffic explanatory variables in the models were daily ramp volume or estimated hourly ramp volume, which cannot effectively represent the real-time traffic condition of the ramp. Second, there was no weather information in the model. Yet, as we stated at the beginning of this paper, weather would significantly influence the ramp crash occurrence.

3 RESEARCH METHODOLOGY

This study built Bayesian Logistic Regression models to estimate ramp crash likelihood. The traditional and standard logistic regression models treat the variable coefficients as fixed values. However, the Bayesian model assumes that there are distributions for the coefficients. It also makes use of the knowledge gained from observations to update the behavior of the coefficients and then assess their distributional properties. Furthermore, the Bayesian inference can effectively avoid the overestimated odds ratio which occurs when the sample size is limited (*22*).

In this study, Bayesian logistic regression models were used to estimate the relationship between the binary response variable (crash or non-crash) and explanatory variables. The binary responses, crash and non-crash, were converted into the probabilities p(y=1) and 1-p, respectively. Bayesian logistic regression models are as follows,

$$\log\left(\frac{p}{1-p}\right) = \beta_0 + \beta_j x_j$$

Where β_0 is the intercept, x_j is the value of explanatory variable, and β_j is the coefficient of x_j. A common choice for the β_0 and β_j distribution is normal distribution (*23*):

$$\beta_0 \sim N(\mu_0, \sigma_0^2)$$

$$\beta_j \sim N(\mu_j, \sigma_j^2)$$

In general, there are three kinds of prior distribution depending on the availability of prior information. Informative prior distribution is used if the possible values of coefficients are known. When little or nothing is known about the coefficient values, or if the author wishes to know what will the data provide as inferences, the vague prior or non-informative priors are preferred. In this research, the non-informative priors, which follow normal distribution, are used. The following are their form:

$$\beta_0 \sim N(0, 10^6)$$

$$\beta_j \sim N(0, 10^6)$$

All real-time ramp crash prediction models were estimated by Bayesian inference which was carried out by Winbugs in R (*24; 25*). In each model, there were three chains of 10,000 iterations. In order to eliminate the concern that early values didn't represent the true posterior distribution, the first half of the simulation iterations were discarded (*23*).

The deviance information criterion (DIC) was widely used for the Bayesian model selection. The model with the smallest DIC stands for the model that would best predict a replicate dataset that has the same structure as the current sample (*26*). Finally, analysis of the Area Under the Curve (AUC) was also used to compare and select the possibly optimal models.

4 EXPERIMENTAL DESIGN AND DATA DESCRIPTION

4.1 Experimental Design

To accomplish the study objectives, the researchers chose the following three expressways: State Roads 408 (SR408), 417 (SR417), and 528 (SR528), all located in Central Florida. About 14.2 miles of SR408, 26.9 miles of SR417 and 7.6 miles of SR528 are covered by Orlando International Airport's (MCO) and Orlando Executive Airport's (ORL) 7.0 miles coverage buffer.

Within this buffer, the airport weather equipments were used to provide sufficiently accurate weather information for the crash and non-crash observations (*17*).

In order to reduce data noise, the traffic data were aggregated into 5-minute intervals. The researchers extracted the traffic data which were 0-5 minute and 5-10 minute prior to crash and non-crash cases. For example, if a crash occurred at 8:00AM, the traffic data extracted were from 7:55 to 8:00AM and from 7:50 to7:55AM. The traffic data which were 5-10 minutes prior to cases provided better model performance, and also increased the practical application of the model by providing sufficient time for the traffic management center to analyze, react and announce warning information to the drivers. In the following parts, the traffic data utilized are 5-10 minutes prior to the crash cases.

To generate the non-crash observations, we used SAS to select 0.05% of the 11,207,808 (12interval×24hours×276days×141ramps) 5-minute intervals randomly. At the same time, if any crash had happened within 2 hours from the time of a non-crash data point then this non-crash data point would be excluded to ensure the purity of the non-crash traffic flow data.

4.2 Data Description and Combination

The data collected in this study are as follows: detailed information for every crash, traffic flow data, ramp geometric properties and weather information. The definitions and acronyms of these variables are shown in Table 1. Their detailed information is described below.

TABLE 1 Variables Considered for the Model

Data	Symbol	Description
Traffic Flow	Spd	Average speed in a 5-minute interval (mile/h)
	Std_spd	Standard Deviation of speed in a 5-minute interval (mile/h)
	Vehcnt	Vehicle count in a 5-minute interval (veh/5minutes)
	Occ	Average lane occupancy in a 5-minute interval (%)
	Std_occ	Standard Deviation of occupancy in a 5-minute interval (%)
	P_truck	Percentage of trucks in a 5-minute interval (%)
Ramp Geometric	Type	1=if the ramp is an off-ramp; 0=otherwise
	Configuration	1=if the ramp is a diamond-ramp; 0=otherwise
	Toll	1=if there is a toll booth on the ramp; 0=otherwise
	Length	Ramp length which is from the painted gore point to the intersection of ramp and street
Weather	Visibility	The distance at which an object or light can be clearly discerned (mile)
	Surface	1=if the road surface condition is wet; 0=otherwise

Crash data: The raw ramp crash data were obtained from Signal Four Analytics. The dataset contained detailed information for all reported crashes in the time period from July 2013 to March 2014. The information included: exact time of crash, crash coordinate, crash street and intersecting street, number of vehicles involved, type and severity of the crash, number of injuries and/or fatalities involved, weather, road surface and light condition, etc. Seventy nine SV crashes and 58 MV crashes were identified and data collected. Among the SV crashes, sixty four (81%) were off road crashes, nine (11%) were rollover crashes and six (8%) crashes were missing the type information. As for the MV crashes, there are forty five rear end crashes (78%), ten (17%) sideswipe crashes, three (5%) crashes with unknown type.

Traffic flow data: The traffic flow data were provided by the Central Florida Expressway Authority (CFX). The traffic data, e.g., volume, speed and lane occupancy, were calculated every minute automatically by MVDS, which additionally recognizes the length of passing vehicles and categorizes them under four groups: the vehicles which are less than 12 feet long belong to group 1; between 12 and 24 ft to group 2; between 24.1 and 40 feet to group 3; and greater than 40 feet to group 4. In our research, we used the term passenger cars for groups 1, 2, and 3, and trucks for group 4. There were 124 MVDS detectors along the selected expressways in the study area with an average spacing of 0.785 miles.

Geometric data: The geometric data of ramps were collected manually by using ArcGIS map. There were 141 ramps, and each ramp had four variables: ramp type, ramp configuration, the presence of a toll booth, and ramp length. Seventy out of the 141 ramps were off-ramps, 71 were diamond ramps, and 39 with a toll booth, the mean of ramp length was 0.347 miles. Nearly every ramp had one MVDS used to collect traffic flow data.

Weather data: Airport weather data were collected from the National Climate Data Center (NCDC). The weather data were monitored continuously, and if the weather parameters did not change, the data would be recorded every one hour. When the weather parameters changed, the weather station would record the new weather state. The dataset included: sky condition, weather type, wind direction and speed, pressure, humidity, temperature, visibility and hourly precipitation. Visibility, weather type and hourly precipitation were used in this study.

Integrating crash, traffic, geometric and weather data together was an important part in this study. Every ramp was first assigned an ID for the geometric data. Then, for every crash, we manually added an ID variable, which was the same as the ID in geometric data and would stand for the ramp where crash happened. All traffic data at the same ramp would have a same ID which was the same ID in geometric and crash data. Based on this ID variable, crash, geometric and traffic data were combined, and the traffic data were then integrated into 5-min interval. The last step was adding weather data into the former combined data. Ramps' weather data was from the airport which was closest to ramps. As for the visibility parameter, all crash and non-crash cases were matched with the visibility data whose time was the closest prior to the data point. As for the road surface condition parameter, if hourly precipitation was higher than zero, or weather type contained TS (thunderstorm), RA (rain) and so on, we assumed that the road surface condition of all crash and non-crash cases was wet in this following hour.

Two hundred and eleven crashes and 5603 non-crash cases were filtered out in the study area. However, thirty four crashes were deleted because of the absence of clear location information. Combining all the datasets together produced 137 crash observations and 4907 non-crash observations. Each of them contained complete traffic flow, ramp geometric, and weather information. Two thousand eight hundred and thirty non-crash observations were randomly assigned to SV and 2077 to MV crashes. As a result, the sample size of SV crashes model-building dataset was 2909, and the sample size of MV crashes model-building dataset was 2135.

4.3 Descriptive and Exploratory Analysis

Table 2 summarizes the continuous variables' descriptive statistics for SV and MV crashes. The t-test shows that there is no significant difference for six variables, i.e., speed, Logarithm of vehicle count, occupancy, truck percentage and ramp length. Yet, the speed standard deviation of MV crashes is significantly higher than that of SV crashes at 90% confidence interval. This indicates that when the speed fluctuation is small, it is less likely to have MV crashes. Meanwhile, the mean of visibility of SV crashes is significantly less than that of MV crashes at the 95% confidence interval. These findings confirm that estimating SV and MV crash prediction models separately would be helpful in exploring the specific variables' impact on different crash types.

TABLE 2 Summary of Continuous Variables' Descriptive Statistics for Crashes

Variables	**Spd**		**Std_spd**		**Log (Vehcnt)**		**Occ**		**Std_occ**		**P_truck**		**Visibility**		**Length**	
	SV	MV	SV	MV	SV	MV	SV	MV	SV	MV	SV	MV	SV	MV	SV	MV
Mean	56.5	56.5	3.3	4.5	3.2	3.2	3.7	3.8	1.6	1.5	0.03	0.06	4.4	7.9	0.5	0.4
Std Dev	6.3	7.8	2.2	5.2	0.8	0.7	3.6	3.9	1.8	1.1	0.05	0.16	3.9	3.5	0.5	0.3
Min	36.4	29.4	0.3	0.0	1.4	1.4	0.0	0.0	0.0	0.0	0.00	0.00	0.0	0.0	0.1	0.1
Max	66.3	82.0	11.1	29.9	5.1	4.4	22.0	28.1	12.5	5.3	0.36	1.00	10.0	10.0	1.7	1.7
t-value	-0.00		-1.67		0.36		-0.02		0.39		-1.30		-5.43		1.27	
p-value	0.9997		0.0992		0.7192		0.9855		0.7008		0.1971		<0.0001		0.2062	

In Table 3, the significant difference between SV and MV crash numbers for different ramps and road surface conditions is notable, and suggests the existence of heterogeneity across the geometric characteristics and weather types. This also demonstrates that dividing crashes into SV and MV crashes is better for model estimation.

TABLE 3 Exploratory Statistics of Crashes on Categorical Variables

Variables	**SV crash**	**MV crash**	**Chi-square**	**P-value**
Ramp type				
On-ramp	47	16	13.7084	0.0002
Off-Ramp	32	42		
Ramp configuration				
Curved ramp	71	31	23.3297	<.0001
Diamond ramp	8	27		
Toll				
No toll booth on ramps	70	47	1.5385	0.2148
With toll booth on ramps	9	11		
Road surface condition				
Dry	13	41	41.1952	<.0001
Wet	66	17		

The Chi-square and p-value of these four contingency tables show that ramp type, ramp configuration, and road surface condition play significant roles in determining crash type. The odds ratio of MV crashes for on-ramps relative to off-ramps was 0.259. It could be inferred that MV crashes are more likely to happen at off-ramps than SV crashes. Vehicles at off-ramps need to decelerate to accommodate the speed on streets, so a rear-end crash may occur if the following vehicle doesn't decelerate in time. The odds ratio of SV crash at a curved ramp relative to diamond ramp was 7.730, on a wet surface relative to dry surface were 12.244. These suggest that SV crashes are more likely to happen on curved ramps or/and wet surface ramps, because the chance of a vehicle skidding off the road and being involved in a SV crash would increase significantly on these ramps.

In the case of toll booths, their presence on a ramp is not statistically related to the crash type. The reasons for this are straightforward. All toll booths are located at the end of off-ramps or at the beginning of on-ramps. Driving speed at these locations is low, and thus drivers are in a high level of control. At the same time, around 90% of vehicles at these expressways will use Electronic Toll Collection (E-pass) at the toll booths and will not stop. There will be no significant deceleration, and the opportunity for the following vehicle to run into the leading vehicle does not increase. Thus, the existence a toll booth has a minor or insignificant influence on crashes. Furthermore, the absence of toll booths will not influence the crash types.

5 MODEL ESTIMATION AND VARIABLE IMPORTANCE

As mentioned earlier, the objective of this paper is to estimate the relationship between the likelihood of a ramp crash and the variables of traffic, weather, and geometrics, while distinguishing different crash sub-groups. Two Bayesian logistic regression models were built, one was a real-time SV crash prediction model, and the other was a real-time MV crash prediction model. Both SV and MV crash model-building datasets were split into training and validation datasets with a ratio of 70:30.

In order to prevent high correlation between traffic predictors for SV and MV crash prediction models, Pearson correlation test was done before the model building. The result shows that, for both SV and MV crashes, occupancy is correlated with Logarithm of vehicle count, speed and speed standard deviation, the absolute of correlation coefficient value is higher than 0.3. Meanwhile, standard deviation of occupancy is also correlated with truck percentage, speed and Logarithm of vehicle count for both SV and MV crashes. Meanwhile, in SV crashes, standard deviation of speed is highly correlated with speed with a -0.45 correlation coefficient. Hence, in the real-time SV crash model, only Logarithm of vehicle count, speed and truck percentage were taken into consideration; in the MV crash model, Logarithm of vehicle count, speed, standard deviation of speed and truck percentage were inputs in the model.

5.1 Real-time Single-Vehicle Crash Model

Estimation results for the real-time SV crash-prediction model are shown in Table 4. Five variables are found to be significant in the model at the 95% confidence interval. AUC area for training and validation are 0.9346 and 0.9710, respectively. The overall accuracy for training and validation are 0.8900 and 0.9049, respectively, when the cutoff-point is 0.020. These results demonstrate that the model's predictive accuracy for discriminating between crashes and non–crashes is excellent.

TABLE 4 Real-time SV Crash Prediction Model

Node	Mean	Std	2.50%	97.50%
Intercept	-8.805	2.113	-13.4	-5.14
Log(Vehcnt)	0.9588	0.2619	0.4411	1.507
Spd	0.06087	0.02652	0.01283	0.1211
Configuration	-1.737	0.4787	-2.723	-0.8641
Surface	3.087	0.4763	2.134	4.036
Visibility	-0.238	0.05056	-0.3399	-0.1453
	$\bar{D}$	**p$_D$**	**DIC**	
	247.222	6.324	253.547	
	AUC	**Sensitivity**	**Specificity**	**Accuracy**
Training	0.9346	0.8491	0.8911	0.8900
Validation	0.9710	0.9231	0.9044	0.9049

The Logarithm of vehicle count in 5-minute intervals is positive, indicating that high volume might increase the likelihood of SV crashes on a ramp. Speed is found to be significant with a positive sign. When the vehicles are at high speed, if the driver are distracted or influenced by unexpected occurrences, they may suddenly brake or turn the wheel. These could be hazardous on ramp and drivers may lose control of vehicles and resulting in SV crashes, because ramp has steep slope and/or small turning radius.

Ramp configuration is significant and proven to be negatively related to SV crashes, since curved ramps have smaller turning radii compared to diamond ramps, and can lead to a loss of vehicles' control and result in SV crashes. Wet road surfaces have smaller friction and result in longer braking distances than on dry surfaces. It can easily result in vehicles spinning out of control. Consequently, wet road surfaces may contribute to an increased potential for SV crashes. Visibility is statistically significant and found to be negatively related to SV crash occurrence, which suggest that SV crashes are more probable during poor visibility conditions.

5.2 Real-time Multi-Vehicle Crash Model

Estimation results for the real-time MV crash-prediction model are shown in Table 5. In the model, four variables are significant at the 95% confidence interval, and the standard deviation of speed is found to be significant at the 90% interval. AUC area for training and validation are 0.8134 and 0.8095, respectively. The overall accuracy for training and validation are 0.7644 and 0.7600 when the cutoff-point is 0.035, which means the model's predictive accuracy for discriminating between crashes and non–crashes is good.

An important factor of MV crashes is the variation of speed along the segment (*15*), so almost all previous research had at least two traffic data collection stations which were at the upstream and downstream of crashes. A potential restriction in our research, particularly for MV crashes, is that there is only one traffic data collection station located at the ramp.

TABLE 5 Real-time MV Crash Prediction Model

node	Mean	Std	2.50%	97.50%
intercept	-8.959	1.493	-12.07	-6.124
Log(Vehcnt)	1.157	0.2208	0.7252	1.589
Spd	0.04775	0.01872	0.01056	0.08542
Std_spd	0.0646	0.03278	-0.00443	0.1244*
Type	0.8447	0.3483	0.1939	1.546
Visibility	-0.1467	0.05222	-0.2428	-0.03851
	$\bar{D}$	p_D	**DIC**	
	350.385	5.922	356.307	
	AUC	**Sensitivity**	**Specificity**	**Accuracy**
Training	0.8134	0.7500	0.7648	0.7644
Validation	0.8095	0.6429	0.7640	0.7600

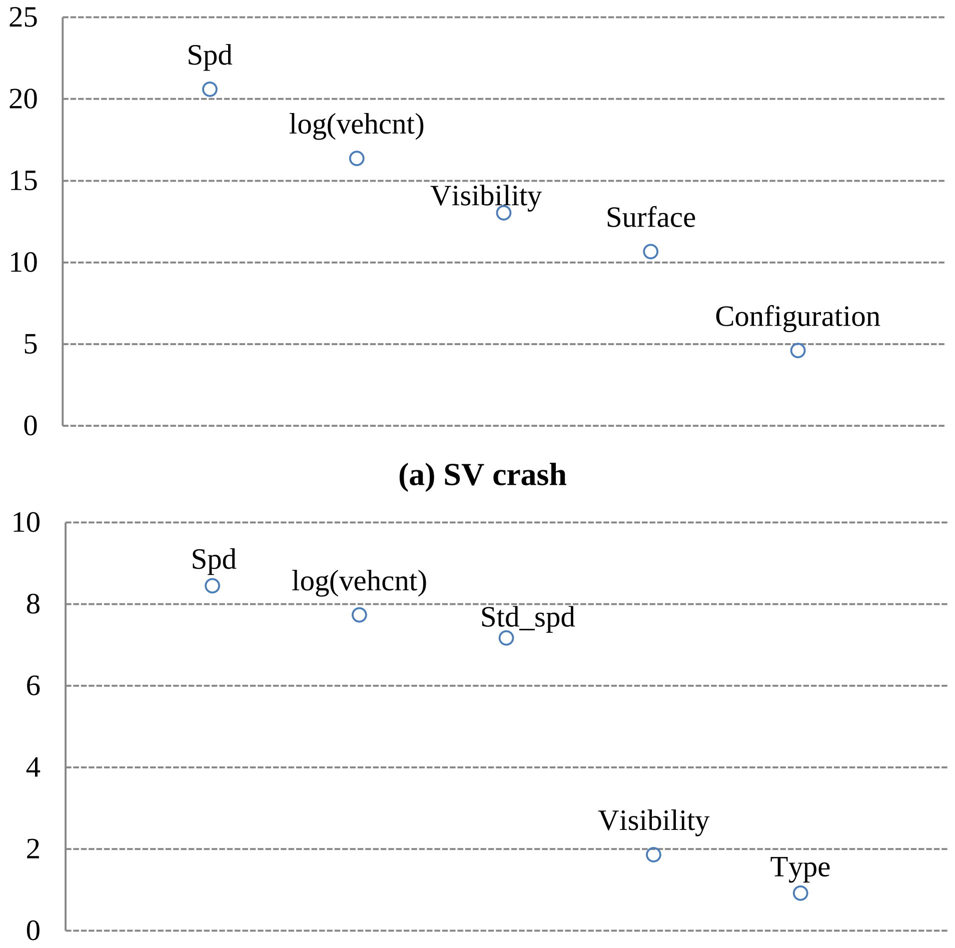
** variable significant at 90% interval*

The Logarithm of vehicle count in a 5-minute interval is positive, which indicates high volume may increase the total interactions between vehicles and increase the likelihood of MV crashes. Speed is found to be significant with a positive sign. As the speed increases, so does the stopping sight distance. Since speed will definitely increase both the braking distance and the reaction distance, a vehicle travelling at a higher speed will more likely have a collision with the vehicle ahead of it. Hence, higher speed would increase the possibility of MV crashes. The standard deviation of speed is a good indicator of traffic turbulence. When there is a significant speed difference, deceleration or acceleration action would need to be taken to guarantee a safe following distance. Under these circumstances, rear-end crashes are likely to occur on ramps.

Ramp type is significant and proven to be positively related to MV crashes. Vehicles on the off-ramps need to slow down to adjust to the lower surface street speed. If the following vehicle does not react and decelerate in time, it will run into the leading vehicle, and have a MV crash. Visibility is significant with a negative sign. Under poor visibility, car-following and lane-changing are much harder, so vehicles may have rear-end or sideswipe crashes.

5.3 Variable Importance

This study employed Random Forest to rank the importance of variables which are significant in the real-time SV and MV crash models according to the Bayesian logistic regression models. The results are illustrated in Figure 1.



(a) SV crash

(b) MV crash

FIGURE 1 Variable Importance

From Figure 1, we observe that traffic variables are more important than weather and ramp geometric variables for both models. Meanwhile, speed is the most important factor in both SV and MV models. Thus, informing drivers of reducing speeds through Dynamic Message Signs (DMS) may be the most effective way to reduce crash likelihood. Weather variables' impact on real-time crashes is moderate. Warning drivers that the road surface is wet can significantly reduce SV crash likelihood, and informing drivers that they should be careful in low visibility may reduce both SV and MV crashes. Ramp geometric variables have a significant, but the least, impact on crash occurrence. The effects of warnings on ramp type or configuration may not be as efficient as those regarding regulating speed or the presence of severe weather.

Regardless of crash type (SV or MV), the essential factors used in real-time crash prediction on ramps are traffic variables, e.g., volume, speed, and standard deviation of speed. This is the reason why all the models in the previous work provided good real-time crash predictions with only traffic information. However, if real-time weather information along with ramp geometric characteristics can be used in building the crash prediction model, this would be more ideal than just including traffic parameters, since the weather and geometric variables are statistically significant and important factors in predicting crashes on ramps.

6 SUMMARY AND CONCLUSIONS

No research has been conducted on real-time crash prediction for expressway ramps with combined real-time traffic, weather and geometric information. This paper implemented two Bayesian logistic models to predict in real time the likelihood of SV and MV crashes on expressway ramps based on MVDS data, airport weather data, and ramp geometric information.

The descriptive and exploratory analysis show that the crash type is linked to the standard deviation of speed, ramp type, ramp configuration, road surface condition and visibility. This finding corroborates the importance of distinguishing between SV and MV crashes, since the crash type is obviously not homogeneous across the traffic, geometric and weather parameters. Curved or/and wet road-surface ramps are more likely to have SV crashes. Off-ramps increase the possibility of MV crashes.

The occurrence of SV and MV crashes is significantly influenced by the Logarithm of vehicle count, average speed in 5-minute intervals, and visibility. If the Logarithm of vehicle count or average speed increases or visibility decreases, the likelihood of SV and MV crashes will obviously increase. When the Logarithm of vehicle count increases by one unit, the odds ratio of an SV crash is 2.6, and that of an MV crash is 3.18. This means that the Logarithm of vehicle count has a greater positive impact on the occurrence of MV crashes. On the contrary, speed and visibility have greater impact on odds ratio of SV crashes than on that of MV crashes. The standard deviation of speed is only significant in the MV crash prediction model. When it increases, the likelihood of MV crashes increases significantly. As for the categorical variables, the Bayesian logistic regression models' results are the same as that of the exploratory analysis. Ramp configuration and road surface condition have significant impact on the occurrence of SV crashes, and ramp type would obviously influence the MV crash risk.

Variable importance analysis indicates that the most important factors for SV and MV models are traffic variables; the least important but still significant factors were ramps' geometric characteristics. In practice, when traffic conditions are poor and weather is severe simultaneously, traffic-related warning information should be given the priority on DMS. Real-time changing messages and colors based on the condition and risk should also be considered.

Since the real-time crash prediction models in this study contain geometric information, it's possible to get the relative crash risk for different types of ramps. Meanwhile, we conclude that MV and SV crashes on ramps have different precursors and these precursors' impacts are different, in other words, their crash mechanisms are not exactly the same. When implementing ITS (Intelligent Transportation Systems) to decrease crash risk on ramps, it is advisable to calculate the crash risk for both MV and SV crashes, and then show the warning information based on the higher calculated risk value. Since speed is the most important factor affecting crash occurrence for both SV and MV models, informing drivers of adapting their speed through DMS may be one of the most effective ways to reduce crash likelihood.

There are some limitations of this study. Only four ramp geometric parameters are included, they are ramp type, configuration, length and presence of a toll booth. In future work, more geometric variables, e.g., lane and shoulder widths, curvature, gradient, may be attempted in the model estimation.

ACKNOWLEDGEMENT

The authors thank the Central Florida Expressway Authority (CFX) for funding this research and providing data. The authors also thank the Southeastern Transportation Center UTC consortium for partial funding of this research. The authors wish to thank Charles R. Lattimer, PE, PMP Sr. Project Manager, Intelligent Transportation Systems at ATKINS for his suggestions.

REFERENCE

[1] Lee, C., and M. Abdel-Aty. Analysis of Crashes on Freeway Ramps by Location of Crash and Presence of Advisory Speed Signs. *Journal of Transportation Safety & Security*, Vol. 1, No. 2, 2009, pp. 121-134.

[2] Pande, A., and M. Abdel-Aty. Assessment of freeway traffic parameters leading to lane-change related collisions. *Accident Analysis & Prevention*, Vol. 38, No. 5, 2006, pp. 936-948.

[3] Yu, R., and M. Abdel-Aty. Multi-level Bayesian analyses for single- and multi-vehicle freeway crashes. *Accident Analysis & Prevention*, Vol. 58, No. 0, 2013, pp. 97-105.

[4] Oh, C., J.-S. Oh, S. G. Ritchie, and M. Chang. Real-time estimation of freeway accident likelihood. Presented at 80th Annual Meeting of the Transportation Research Board, Washington, DC, 2001.

[5] *How Do Weather Events Impact Roads*? Federal Highway Administration. http://ops.fhwa.dot.gov/Weather/q1_roadimpact.htm. Accessed June 2, 2014.

[6] Goodwin, L. C. Weather impacts on arterial traffic flow. *The Road Weather Management Program*, FHWA, US Department of Transportation, Washington DC, 2002.

[7] Abdel-Aty, M., N. Uddin, A. Pande, F. M. Abdalla, and L. Hsia. Predicting freeway crashes from loop detector data by matched case-control logistic regression. *Transportation Research Record: Journal of the Transportation Research Board*, Vol. 1897, No. 1, 2004, pp. 88-95.

[8] Abdel-Aty, M., H. Hassan, and M. Ahmed. Real-Time Analysis of Visibility Related Crashes: Can Loop Detector and AVI Data Predict Them Equally? Presented at Transportation Research Board 91st Annual Meeting, Washington, DC, 2012.

[9] Hossain, M., and Y. Muromachi. A real-time crash prediction model for the ramp vicinities of urban expressways. *IATSS Research*, Vol. 37, No. 1, 2013, pp. 68-79.

[10] Hossain, M., and Y. Muromachi. A Bayesian network based framework for real-time crash prediction on the basic freeway segments of urban expressways. *Accident Analysis & Prevention*, Vol. 45, 2012, pp. 373-381.

[11] Hossain, M., and Y. Muromachi. Understanding crash mechanism on urban expressways using high-resolution traffic data. *Accident Analysis & Prevention*, Vol. 57, 2013, pp. 17-29.

[12] Xu, C., A. P. Tarko, W. Wang, and P. Liu. Predicting crash likelihood and severity on freeways with real-time loop detector data. *Accident Analysis & Prevention*, Vol. 57, 2013, pp. 30-39.

[13] Zheng, Z., S. Ahn, and C. M. Monsere. Impact of traffic oscillations on freeway crash occurrences. *Accident Analysis & Prevention*, Vol. 42, No. 2, 2010, pp. 626-636.

[14] Madanat, S., and P.-C. Liu. *A prototype system for real-time incident likelihood prediction. ITS-IDEA Program Project Final Report*, 1995.

[15] Lee, C., F. Saccomanno, and B. Hellinga. Analysis of crash precursors on instrumented freeways. *Transportation Research Record: Journal of the Transportation Research Board*, Vol. 1784, No. 1, 2002, pp. 1-8.

[16] Abdel-Aty, M., and R. Pemmanaboina. Calibrating a real-time traffic crash-prediction model using archived weather and ITS traffic data. *IEEE Transactions on Intelligent Transportation Systems* , Vol. 7, No. 2, 2006, pp. 167-174.

[17] Ahmed, M. M., M. Abdel-Aty, R. Yu, and J. Lee. Exploring the Feasibility of Using Airport Data in Real-Time Risk Assessment. Presented at Transportation Research Board 92nd Annual Meeting, Washington, DC, 2012.

[18] Christoforou, Z., S. Cohen, and M. G. Karlaftis. Identifying crash type propensity using real-time traffic data on freeways. *Journal of Safety Research*, Vol. 42, No. 1, 2011, pp. 43-50.

[19] Lord, D., and J. A. Bonneson. Calibration of predictive models for estimating safety of ramp design configurations. *Transportation Research Record: Journal of the Transportation Research Board*, Vol. 1908, No. 1, 2005, pp. 88-95.

[20] Garnowski, M., and H. Manner. On factors related to car accidents on German Autobahn connectors. *Accident Analysis & Prevention*, Vol. 43, No. 5, 2011, pp. 1864-1871.

[21] Lee, C., and M. Abdel-Aty. Temporal Variations in Traffic Flow and Ramp-Related Crash Risk. Presented at Applications of Advanced Technology in Transportation. The Ninth International Conference, 2006.

[22] Nemes, S., J. M. Jonasson, A. Genell, and G. Steineck. Bias in odds ratios by logistic regression modelling and sample size. *BMC medical research methodology*, Vol. 9, No. 1, 2009, p. 56.

[23] Gelman, A., J. B. Carlin, H. S. Stern, D. B. Dunson, A. Vehtari, and D. B. Rubin. *Bayesian data analysis*. CRC press, 2013.

[24] Sturtz, S., U. Ligges, and A. E. Gelman. R2WinBUGS: a package for running WinBUGS from R. *Journal of Statistical software*, Vol. 12, No. 3, 2005, pp. 1-16.

[25] Lunn, D. J., A. Thomas, N. Best, and D. Spiegelhalter. WinBUGS-a Bayesian modelling framework: concepts, structure, and extensibility. *Statistics and computing*, Vol. 10, No. 4, 2000, pp. 325-337.

[26] Spiegelhalter, D. J., N. G. Best, B. P. Carlin, and A. Van Der Linde. Bayesian measures of model complexity and fit. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, Vol. 64, No. 4, 2002, pp. 583-639.