Identification of High Crash-Risk Zones in Southern California

Yongping Zhang, Ph.D., P.E.
Assistant Professor, Department of Civil Engineering
California State Polytechnic University, Pomona
3801 W. Temple Ave., Pomona, CA 91768
Tel: (909) 869-2632, Email: yongpingz@cpp.edu

Tom Vo, M.S.C.E.
Regional Planner/GIS Analyst
Southern California Association of Governments
818 West 7th Street, 12th Floor, Los Angeles, CA 90017
Tel: (213) 236-1800, Email: Vo@scag.ca.gov

Frank Wen, Ph.D.
Manager of Research and Analysis
818 West 7th Street, 12th Floor, Los Angeles, CA 90017
Tel: (213) 236-1800, Email: Wen@scag.ca.gov

Gurdiljot Singh Gill
Graduate Student, Department of Civil Engineering
ITE Student member, California State Polytechnic University, Pomona
3801 W. Temple Ave., Pomona, CA 91768
Tel: (909) 869-4312, Email: gurdiljotg@cpp.edu

Wen Cheng, Ph.D., P.E., T.E., PTOE
Associate Professor, Department of Civil Engineering
ITE Student chapter faculty advisor, California State Polytechnic University, Pomona
3801 W. Temple Ave., Pomona, CA 91768
Tel: (909) 869-4312, Email: wcheng@cpp.edu

Abstract

This study introduced traffic crash prediction module to Southern California Association of Governments (SCAG) 's existing transportation model structure such that traffic crashes can be consistently predicted at the regional planning level using the existing transportation analysis zones (TAZs) as unit of analysis in the SCAG region. To this end, this study identified the influence of built environment on collisions by type (i.e. collisions between automobiles and between automobile and pedestrian or bicycle) in the SCAG region. To quantify the crash counts of various transportation modes at TAZ level, Multivariate crash frequency model was developed to jointly estimate the crash risk of different mode users. Then, three tests were introduced to evaluate the adopted Hot Spot Identification (HSID) method. All three test results show that the multivariate modeling method is superior to the commonly used observed count method.

Keywords: TAZ; multivariate correlation; crash prediction; Bayesian; HSID

1. Introduction

During the year of 2014, 32,675 fatalities and more than 2,338,000 injuries occurred on the US roads. Road accidents were the leading cause of death among ages 16 through 24 in 2014 in the U.S. (1). In Southern California Association of Governments (SCAG) region, each year about 1,500 fatalities and 120,000 injuries occur due to traffic accidents. These fatalities and injuries reflect a significant proportion of healthy lives which could have been saved by the application of appropriate safety countermeasure treatments. The traffic management processes which address safety issues include network screening, problem diagnosis, countermeasure identification, and project prioritization.

MAP–21 and FAST Act call for establishing performance measures and standards on traffic safety. FHWA is now requiring DOTs to work with MPOs to assess fatalities and serious injuries on all public roads and to set annual performance measures. The most current SCAG 2016-2040 Regional Transportation Plan/Sustainable Communities Strategies (RTP/SCS) was adopted in April 2016. In this regional planning document, traffic safety issue is considered as an important subject. The rate of fatal and injury collisions is briefly reviewed and presented as heat maps in the region (2). While SCAG intends to concentrate more on road safety and develop guidance to improve road safety, there is still a significant gap in modeling traffic accidents at the regional planning level. By quantifying the relationship between various traffic accidents and the built environment variables, modeling and forecasting traffic accidents can help identify hot spots of accident locations, allocate scarce resources to the most critical locations more accurately and efficiently, understand the impacts of safety countermeasures and initiate other useful planning research studies such as environmental justice analysis.

Over the past decades, SCAG has successfully developed and implemented state-of-the-art models to forecast both socioeconomic and transportation activities. The outcomes of these models can be utilized to forecast the traffic accident rates related to both motorized and non-motorized modes.

In this context, this study introduced traffic crash prediction module to SCAG's existing transportation model structure such that traffic crashes can be consistently predicted at the regional planning level using the existing transportation analysis zones (TAZs) as unit of analysis in the SCAG region. To this end, this study identified the influence of built environment on collisions by type (i.e. collisions between automobiles and between automobile and pedestrian or bicycle) and quantified the potential impacts of future built environments on safety in the SCAG region. It is intended to meet the SCAG's Strategic Plan to develop, maintain and promote the utilization of state-of-the art models, information systems and communication technologies. (3)

2. Data

Crashes which occurred in the SCAG region (i.e., Counties of Los Angeles, Orange, Ventura, Riverside, San Bernardino, and Imperial) in the period of 2012–2013 were analyzed for the study. In specific, 2012 data were used for detection of hot zone and modeling development, and 2013 data were employed for HSID performance evaluation. As demonstrated in previous research (Abdel-Aty et al., 2013), compared with other geographic units such as block groups and census tracts, TAZs have benefits of better homogeneity and easy integration into the transportation planning process. Therefore, TAZs were selected as the base units, and the crash data were aggregated at the TAZ-level. Overall, there are 11,267 tier-2 TAZs in the SCAG regional planning area. Three different transportation mode-related crashes were collected from SWITRS (California Statewide Integrated Traffic Records System) which include pedestrian, bicyclist, and motorized vehicle only crashes. Shapefiles of TAZ boundary and TAZ characteristics were provided by SCAG.

The variables used for model development and their description are shown in Table 1. The numbers of various transportation mode-involved crashes were used as the dependent variables. DVMT was utilized as the exposure variable. The explanatory variables were the predictors commonly used in previous regional safety analyses which include socioeconomic, transportation-related, and environment-related factors, and so on.

TABLE 1. Description of Variables

Var	Description
Cnty	County
TAZ	TAZ (tier2)
Acre	TAZ area in acre
Density	
Pop_den	Population density (persons/acre)
HH_den	Household density (hh/acre)
Emp_den	Employment density (jobs/acre)
Ret_den	Rtail job density
RetSer_den	Retail+ Service (retail + FIRE + ArtsFood + Other Serv.) job density
Diversity / Mixed Use of Land	
Jobmix13	Job mix (13sectors); 1 = highest mex (jobs are equal for all sectors)
Jobmix9	Job mix (9 sectors)
Emix13	Job mix (13sectors); 1 = highest mex (jobs are equal for all sectors)
Emix9	Job mix (9 sectors)
EH_ratio	Job/Household ratio
EP_ratio	Job/Pop ratio
Built Environment / Access to Active Transportation & Transit	
Int34_Den	Intersection density (3- and 4- legs)
BKlnAcc	Bike lane access (1=if a TAZ has bike lane)
Rail	1=at least one rail station in a TAZ
ExBus_D	Stop density for Express Bus and BRT
HFLbus_D	High-Frequency Bus Stop Density (local bus headway <= 20 mins)
TTbus_D	Total Bus Stop Density

Land Development Characteristics: TOD (HQTA / TPA)	
Mlt_pct	% of households living in multiple unit
HQTA_pct	% of TAZ area are in non-freeway HQTA (high-quality transit area)
TPA_pct	% of TAZ area are in TPA
Additional Biking / Walking Related Built Environment Variable	
BLdenIND	Bike Lane Density Indicator = Sum (Bike Lane Denity/Distance to Home TAZ within 3 mile) Bike Lane Density for Each TAZ = ((Street15-25mpg)*1 + (Street35mpg)*2 + Bike Lane Class1 * 3 + Bike Lane Class2*4 Bike Lane Class3 *5) / Total TAZ area (excluding speed >60mph)
Blck_len	Estimated block length = LocalSt/Int34new (Total street length / number intersection) - highways are exluded
WalkAcc	Walk Accessibility (RS_den2/block_len) = (weighted retail + service density)/ estimated block length
Pct_Art	Percent of main arterial (45-55mph) of TAZ - higher % means more difficult to across street (also larger block to across steet); can be used with WalkAcc

3. Method

3.1 Identifying the Collision Hotspots by Different Transportation Modes

An efficient identification of hotspots in terms of transportation modes relies on an accurate **Crash Prediction Model** (i.e., crash frequency model). Hence, the research team quantified the crash counts of various transportation modes at TAZ level using the Multivariate crash frequency model, which can jointly estimate the crash risk of different mode users. The simultaneous modeling is essential given the unobserved heterogeneity shared by various transportation modes. Ignorance of such correlation structures has been illustrated to reduce the efficiency of the model due to lesser precise parameters (Cheng et al., 2017a; Park and Lord, 2007). In comparison with the large number of univariate models dedicated to various mode users, very few studies have used the joint models to analyze the interaction between different modes. Given the crash count nature, this study will utilize the Multivariate Poisson-Lognormal Model (MVPLN).

This model assumes that crash count of certain modal crash *j* at a given location *i* in time *t* (in years), y^j_{it}, obeys Poisson distribution, while the corresponding observation specific error term ε_{ijt} follows a multivariate normal distribution:

$$\boldsymbol{y_{ijt}}|\boldsymbol{\lambda_{ijt}} \sim Poisson\ (\boldsymbol{\lambda_{ijt}}) \quad (1)$$

$$\ln(\boldsymbol{\lambda_{ijt}}) = \boldsymbol{X'_{ijt}\beta} + \boldsymbol{\varepsilon_{ijt}} \quad (2)$$

$$\boldsymbol{\varepsilon_{ijt}} \sim Normal\ (\mathbf{0},\ \textstyle\sum) \quad (3)$$

Where $\boldsymbol{y_{ijt}} = \begin{pmatrix} y^1_{it} \\ y^2_{it} \\ y^3_{it} \end{pmatrix}$, $\boldsymbol{\lambda_{ijt}} = \begin{pmatrix} \lambda^1_{it} \\ \lambda^2_{it} \\ \lambda^3_{it} \end{pmatrix}$, $\boldsymbol{\varepsilon_{ijt}} = \begin{pmatrix} \varepsilon^1_{it} \\ \varepsilon^2_{it} \\ \varepsilon^3_{it} \end{pmatrix}$, $\sum = \begin{pmatrix} \sigma_{11} & \cdots & \sigma_{13} \\ \vdots & \ddots & \vdots \\ \sigma_{31} & \cdots & \sigma_{33} \end{pmatrix}$ (4)

In above equations, ***X'*** is the matrix of risk factors, **β** is the vector of model parameters, ε^j_{it} is the independent random effect which captures the extra-Poisson heterogeneity among locations. ∑ is called the covariance matrix. The diagonal element $\boldsymbol{\sigma_{jj}}$ in the matrix represents the variance of $\boldsymbol{\varepsilon_{ij}}$, where the off-diagonal elements represent the covariance of crash counts of different severities. The inverse of the covariance matrix represent the precision matrix and has the following distribution:

$$\textstyle\sum^{-1} \sim Wishart(I, J) \quad (5)$$

Where I is the J x J identity matrix (*Congdon, 2006*), and J is the degree of freedom, J=3 herein representing 3 crash outcomes.

The Deviance Information Criterion (DIC) developed by Spiegelhalter et al. (*2003*) was employed to assess the complexity and fit of the models. The DIC is computed as the sum of the posterior mean deviance and estimated effective number of parameters:

$$DIC = \overline{D} + p_D \quad (6)$$

Where $\overline{D}$ is the sum of the posterior mean deviance which measures how well the model fits the data; the smaller the$\overline{D}$, the better the fit. p_D represents the effective number of parameters. In general, $\overline{D}$ will decrease as the number of parameters in a model increases. Therefore, the p_D term is mainly used to compensate for

this effect by favoring models with a smaller number of parameters. This idea is analogous to other penalized fit criteria such as Akaike information criterion and Bayesian information criterion.

The MVPLN model was implemented in freeware WinBUGS package (Spiegelhalter et al., 2003) which employs an MCMC algorithm for estimation of parameters. For the model calibration, two Markov chains were utilized to visually inspect trace plots of posterior estimates for convergence, which was further checked by using the Gelman-Rubin convergence statistic (*Flegal et al., 2015*). The following two subsections describe the modeling development and HSID results in order.

3.1.1 Modeling Development Results

In general, the larger the effective number of parameters is, the easier it is for the model to fit the data. To obtain a parsimonious model and avoid risk of overfitting, backward stepwise methods were employed in selecting covariates. Besides, a correlation matrix for the variables entered in the final models has been checked to avoid multicollinearity issues. Results of parameter estimation and associated uncertainty estimates of significant variables in the final models are presented in Table 2. It is known that the same significant variables are identified for all crash types.

TABLE 2 Estimates of Regression Coefficients Obtained by Bayesian Multivariate Model

Crash Types	Variables	Mean	Standard Deviation	2.5%	97.5%
Pedestrian	Constant	-1.692	0.0459	-1.777	-1.599
	Blck_Len	-1.112	0.1105	-1.337	0.8978
	WalkAcc	0.0175	0.0018	0.014	0.021
	Pct_Art	-0.9135	0.2817	-1.46	-0.3825
	Pop_den	0.0456	0.0025	0.0402	0.0498
	HH_den	-0.0239	0.0062	-0.0341	-0.0104
	Emp_den	0.0046	8.026E-4	0.0030	0.0062
	VMT_total (unit: 1000veh-miles)	0.0462	0.0025	0.0411	0.0501
Bike	Constant	-1.631	0.0457	-1.727	-1.544
	Blck_Len	-0.7477	0.0822	-0.9152	-0.5854
	WalkAcc	0.017	0.0018	0.013	0.020
	Pct_Art	-1.27	0.2767	-1.818	-0.7241
	Pop_den	0.1408	0.0027	0.0358	0.0455
	HH_den	-0.0282	0.0070	-0.0409	-0.0131
	Emp_den	0.0054	8.02E-4	0.0037	0.0068
	VMT_total (unit: 1000veh-miles)	0.0446	0.0024	0.0397	0.0493
Motorized	Constant	0.6249	0.0204	0.5852	0.6630
	Blck_Len	-0.3078	0.0253	-0.3555	-0.2578
	WalkAcc	0.011	0.0013	0.008	0.013
	Pct_Art	1.489	0.1193	1.254	1.733
	Pop_den	0.0194	0.0021	0.0146	0.0232
	HH_den	-0.0150	0.0057	-0.0248	-4.223E-4
	Emp_den	0.0035	5.727E-4	0.0023	0.0046
	VMT_total (unit: 1000veh-miles)	0.0729	0.0013	0.0702	0.0751
	$\bar{D}$	79247.400			
	p_D	10571.200			
	DIC	89818.600			

Note: Refer to Table 1 for detailed description of variables

Similar to previous literature (Cheng et al., 2017b), the variance estimates of all three crash types are statistically significant at the 0.05 level of significance which indicate the presence of over-dispersion and strong correlation in all modal crashes. The covariance results for the model are shown in Table 3. It exhibits

that the correlations are statistically significant for heterogeneity error term among various crash types, demonstrating that the occurrence of various crash types is highly correlated. The highest correlation is observed between Bike and Pedestrian crashes (0.918), which might suggest that the bicyclists and pedestrians have closer behaviors than motorized vehicles.

TABLE 3 The Estimate of the Covariance among Crash Types

	Pedestrian	Bike	Motorized Vehicle
Pedestrian	**0.940 (0.041)**		
Bike	**0.918 (0.027)**	**1.045 (0.034)**	
Motorized Vehicle	**0.714 (0.023)**	**0.742 (0.022)**	**0.849 (0.018)**

Notes: 1.The standard deviations are shown in parentheses.
2. The statistically significant correlation coefficients are shown in bold.

3.1.2 Hot Zones Identification Results

Once the above model are developed jointly, the model output (i.e., estimated $\boldsymbol{\lambda ijt}$) can be used to rank the TAZ's for identification of the hot zones in terms of each transportation mode. For illustration purpose, Tables 4-6 exhibit the top 30 TAZs in terms of Bayesian modeling-estimated crash counts for each transportation mode, respectively.

TABLE 4. Illustration of Top 30 TAZs in terms of Bayesian-estimated Pedestrian Crash Counts

Ranks	County	TAZ	Observed Pedestrian Crash Counts	Bayesian-estimated Pedestrian Crash Counts
1	37	21950100	15	16.73
2	37	20953200	14	10.1
3	37	21119100	17	10.08
4	37	20953100	12	8.928
5	37	21555300	7	8.554
6	37	21055200	9	8.523
7	37	20937100	9	7.941
8	59	32955100	9	6.898
9	37	20989100	9	6.87
10	37	21393100	9	6.678
11	37	21024100	7	6.338
12	37	21062100	5	6.072
13	37	21107200	8	6.015
14	37	21918100	10	6.013
15	37	21937100	5	5.979
16	37	20899300	12	5.958
17	37	21544300	9	5.892
18	59	32521200	7	5.871
19	37	21953100	5	5.852
20	37	21922100	5	5.769
21	37	21002100	10	5.713
22	37	21097100	7	5.705
23	37	21017200	4	5.54
24	37	21944100	4	5.431
25	37	20749100	6	5.394
26	59	32728200	5	5.371
27	37	20996200	8	5.165

28	37	20531100	9	5.151
29	37	20519100	8	5.086
30	37	21022100	4	5.071

TABLE 5. Illustration of Top 30 TAZ's in terms of Bayesian-estimated Bike Crash Counts

Ranks	County	TAZ	Observed Pedestrian Crash Counts	Bayesian-estimated Pedestrian Crash Counts
1	37	21944100	14	9.969
2	59	32728200	14	9.399
3	37	20953200	7	7.703
4	37	21937100	10	7.251
5	37	21055200	7	6.897
6	37	20749100	10	6.781
7	37	21950100	7	6.688
8	37	21119100	6	6.592
9	59	32521200	7	6.358
10	37	21024100	8	6.213
11	37	20534100	8	6.101
12	59	32955100	6	6.078
13	37	20748100	9	5.961
14	37	21953100	7	5.869
15	37	20772100	9	5.736
16	37	21393100	6	5.691
17	37	21544300	6	5.684
18	37	20953100	4	5.595
19	37	21062100	7	5.549
20	59	32934100	8	5.479
21	59	32932100	9	5.47
22	59	32949100	7	5.44
23	37	21052200	8	5.338
24	59	32733100	7	5.329
25	59	32736300	10	5.215
26	59	32728100	10	5.164
27	37	21939100	7	5.016
28	37	20794100	7	4.999
29	37	21929100	9	4.937
30	37	21508100	8	4.917

TABLE 6. Illustration of Top 30 TAZ's in terms of Bayesian-estimated Motorized Crash Counts

Ranks	County	TAZ	Observed Motorized Crash Counts	Bayesian-estimated Motorized Crash Counts
1	37	20505100	232	242.8
2	37	21957100	194	211.7
3	37	21587200	104	103.2
4	71	53860200	94	103.1
5	37	21119100	90	95.98

6	65	43147200	93	95.1
7	59	32889100	82	85.36
8	59	32905100	77	83.35
9	59	32943100	80	81.3
10	37	20586100	65	65.18
11	71	53885100	60	64.45
12	37	20774100	56	63.91
13	59	32987100	65	63.74
14	37	20227100	60	62.37
15	37	20395100	60	62.12
16	59	32975400	57	55.97
17	37	22022100	56	54.73
18	37	22261200	55	54.65
19	59	32878300	57	54.23
20	37	22281100	52	54.14
21	37	21714200	53	54.08
22	37	22257100	51	53.23
23	71	53884100	46	51.45
24	37	21326100	49	50.57
25	59	32585100	52	50.41
26	37	20530300	53	49.68
27	37	21142100	50	49.14
28	37	20622100	48	49.1
29	111	60088100	51	48.6
30	59	32582100	50	48.58

3.2 Evaluation of the Adopted Hot Spot Identification (HSID) Method

Compared with the large number of studies focused on the development of various HSID methods, considerably less research has been dedicated to devising the evaluation criteria for comparing the performance of various methods. Cheng and Washington (2008) developed three unique evaluation criteria containing the Site Consistency Test (SCT), the Method Consistency Test (MCT), and the Total Rank Difference Test (TRDT), which were used in the study to evaluate the safety performance of hotspots identified by joint models (i.e., $\boldsymbol{\lambda_{ijt}}$) and observed crash counts (i.e. $\boldsymbol{y_{ijt}}$), respectively. The following subsections present the succinct description of these tests.

3.2.1Evaluation Criteria

T1: Site Consistency Test

The Site Consistency Test (T1) is used to measure the ability of a HSID method to consistently identify a site as high risk over subsequent observation periods. The test rests on the premise that a site identified as high risk during time period 1 should also reveal inferior safety performance in a subsequent time period 2, given that no significant changes have occurred at the site and given that the site is in fact high risk. T1 simply requires a comparison of the sum of observed crashes occurring on $n\alpha$ high risk sites during future time period $i + 1$ identified by method j (during time period i) to crashes occurring at high risk sites (in time period $i + 1$) identified by other 'competing' HSID methods. In equation form, this comparison is given as:

$$T1 = \sum_{k=n-n\alpha}^{n} C_{k,Method=j(i),i+1} > \sum_{k=n-n\alpha}^{n} C_{k,Method\neq j,i+1} \quad (7)$$

where, n is the total number of sites being compared (e.g. crashes occurring on n=100 rural signalized intersections), C is the crash count for site ranked site k, α is the threshold of identified high risk sites (e.g. $\alpha = 0.05$ corresponds with top 5% of n sites identified as high-risk, and $n\alpha$ is the number of identified high risk sites), j = HSID method being compared (e.g. j = 1 could be the observed count method, j = 2 the multivariate modeling method), and i is observation period (e.g. i = 2012, i +1= 2013). In this comparison, the method j that identifies sites in a future period with the highest crash count is the most consistent method for identifying underlying safety problems.

T2: Method Consistency Test

In the Site Consistency Test previously discussed, crash counts in Period 2 are used as a benchmark to compare different HSID methods. The underlying assumption is that the identification performance of the HSID methods is revealed through the safety performance of the corresponding identified hot spots. The Method Consistency Test, in contrast, is designed to evaluate a method's performance by measuring the number of the same hot spots identified in both periods. Inasmuch as the two periods are close in time, it is assumed that road sections are in the same or similar underlying operational state (similar traffic volumes, driver populations, geometric designs, weather fluctuations, etc.) and their expected safety performance remains virtually unaltered over the two periods. Under this homogeneity assumption, a good HSID method will identify the same set of hot spots across two different periods. The greater the number of hot spots that are identified in both periods the more reliable and consistent is the performance of HSID method.

Analytically, test T2 is simply the intersection of ranked sites k identified in subsequent time periods i and i + 1 that are high risk, or:

$$T2_j = \{k_{n-n\alpha}, k_{n-n\alpha+1}, \ldots\ldots, k_n\}_{j,i} \cap \{k_{n-n\alpha}, k_{n-n\alpha+1}, \ldots\ldots, k_n\}_{j,i+1}, \quad (8)$$

where, terms are as defined previously. Note that only sites identified in the top threshold α are compared.

In this test, the intersection of sites identified as high risk in two subsequent periods is compared over methods j=1 to J, and the method yielding the largest intersection of sites is said to be the most consistent.

T3: Total Rank Differences Test

In the Method Consistency Test (T2), the number of consistently identified hot spots common to both periods is used to measure a method's reliability. The Total Rank Differences Test, building on T2, takes into account the rankings of safety performances of road sections in the two periods. To illustrate, consider a group of 100 road sections, among them there is one site whose AR ranks 1st in Period 1 and ranks 10th in Period 2. If the top 10 sites are identified as hot spots, then the AR method has identified a site in both periods even though the rankings have changed between periods. The Total Rank Differences Test, T3, is proposed to account for the rankings. The test is conducted by calculating the sum of total rank differences of the hazardous road sections identified across the two periods. The smaller is the total rank difference, the more consistent is the HSID method on this test—reflecting consistent ranking of sites across periods. As for the prior tests, this test is based on the assumption that no safety treatments are implemented on the road sections and the underlying safety of the road section during the two periods is constant. Hence, it is of great importance to ensure all the data outliers (i.e., road sections that are treated during Period 2) have been identified and removed when conducting this test.

In equation form, this test statistic is given as:

$$\boldsymbol{T3_j = \sum_{k=n-n\alpha}^{n} |R(k_{j,i}) - R(k_{j,i+1})|} \quad (9)$$

where R is the rank of site k in period i for method j. Note that the difference in ranks is summed over all identified sites for threshold level α for period i.

3.2.2 Test Results

The three tests described previously (T1 through T3) were used to assess the relative performance of two HSID methods, the commonly used observed count method (Method 1) and the multivariate modeling method (Method 2, as described in the Methodology Section).

The evaluation follows the following procedure, which closely mimics how reactive safety management programs are conducted in practice:

1. For each HSID method, all TAZ's are sorted in descending order of estimated safety in terms of each of the three transportation modes (noting that the four HSID methods rank sites according to different criteria).
2. TAZs with the highest rankings are flagged as sites with promise (in practice these sites will be further scrutinized). Typically, a threshold is assigned according to safety funds available for improvement, such as the top 10% of zones. In this evaluation, both the top 10% and 5% zones are used as experimental values.

T1: Site Consistency Test Results

Using test T1 it is shown in Table 7 that the Multivariate Modeling Method outperforms the other one in identifying both top 5% and 10% hot spots (zones) with consistently higher average counts in Period 2 (Yr 2013). For example, under Method 1, all identified top 5% zones based on 2012 crash data reported 1263 pedestrian-related crash counts, while the corresponding counts are 1405 in the case of Method 2. Such phenomenon indicates Method 2 is more efficient in screening out the zones where more crashes would occur in the future period should no major changes happen in these zones.

TABLE 7: SCT Results for various transportation modes and HSID methods

	Method 1: Observed Counts		Method 2: Multivariate Modeling	
	Top 5%	Top 10%	Top 5%	Top 10%
Pedestrian	1263	1962	1405	2308
Bicyclist	1240	2014	1421	2260
Motorized Drivers	15095	23364	15137	23569
Average	5866	9113	5988	9379

Note: the values in the table represent the total accident counts of all identified zones (top 5% or 10%) in the after Period (Year 2013).

T2: The Method Consistency Test Results

Table 8 shows the number of commonly identified hot zones by alternate HSID methods over the two periods. Method 2 is superior in this test by identifying the larger number of the same hot zones in both cases of α = 0.90 and α = 0.95. In other words, Method 2 identified more zones in 2012 that were also identified as hot zones in 2013. For instance, out of the top 5% zones (11267*5%=563) in terms of bicyclist-related crashes, there are only 195 zones were identified as hot zones based on both Before and After period under Method 1. Nonetheless, the common zones for the same situation under Method 2 are 291. This fact illustrates the consistency of the method in detecting the hot zones.

Table 8: MCT Results for various transportation modes and HSID methods

	Method 1: Observed Counts		Method 2: Multivariate Modeling	
	Top 5%	Top 10%	Top 5%	Top 10%
Pedestrian	195 (34.6%)	495 (43.9%)	322 (57.2%)	727 (64.5%)
Bicyclist	191 (33.9%)	522 (46.3%)	291 (51.7%)	683 (60.6%)
Motorized Drivers	382 (67.9%)	798 (70.8%)	404 (71.8%)	823 (73.0%)
Average	256 (45.4%)	605 (53.7%)	339 (60.2%)	744 (66.0%)

Note: The number represents zones identified by methods in both periods, the percent shown in parenthesis stands for the percentage of consistent hot zones, or the percentage of hot zones identified in 2012 that were also identified in 2013.

Also shown in Table 8 are differences between percentages (shown in the parenthesis) for the three transportation modes and associated average values. There is a consistent drop in percentages as threshold values drop. The explanation is that the top zones suffer from greater random fluctuations in crashes, and thus the higher the threshold the larger are the random fluctuations and the likelihood of not being identified in a prior year.

T3: Total Rank Differences (TRD) Test Results

Table 9 illustrates that Method 2 is vastly superior using the Total Rank Differences Test. In both the cases of α = 0.90 and α = 0.95, Method 2 has significantly smaller summed ranked differences; by about 400%, 300%, and 10% compared to the Method 1, with respect to Pedestrian, Bicyclist, and motorized drivers, respectively. This result suggests that the multivariate modeling method is the better HSID method for ranking zones consistently from period to period.

Table 9: TRD Results for various transportation modes and HSID methods

	Method 1: Observed Counts		Method 2: Multivariate Modeling	
	Top 5%	Top 10%	Top 5%	Top 10%
Pedestrian	1194881	2861364	294469	831137
Bicyclist	1209692	2803658	342220	964315
Motorized Drivers	157903	649739	135216	539460
Average	854159	2104920	257302	778304

In conclusion, all of the three test results show that Method 2 is superior to Method 1.

4. Conclusions

To consistently predict traffic crashes in the SCAG region based on SCAG's existing land use model and travel demand models, transportation analysis zones (TAZs) are deployed as unit of analysis in developing traffic crash prediction module to supplement SCAG's existing transportation model structure. To this end, this study identified the influence of built environment on collisions by type (i.e. collisions between automobiles and between automobile and pedestrian or bicycle) in the SCAG region. To quantify the crash counts of various transportation modes at TAZ level, Multivariate crash frequency model was developed to jointly estimate the crash risk of different mode users. The simultaneous modeling is essential given the unobserved heterogeneity shared by various transportation modes. Ignorance of such correlation structures has been illustrated to reduce the efficiency of the model due to lesser precise parameters.

Given the crash count nature, this study utilized the Multivariate Poisson-Lognormal Model (MVPLN). Since Bayesian estimation method has shown significant advantages over traditional maximum likelihood estimation method, the MVPLN model was implemented in freeware WinBUGS which employs an MCMC algorithm for estimation of parameters. For the model calibration, two Markov chains were utilized to visually inspect trace plots of posterior estimates for convergence. Only year 2012 crash data were used to estimate the model, leaving year 2013 data for model evaluation. After the crash prediction model was created, three tests were introduced to evaluate the adopted Hot Spot Identification (HSID) method. All three test results show that the multivariate modeling method is superior to the commonly used observed count method.

Since all explanatory variables can be easily obtained from SCAG's future year travel demand models, future year high crash-risk zones can be identified using crash prediction model.

References

1. NHTSA Quick Facts 2014, National Highway Traffic Safety Administration, 2016.
2. Bicyclist Injuries Heat Map for the SCAG Region, SCAG 2016-2040 RTP/SCS http://scagrtpscs.net/Documents/2016/draft/d2016RTPSCS_ActiveTransportation.pdf
3. Scope of Work, SCAG RFP 17-037
4. SCAG Regional Travel Demand Model and 2012 Model Validation, http://www.scag.ca.gov/DataAndTools/Pages/Documents.aspx

2 Cheng, W., Gill, G. S., Dasu, R., Xie, M., Jia, X., & Zhou, J. (2017). Comparison of Multivariate Poisson lognormal spatial and temporal crash models to identify hot spots of intersections based on crash types. Accident Analysis & Prevention, 99, pp. 330-341

3 Park, E., & Lord, D. (2007). "Multivariate poisson-lognormal models for jointly modelling crash frequency by severity." Transportation Research Record, 2019(2019), pp. 1-6

4 Cheng, W., and S. Washington. New Criteria for Evaluating Methods of Identifying Hot Spots. In Transportation Research Record 2083, TRB, National Research Council, Washington, D.C., 2008, pp. 76–85

5 Congdon, P., (2006). Bayesian Statistical Modeling. 2nd ed. Wiley, New York.

6 Abdel-Aty, M., Lee, J., Siddiqui, C., Choi, K., (2013). Geographical unit based analysis in the context of transportation safety planning. Transp. Res. Part A 49, 62–75.

7 Spiegelhalter, D., Thomas, A., Best, N. and Lunn, D., 2003. WinBUGS user manual.

8 Flegal, J. M., Haran, M., & Jones, G. L. Markov chain Monte Carlo: Can we trust the third significant figure?. Statistical Science, 2016, pp.250-260.

9 Cheng, W., G. Gill, T. Sakrani, M. Dasu, J. Zhou (2017). "Predicting Motorcycle Crash Injury Severity using Weather Data and Alternative Bayesian Multivariate Crash Frequency Models". Journal of Accident Analysis and Prevention, 108, pp. 172-180.