Spatial Models for High-Accurate Hot Zone Identification for E-bikes

Xuesong Wang, Yongfeng Tang, Tongji University, China; Mohammed Quddus, Loughborough University, United Kingdom; Qingya Zhou, Guangzhou Urban Planning & Design Survey Research Institute Guangzhou, China; Xinyi Hou, Traffic Police Headquarters of Shanghai Public Security Bureau, China

TRB

PAPER NO: 21-01968

BACKGROUND

- The annual sale of e-bikes worldwide is predicted to grow from 3.3 million in 2016 to nearly 6.8 million in 2025, with a compound annual growth rate of 8.2%.
- In 2016, approximately 2 million e-bikes were sold in Europe and 210 million e-bikes were used daily in China.
- According to the statistical annual report of China's road traffic accidents in 2017, the numbers of e-bike crashes in 2011 and 2016 were 10,347 and 17,747, respectively, and the number of deaths increased by 71.5% in the five years. In addition, the number of e-bike crashes was 8.2 times larger than that of bicycle crashes and 5.4 times larger than that of pedestrian crashes.
- In Switzerland, the number of e-bikes sold increased from about 3,000 in 2006 to over 75,000 in 2016, and between year 2011 and year 2016, the numbers of injured e-bikers in police reports have more than tripled to a total of almost 700 in 2016

DATA DESCRIPTION

Summary Information of Independent Variables

Data	Variable	Description	Mean	S.D.
Road network	Road mileage	The total length of roads (km)	89.89	6595.17
	Road density	Road length per unit area (km/km²)	4.57	2.90
	Number of intersections	Total number of intersections	116.51	92.91
	Density of intersections	Number of intersections per unit area (/km²)	9.78	8.33
	Percentage of 4-legged intersections	Proportion of 4-legged intersections to total intersections (%)	42.15	11.92
Land use	Number of metro stations	Number of metro stations in the administration unit	1.70	1.81
	Urban or suburban	Location of the administration unit (0 means suburban while 1 denotes urban)	0: 63.85% 1: 36.15%	
Socio-economic	Area	Area of administration unit (km²)	30.77	33.26
	Number of households	Registered number of households (×10³)	24.56	12.43
	Registered population	Registered number of population (×10³)	63.59	33.04

METHODOLOGY

◆ **PLN CAR Models**

In order to take the potential spatial correlations for e-bike crashes into consideration, three Poisson log-normal Conditional Autoregressive (PLN CAR) models were developed. It assumed that the crash frequency followed the Poisson distribution as:

$$Y_i|\theta_i \sim Poisson(\theta_i)$$

where Y_i is the crash frequency for unit i; and θ_i denotes the expected crash frequency for unit i.

A Conditional autoregressive (CAR) model that captures the spatial dependence of crashes in adjacent units is appropriate given the research objective and the data. Based on the Bayesian framework, the prior conditional distribution can be defined as follows:

$$\phi_i|\phi_{(-i)} \sim N\left(\sum_j \frac{w_{j,i}}{w_{i+}}\phi_j, \frac{1}{\tau_c w_{i+}}\right)$$

Spatial weight matrix can take many different forms. In this study, the reciprocal of distance between centroids is used as the spatial weight matrix of the CAR model. This weight matrix is based on the assumption that the spatial correlation among units increases with the decrease of distance, which was basically in accordance with the actual situation. The following is an example of the weight matrix for a three-unit dataset, where d_{ij} denotes the distance between centroid of unit i and unit j.

$$W = \begin{bmatrix} 0 & 1/d_{12} & 1/d_{13} \\ 1/d_{21} & 0 & 1/d_{23} \\ 1/d_{31} & 1/d_{32} & 0 \end{bmatrix}$$

◆ **Model 1: Area is considered as an independent variable.**

The logarithm link was used as a function to link the expectation of Y_i with explanatory variables as:

$$ln(\theta_i) = \beta_0 + \sum_{m=1}^{M} \beta_m x_{im} + e_i + \phi_i$$

where β_0 denotes the intercept; M is the total number of independent variables; x_{im} is the value of the mth independent variable for unit i; e_i was used to account for the unobserved heterogeneity for unit i.

◆ **Model 2: Area is regarded as an exposure variable.**

The area variable is transformed into a logarithmic scale and its parameter was set to 1.0. Area in this form is able to emphasize its impact on crash prediction and the influence of different unit areal size can be well captured in this model. Other independent variables should be selected carefully to avoid the multicollinearity with the area.

The logarithm link function in Model 2 was changed to:

$$ln(\theta_i) = \beta_0 + ln(area_i) + \sum_{m=1}^{M} \beta_m x_{im} + e_i + \phi_i$$

where $area_i$ is the area of unit i; Other variables have the same meaning with Model 1.

◆ **Model 3: Area is excluded as a variable in the model.**

Other variables are normalized by dividing them with area. In this way, all the other variables in the research unit are converted into the corresponding variables per unit area. But some variables such as the percentage of 4-legged intersections were not required to divide by area because they were independent of area. Therefore, the results of hot-zone identification would not be influenced by different area scales.

The logarithm link function can be defined as:

$$ln\left(\frac{\theta_i}{area_i}\right) = \beta_0 + \sum_{m=1}^{M} \frac{\beta_m x_{im}}{area_i} + e_i + \phi_i$$

where variables have the same meanings as of Model 2.

MODEL COMPARISON AND ASSESSMENT

Mean absolute deviance (MAD), mean square prediction error (MSPE), and deviance information criterion (DIC) were utilized to evaluate the prediction and fitting performance of the models. The definitions of MAD and MSPE were as follows:

$$\mathrm{MAD} = \frac{1}{n}\sum_{\forall i} |Y_i^{pred} - Y_i^{obs}|$$

$$\mathrm{MSPE} = \frac{1}{n}\sum_{\forall i} \left(Y_i^{pred} - Y_i^{obs}\right)^2$$

where Y_i^{obs} denotes the observed crash number for unit i while Y_i^{pred} is the predicted crash number for unit i. Clearly, lower values of MAD and MSPE are preferred.

DIC is a Bayesian measure of model fitting and complexity. Smaller DIC is preferred, and it can be defined as follows:

$$\mathrm{DIC} = \overline{D(\theta)} + p_D$$

where $D(\theta)$ denotes the Bayesian deviance of the estimated parameter, and $\overline{D(\theta)}$ is the posterior mean of $D(\theta)$. $\overline{D(\theta)}$ denotes a measure of model fitting, p_D can be viewed as the effective number of parameters, which indicates the complexity of the model.

MODELING RESULTS

	Model 1			Model 2			Model 3		
	Mean	S.D.	95% BCI	Mean	S.D.	95% BCI	Mean	S.D.	95% BCI
Intercept	2.06[a]	2.00	(0.22,3.88)	0.38[a]	0.53	(0.20,0.70)	0.89	0.36	(0.48,1.32)
Road density	0.76	0.82	(0.01,2.04)	0.28	0.21	(0.13,0.37)	0.20	0.15	(0.07,0.30)
Density of intersections	—	—	—	-0.02	0.03	(-0.05,0)	-0.03	0.03	(-0.06,-0.01)
Percentage of 4-legged intersections	—	—	—	0.02	0.02	(0.01,0.04)	0.02	0.01	(0.01,0.04)
Number of household	—	—	—	0.02[a]	0.02	(0.01,0.03)	0.08[a]	0.06	(0.04,0.12)
Registered population	0.01	0.02	(0,0.02)	—	—	—	—	—	—
Number of metro stations	—	—	—	0.11	0.09	(0.02,0.20)	0.21	0.07	(0.14,0.28)
Urban or suburban	-0.41[a]	0.73	(-0.68,-0.08)	1.01	0.45	(0.63,1.43)	0.59	0.34	(0.20,1)
Area	0.01[a]	0.02	(0,0.02)	—	—	—	—	—	—
MAD	3.41			0.82			1.01		
MSPE	21.43			1.11			1.63		
DIC	5872.15			3316.82			3337.31		

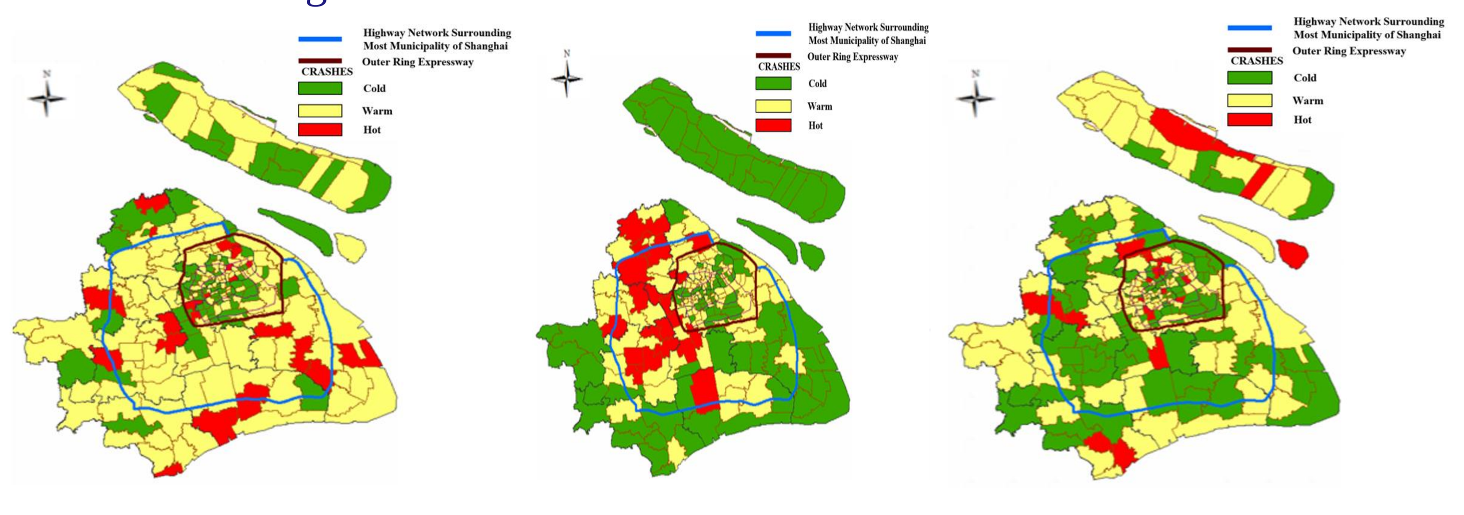
[a] : Significant at 90% BCI.

S.D.: standard error.

Based on the MAD, MSPE and DIC values, the best model was Model 2 (MAD=0.82, MSPE=1.11, DIC=3316.82), followed by Model 3 (MAD=1.01, MSPE=1.63, DIC=3337.31). Thus, Model 2 outperformed all the other models.

HOTSPOT IDENTIFICATION RESULTS

PSI was utilized as the performance measure to rank units with promise. In order to achieve a reliable result, the PSI estimates of the three models were aggregated separately. All units in the study area were classified into three categories based on the PSIs: hot, warm and cold zones. Hot zones refer to those with a top 10% PSI, warm zones are defined as the ones with a PSI between 0 and top 10%, and the remaining zones with PSI values less than 0 could be treated as the cold zones. 10% was commonly used as the threshold in many studies. The spatial distribution of units by this zone category is illustrated in Figure 1.



Model 1 Model 2 Model 3

FIGURE 1 Crash hot zones for e-bikes in three models.

The blue line was the highway network surrounding most of the municipality of Shanghai, while the ring in maroon depicted the Outer Ring Expressway. The highways can be used to roughly distinguish the suburban area, urban fringe area and urban area in Shanghai. Areas in red were those administrative units identified as hot zones, which meant their overall safety level for e-bike crashes was relatively worse than the average level of all areas. Less attention should be paid to the areas in yellow since their safety performance were better, while the overall safety level for areas in green were best in those areas.

In Model 2, the administrative units classified as hot zones mainly concentrated in the urban fringe area (i.e., within the highway network in blue, but outside the maroon line), which was in accordance with the findings of Wang and Zhou. Several administrative units adjacent to the outer ring highway and urban administrative units also had serious e-bike safety problems. In addition, hot zones for e-bike crashes mainly located in the western of Shanghai. For Model 1, large proportion of the units were identified as warm zones or hot zones, which was not reasonable in practice. In Model 3, hot zones mainly located in the urban area with small administrative units.

CONCLUSIONS

- ✓ Based on the MAD, MSPE and DIC values, model 2 outperformed all the other models;
- ✓ Interpretation of Explanatory Variables: road density, number of households, number of metro stations and urban or suburban were positively associated with e-bike crashes;
- ✓ When the area scales of units vary greatly, area should be considered as an exposure variable in the model to obtain better model performance.