Hindawi
Journal of Advanced Transportation
Volume 2019, Article ID 3521793, 13 pages
https://doi.org/10.1155/2019/3521793

Research Article

Comparative Analysis of the Reported Animal-Vehicle Collisions Data and Carcass Removal Data for Hotspot Identification

Xiaoxue Yang,[1] Yajie Zou,[1] Lingtao Wu,[2] Xinzhi Zhong,[1] Yinhai Wang,[3] Muhammad Ijaz,[1] and Yichuan Peng[1]

[1]*Key Laboratory of Road and Traffic Engineering of Ministry of Education, Tongji University, Shanghai 201804, China*
[2]*Texas A&M Transportation Institute 3135 TAMU, College Station, Texas 77843-3135, USA*
[3]*Department of Civil and Environmental Engineering, University of Washington, Washington More Hall 133B, USA*

Correspondence should be addressed to Yichuan Peng; yichuanpeng1982@hotmail.com

Received 13 October 2018; Revised 8 January 2019; Accepted 30 January 2019; Published 1 April 2019

Academic Editor: Md. Mazharul Haque

Two common types of animal-vehicle collision data (reported animal-vehicle collision (AVC) data and carcass removal data) are usually recorded by transportation management agencies. Previous studies have found that these two datasets often demonstrate different characteristics. To accurately identify the higher-risk animal-vehicle collision sites, this study compared the differences in hotspot identification and the effect of explanation variables between carcass removal and reported AVCs. To complete the objective, both the Negative Binomial (NB) model and the generalized Negative Binomial (GNB) are applied in calculating the Empirical Bayesian (EB) estimates using the animal collision data collected on ten highways in Washington State. The important findings can be summarized as follows. (1) The explanatory variables have different effects on the occurrence of carcass removal data and reported AVC data. (2) The ranking results from EB estimates when using carcass removal data and reported AVC data differ significantly. (3) The results of hotspot identification are different between carcass removal data and reported AVC data. However, the ranking results of GNB models are better than those of NB models in terms of consistency. Thus, transportation management agencies should be cautious when using either carcass removal data or reported AVC data to identify hotspots.

1. Introduction

Animal-vehicle collisions (AVCs) have always been one of research frontiers and hot topics. Van der Ree et al. [1] indicated that mortality rate of AVCs is a major concern across most of the developed countries, and it becomes more serious in the developing countries in the next few decades. It was estimated that the number of AVCs per year exceeded 1 million in the 1990s [2]. There are about 155-211 deaths, 13,713-29,000 injuries, and 1 billion dollars property loss per year caused by AVCs [2–5]. The fact that the average number of fatal AVCs was increasing year by year was inferred from the record from the NHTSA Fatality Analysis Reporting System (FARS) [4]. Previous studies found that the number of wild animals decreased significantly due to AVCs [6–8], and billions of wild animals died annually in the collision with vehicle and other types of transportation mode [9, 10].

To implement reasonable management measures with limited resources, hotspot identification (HSID), identifying sites with higher collision risk as hotspots, is an important task in the overall road safety improvement process. In recent years, researchers have proposed various HSID methods, e.g., accident frequency (AF), accident rate (AR), accident reduction potential (ARP), and Empirical Bayesian (EB). Among these methods, the EB method is adopted in this study [11–15]. Previously, researchers have mainly investigated two types of animal-vehicle collision data (number of carcass removal and reported AVCs) [16, 17]. In order to reduce the risk of AVCs and formulate effective countermeasures, transportation safety researchers have tried various statistical models to study the influence of quantitative explanation on AVCs [18], such as Poisson regression [19–22], Negative Binomial (NB) regression [23–27], Poisson–lognormal regression model [28], and Gamma regression model [29, 30].

Most previous AVCs studies considered either the reported AVC data or the carcass removal data. For the carcass removal data, Gkritza et al. [31] evaluated the effect of deterministic factors on the occurrence frequency and severity of AVCs using Poisson regression model and NB regression model. For the reported AVC data, a stepwise logistic regression model was used to identify the important factors and the high risk collision points [32–34]. Seiler [35] predicted the collision of nonincident control points through the reported AVCs data using a multiple logistic regression model. Researchers like Lao et al. [36] found a carcass found on the road is likely caused by collisions with vehicles. However, many previous studies found that the number of the carcass removal differs from the number of the reported AVCs [37–39]. The discrepancy of two AVC data sources is explained as follows. First, not all the wild animals related to the AVCs are died. Second, not all the carcasses are reported through the media.

Meanwhile, there are several researchers focusing on the difference and relationship between two datasets. A fuzzy logic–based mapping algorithm is used to merge the two incomplete datasets [36]. Lao et al. [40] developed a diagonal inflated bivariate Poisson regression model to consider the two datasets simultaneously. To predict AVCs risk, Visintin et al. [41] proposed a model that considers two types of factors: vehicles and animals.

However, few studies have compared the hotspot identification results obtained from the carcass removal and the reported AVCs data. Thus, the primary objective of this paper is to examine the difference in hotspot identification and the effect of the explanation variables on the carcass removal and the reported AVCs. To complete the objective, both the traditional NB model and the generalized Negative Binomial (GNB) are applied in calculating the EB estimates. The dispersion parameter of the NB model is fixed, while the GNB assumes the dispersion parameter varies from site to site. This study analysed the crash data collected at ten highways in Washington State.

The rest of the paper is organized as follows. The second section introduces the methodology of the EB method based on NB model and GNB model used in this study. The third section provides the data description and preliminary data analysis. The following section displays model results. The reported AVC and the carcass removal are also compared by the EB method based on the NB model and GNB model. Finally, the model results are discussed and summarized.

2. Materials and Methods

The following two sections introduce Negative Binomial model based and generalized Negative Binomial model based EB methods, respectively.

2.1. Negative Binomial Model Based Empirical Bayesian Method. The EB estimate of a site consists of two parts: predicted number of crashes from similar sites and observed number of crashes at the site. The prediction is usually based on safety performance functions (SPFs), which commonly assume the traffic counts follow some probability distributions. Until now, the NB method is the most popular approach to estimate the EB values. And the weight factor is determined by the dispersion parameter of the NB models. The NB model has the model structure below. Poisson distribution is used to assume the number of crashes during a specific time period, which is defined by

$$p(y \mid \lambda) = \frac{\lambda^y \exp(-\lambda)}{y!} \tag{1}$$

where

λ = mean response of the observation.

If the Poisson rate is assumed to be gamma distributed, the response variable follows a NB distribution. Thus, the NB distribution can be seen as a mixture of Poisson distributions. Hilbe [42] illustrated the whole derivation of the NB model. The probability density function of the NB is defined below:

$$f(y \mid \mu, \alpha) = \frac{\Gamma(y + 1/\alpha)}{\Gamma(1/\alpha)\,\Gamma(y+1)} \left(\frac{\alpha\mu}{1+\alpha\mu}\right)^y \left(\frac{1}{1+\alpha\mu}\right)^{1/\alpha} \tag{2}$$

where

y = response variable;

μ = mean of the observation; and

α = dispersion parameter.

Compared to the Poisson distribution, the NB distribution is appropriate for handling the overdispersion (that is, the variance is larger than the mean). For $y = 0, 1, 2, \ldots, \infty$, the mean of y is $E[y] = \mu$ and variance is $VAR(y) = \mu + \mu^2\alpha$. If $\alpha \longrightarrow 0$, the variance equals the mean and the NB distribution converges to the Poisson distribution.

The dispersion parameter α of the NB model is of great significance in calculating the EB estimates. Thus, the EB method is proposed to calculate the long term mean for the site i by Hauer (1992) [43]. And the EB method is shown as follows:

$$\hat{\hat{\mu}}_i = w_i \hat{\mu}_i + (1 - w_i)\, y_i \tag{3}$$

where

$\hat{\hat{\mu}}_i$=predicted number of crashes per year for site i estimated by EB method;

$\hat{\mu}_i$= predicted number of crashes per year for site i expected by the SPF;

$w_i = 1/(1 + \alpha\hat{\mu}_i)$= weight factor defined as a function of $\hat{\mu}_i$ and dispersion parameter α; and

y_i=observed number of crashes per year at site i.

Table 1: Data collection information.

Data	Data Time Covered	Date Received	Providing Agency
Reported AVCs Data	2000-2006	Apr. 2008 (Jan. 2009 update)	HSIS/ WSDOT
Carcass Removal Data	1999-2007	Jul. 2008	WSDOT
Roadlog Data	2002-2006	Apr. 2008 (Jan. 2009 update)	HSIS

2.2. Generalized Negative Binomial Model Based Empirical Bayesian Method. Traditionally, the NB models assume fixed dispersion parameter α (i.e., all sites share the same dispersion parameter), and it is used to calculate EB estimates. However, in recent years, some studies have found that the dispersion parameter α is related to the explanatory variables. They also discovered that GNB model presents better statistical adaptive performance and describes the dispersion phenomenon better [25, 44]. That is to say, the varying dispersion parameter has an impact on the EB estimates and may potentially improve the EB estimates [45]. For the GNB model, the difference of the EB estimates between the carcass removal and the reported AVCs is shown in this section.

When estimating the EB value, the weight factor will be influenced by the selection of the functional form. As discussed in a previous study [46], we considered several different functional forms to calculate the dispersion parameter α. The functional forms representing dispersion parameter of GNB model are shown as follows:

$$\text{Model 1: } \alpha_i = \gamma_0 L_i \tag{4}$$

$$\text{Model 2: } \alpha_i = \frac{\gamma_0}{L_i} \tag{5}$$

$$\text{Model 3: } \alpha_i = \gamma_0 L_i^{\gamma_1} \tag{6}$$

where

α_i= the dispersion parameter at segment i;

L_i = the segment length in miles for segment i; and

$\boldsymbol{\gamma} = (\gamma_0, \gamma_1)'$ = coefficients to be estimated.

3. Data Description and Preliminary Data Analysis

The collision dataset used in this study was collected at ten highways (I90, US2, SR8, SR20, US97, US101, US395, SR525, US12, and SR970) in Washington State. This dataset includes the reported AVC and the carcass removal data over a five-year period from 2002 to 2006 [40]. In our study, 10475 road segments are chosen as the research targets. That is, the number of the count is 10475. According to specific road characteristics (i.e., median width, lane width, and shoulder type), the highway is divided into road segments with different length. Table 1 shows the data acquisition time covered by the three main datasets used in this study. Reported AVCs dataset is collected from traffic collision records of Washington State Department of Transportation (WSDOT) and Highway Safety Information System (HSIS). Carcass removal dataset is gathered from the maintenance files recorded by the maintenance workers of WSDOT.

However, compared with the actual number of collisions, two datasets are both underreported. The reason is described as follows. (1) For reported AVCs dataset, collision is recorded only when its cost is larger than a threshold. Moreover, due to human factors of drivers, not every collision is reported to police officers. (2) For carcass removal dataset, some carcasses are hidden in roadside facilities and difficult to find. Another cause is that not each carcass is removed by professional maintenance workers. Thus, although two datasets overlap in some extent, there is a great discrepancy shown in Table 2 between two datasets.

Table 2: Frequency distribution of reported AVCs and carcass removal in the Washington data.

Crashes	Observed frequency of reported AVCs	Observed frequency of carcass removal
0	9168	8558
1	840	705
2	235	329
3	101	201
4	57	120
5	27	92
6	20	69
7	10	46
8	9	47
9	3	34
10	2	25
11	0	29
12	1	21
13	0	27
14	0	11
15	1	14
16	0	8
17	0	15
18	0	13
19	0	12
20	0	7
21-25	1	33
26-30	0	23
31-40	0	15
40+	0	21

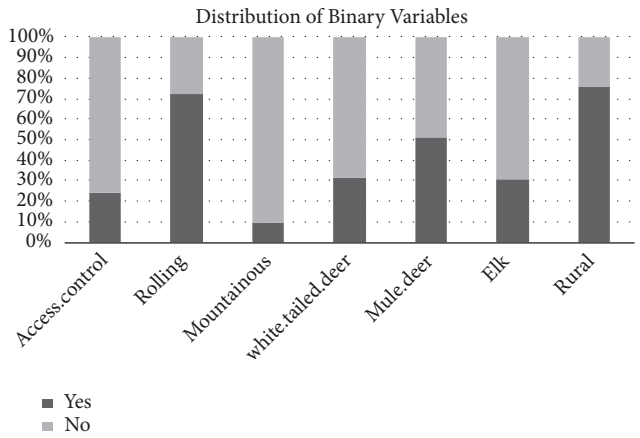


FIGURE 1: The distribution of binary variables in the defined segments.

Table 3 provides the summary statistics of characteristics for reported AVCs and carcasses in the Washington data. Apparently, the reported AVCs and carcass removal datasets differ significantly. And the number of carcass removal records is typically more than the numbers of reported AVCs data. Table 3 also describes the explanatory variables used in the models. Some variables (restrictive access control, rural or urban and terrain type, etc.) are binary variables. The percentage for binary variables is 43.75%. Figure 1 describes the distribution of binary variables in the defined segments. Annual average daily traffic (AADT) is scaled into thousands of vehicles. The model also takes three animal habitats into consideration. These three kinds of habitats include the white-tailed deer habitat, mule deer habitat, and elk habitat, since the deer and the elk are the most common animals in the AVC researches in Washington State.

Table 3 also shows the mean, maximum, minimum, and standard deviation (SD) of each variable. And the distribution of zero observations is provided in Table 4. It can be observed from Tables 3 and 4 that both the number of reported AVCs and the number of carcasses per section of a highway are overdispersed.

4. Results and Discussion

The modelling results for the carcass removal data and reported AVC data are provided in this section. This section is divided into two parts. In the first part, the NB model with fixed dispersion parameter was utilized to compare the difference between the carcass removal and the reported AVC data when estimating the number of crashes for a specific site. In the second part, the difference is analysed using the GNB model (i.e., NB model with a varying dispersion parameter).

4.1. Comparison of the Reported AVC and the Carcass Removal Data Using the NB Model with Fixed Dispersion Parameter. In the NB model, the mean functional form is shown below:

$$\mu_i = \beta_0 L_i F_i^{\beta_1} e^{\beta_2 * AC_i + \beta_3 * SL_i + \beta_4 * TP_i + \beta_5 * NL_i + \beta_6 * TR_i + \beta_7 * TM_i + \beta_8 * LW_i + \beta_9 * LSW_i + \beta_{10} * RSW_i + \beta_{11} * W_i + \beta_{12} * E_i + \beta_{13} * M_i + \beta_{14} * AT_i + \beta_{15} * MW_i} \quad (7)$$

where

μ_i= predicted numbers of collisions at segment i per year;

L_i=roadway length in miles for segment i;

F_i =flow (annual average daily traffic over five years) on segment i;

AC_i =restrictive access control for segment i;

SL_i =posted speed limit for segment i;

TP_i = truck percentage for segment i;

NL_i =total number of lanes for both directions for segment i;

TR_i =terrain type of rolling for segment i;

TM_i =terrain type of mountain for segment i;

LW_i =lane width in feet for segment i;

LSW_i =left shoulder width in feet for segment i;

RSW_i = right shoulder width in feet for segment i;

W_i =white-tailed deer habitat for segment i;

E_i =elk habitat for segment i;

M_i =mule deer habitat for segment i;

AT_i = area type (rural or urban) for segment i;

MW_i = median width for segment i; and,

$\boldsymbol{\beta} = (\beta_0, \beta_1, \beta_2, \beta_3, \beta_4, \beta_5, \beta_6, \beta_7, \beta_8, \beta_9, \beta_{10}, \beta_{11}, \beta_{12}, \beta_{13}, \beta_{14}, \beta_{15})'$ are coefficients to be estimated.

Tables 5 and 6 show the NB modelling results without insignificant variables for carcass removal and reported AVCs, respectively. For the carcass removal data, the insignificant variables are truck percentage, terrain type of mountain and right shoulder width. However, for the reported AVCs, terrain type of rolling, right shoulder width, and rural or urban type are insignificant. In summary, these insignificant variables should be eliminated when obtaining the EB estimates by NB model. On the other hand, restrictive access control is the most significant variable of reported AVCs, while the white-tailed deer habitat is the most significant variable of carcass removal.

For the possibility of the reported AVCs, AADT, speed limit, left shoulder width, white-tailed deer habitat, elk habitat, and mule deer habitat have a positive effect. And AADT, speed limit, terrain type of rolling, lane width, left

Table 3: Summary statistics of characteristics for reported AVCs and carcasses in the Washington data.

Variable code	Variable	Minimum	Maximum	Mean	SD[†]
X[a]	Number of reported AVCs per segment[c]	0	22	0.24	0.81
Y[b]	Number of carcasses per segment[c]	0	95	0.94	3.88
z1	Annual average daily traffic (in thousands)	0.31	148.8	13.85	19.76
z2	Restrictive access control (Yes: 1; No: 0)	-	-	0.24	-
z3	Posted speed limit (mph)	20	70	52.76	10.79
z4	Truck percentage (%)	0	52.28	14.05	8.29
z5	Total number of lanes for both directions	1[d]	9	2.79	1.24
z6	Roadway length (mile)	0.01	6.99	0.22	0.4
z7	Terrain type (rolling: 1; otherwise: 0)	-	-	0.720	-
z8	Terrain type (mountainous: 1; otherwise: 0)	-	-	0.096	-
z9	Lane width (feet)	10	20	12.5	1.88
z10	Left shoulder width (feet)	0	18	2.44	2.04
z11	Right shoulder width (feet)	0	20	4.03	3.52
z12	White-tailed deer habitat (yes: 1; no: 0)	-	-	0.31	-
z13	Mule deer habitat (yes: 1; no: 0)	-	-	0.51	-
z14	Elk habitat (yes: 1; no: 0)	-	-	0.31	-
z15	Median width (feet)	0	60	7.9	15.62
z16	Rural or Urban (urban: 0; rural: 1)	-	-	0.758	-

Note. [†]SD = Standard Deviation.
[a]Reported AVC data record.
[b]Carcass removal data record.
[c]Dependent variable.
[d]Six out of 10,475 segments have only one lane.
- = not applicable.

Table 4: The distribution of zero observations in the defined segments.

Data	Number of the count	Number of zero observations	Percentage of zero observations (%)
Collision Report Data	10475	9168	87.52
Carcass Removal Data	10475	8558	81.69

shoulder width, white-tailed deer habitat, and elk habitat have a positive effect on the possibility of the carcass removal.

AADT is found to increase the likelihood of both carcass removal and reported AVCs. The exposure between traffics and animals is main cause of animal-vehicle collisions. As mentioned above, it is more significant for reported AVCs than carcass removal. The cause of this is described as follows: if there is a heavier traffic flow, it is more likely that the AVCs can be reported timely since more people can notice the AVC occurrence. The coefficient of speed limit is positive, and it is less significant for reported AVCs than carcass removal. Under the condition of high speed limit, drivers prefer to choose a higher speed, and the drivers need a longer stopping distance. Thus, the driver is unlikely to stop at a safe distance to avoid a collision with an animal. As the truck percentage increases, the number of collisions will decrease. First, when a truck is traveling, it may cause a lot of noise to drive away the surrounding animals. Second, compared to smaller cars, trucks have a wider view. Moreover, drivers are likely to drive more carefully when more trucks appear. Consequently, the number of crashes will decline. Restrictive access control has a decreasing effect on the possibility of the carcass removal and the reported AVCs. The number of crashes may be smaller on the road with restrictive access control. This is because that the restrictive access control limits the animal activities. Thus, it is difficult for animals to cross the road. As a result, the number of collisions will decrease. Moreover, it is of greater significance for reported AVCs than carcass removal.

Total number of lanes is found to decrease the likelihood of both carcass removal and reported AVCs. It is more significant for carcass removal than reported AVCs. This may be because the road with more lanes is more difficult for animals to cross. And the more lanes, the easier it is to find carcass removal. Left shoulder width is found to increase the likelihood of both carcass removal and reported AVCs and the effect of left shoulder width is similar for carcass removal and reported AVCs. White-tailed deer habitat and elk habitat both have increasing effects on the possibility of the carcass removal and the reported AVCs. The finding demonstrates that collisions are prone to happen in the site with more animals. In addition, white-tailed deer habitat is less significant for reported AVCs than carcass removal, while elk habitat is of greater significance for reported AVCs than carcass removal.

Table 5: Modelling results of carcass removal for NB models with the Washington data.

Estimates	Coef.	SE∗
Intercept ln(β_0)	-7.065	0.501
ln(Average daily traffic) β_1	0.419	0.044
Restrictive access control β_2	-0.689	0.108
Posted speed limit β_3	0.051	0.003
Total number of lanes for both directions β_5	-0.449	0.041
Terrain type of rolling β_6	0.251	0.067
Lane width β_8	0.067	0.017
Left shoulder width β_9	0.105	0.017
White-tailed deer habitat β_{11}	1.271	0.061
Elk habitat β_{12}	0.351	0.061
Mule deer habitat β_{13}	-0.140	0.062
Rural or Urban β_{14}	-0.421	0.063
Median width β_{15}	-0.014	0.001
α	1.158	0.038
AIC	15714.700	
BIC	15816.290	

Note.∗ SE = standard error.

Table 6: Modelling results of reported AVCs for NB models with the Washington data.

Estimates	Coef.	SE∗
Intercept ln(β_0)	-7.810	0.641
ln(Average daily traffic) β_1	0.668	0.049
Restrictive access control β_2	-1.088	0.111
Posted speed limit β_3	0.031	0.004
Truck percentage β_4	-0.037	0.004
Total number of lanes for both directions β_5	-0.156	0.037
Terrain type of mountain β_7	-0.486	0.116
Lane width β_8	-0.080	0.029
Left shoulder width β_9	0.123	0.016
White-tailed deer habitat β_{11}	0.360	0.063
Elk habitat β_{12}	0.639	0.060
Mule deer habitat β_{13}	0.125	0.062
Median width β_{15}	-0.014	0.001
α	0.024	0.083
AIC	8454.307	
BIC	8555.890	

Note.∗ SE = standard error.

Terrain type of rolling has an increasing effect on the carcass removal, while terrain type of mountain has a decreasing effect on the reported AVCs. The cause of this phenomenon is shown as follows: in roll or mountain area, there are more animals than in level terrain. Another cause is that collision location is likely to be hidden and difficult to find in roll or mountain area [47]. Rural or urban type has a decreasing effect on the carcass removal. This may because that the carcass in urban area is more likely to be found. Median width decreases the likelihood of both carcass removal and reported AVCs and the effect is similar for carcass removal and reported AVCs. With wider median, animal activities are limited and the likelihood of crashes may decrease on the road.

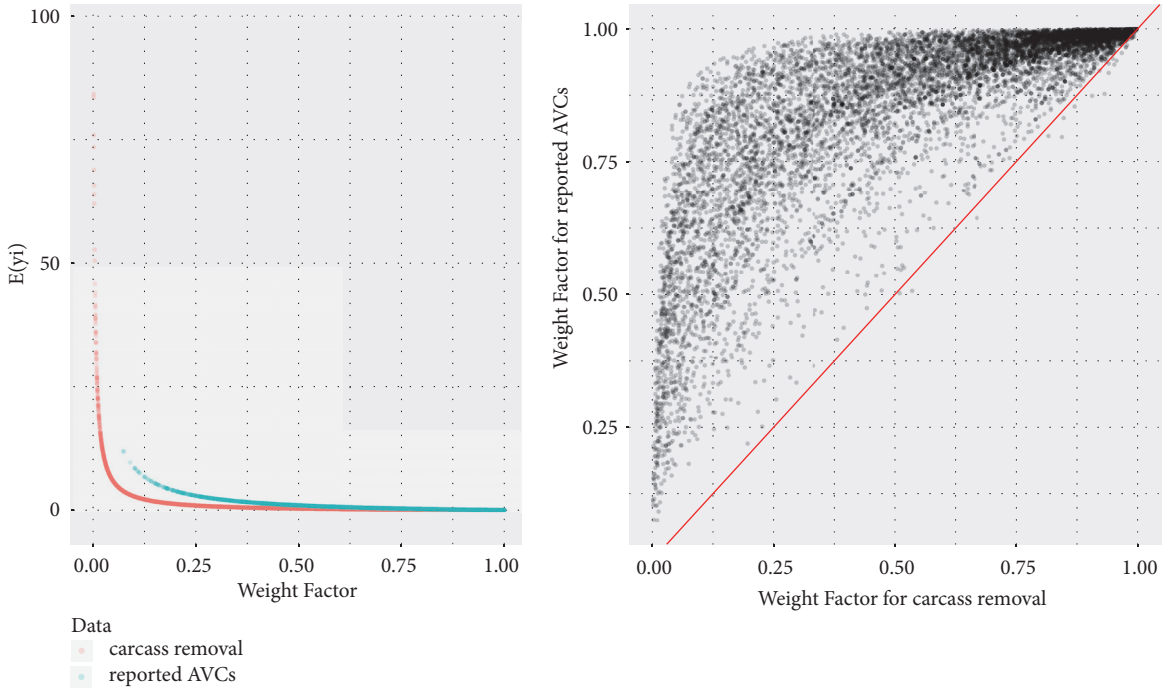
Table 7: Differences in ranking between the carcass removal and the reported AVC using the NB-based EB estimates.

Differences in ranking	Difference and percentage
Non-identical ranking	10,465(99.90%)
Ranking difference beyond 100 positions	9,824 (93.78%)
Ranking difference beyond 500 positions	7,591 (72.46%)
Ranking difference beyond 1,000 positions	5,189 (49.53%)

Note. There are 10,475 road segments in the Washington data.

Figure 2 shows the comparison of EB estimates between carcass removal and the reported AVCs. As demonstrated in Figures 2(a) and 2(b), for carcass removal and reported AVCs data, the expected number of collision and the weight factor show a similar association pattern. And the expected number of collision is inversely proportional to the weight factor. That is, smaller weight parameter is related to larger expected number of collision. In addition, since the dispersion parameter α estimated from carcass removal data is greater than the dispersion parameter α estimated from the reported AVC data, the weight parameter for reported AVC data is generally larger than the weight parameter for carcass removal data. Consequently, the EB estimates from carcass removal data will put more weight on the observed number of carcass removal data than the EB estimates from the reported AVC data.

Figure 3 shows the difference in EB estimate ranking results between the carcass removal and the reported AVC data. Smaller values of the reported AVCs ranking or the carcass removal ranking mean that the site is more dangerous. If both the carcass removal data and the reported AVC data have the similar effect on predicting the number of crashes, the distribution of the scatter in Figure 3 will be concentrated to the red line (i.e., y = x). It can be easily seen from the figure that when using the carcass removal data and the reported AVCs data to identify the hotspots, respectively, the results are very different. Further ranking comparison results are provided in Table 7. Notable difference is that the ranking differs significantly. For example, 49.53% of the results have a ranking difference beyond 1,000 positions between the carcass removal data and the reported AVCs data (note that the number of road segments in the Washington data is



(a) Relationships between the model estimates and the weight factor

(b) Associations of weight factors between carcass removal and reported AVCs

Figure 2: Weight factors produced by NB model.

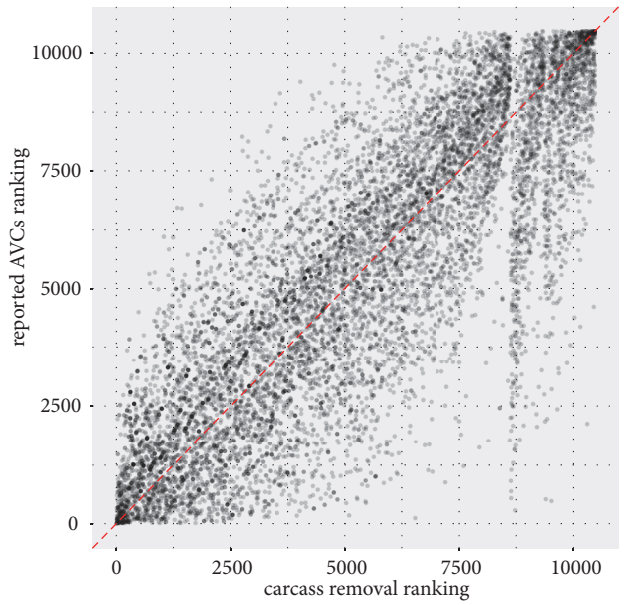


Figure 3: Comparison in HSID ranking by the carcass removal and the reported AVC using NB model.

10,475). On the whole, the type of the data will influence the identification of dangerous sites.

In order to further measure the differences between the carcass removal and the reported AVC data in identifying the hotspot, two evaluation tests are used, which are similar to the tests proposed by Cheng and Washington [48].

4.1.1. Test I. Data Consistency Test. The number of the same sites identified as hotspots using the carcass removal data and the reported AVC data is used to evaluate the performance of two types of data. The number mentioned above is defined in

$$T_I = \{k_{n-cn}, k_{n-cn+1}, \ldots k_n\}_{carcass}$$

Table 8: The number of the same sites identified as hotspots using the carcass removal data and the reported AVCs data.

Threshold level α	Number and percentage
The top 1% of the hotspots(105)	32(30.48%)
The top 5% of the hotspots(524)	219 (41.79%)
The top 10% of the hotspots(1,047)	538 (51.38%)

Note. There are 10,475 road segments in the Washington data.

$$\cap \{k_{n-cn}, k_{n-cn+1}, \ldots k_n\}_{reported} \tag{8}$$

where

T_I is the number of the same sites identified as hotspots using the carcass removal data and the reported AVC data;

n is the total number of sites;

c is the threshold of hotspots; and

k is the site ID.

The test process includes comparison across two types of AVC datasets, and we consider three cases in terms of the number of hotspots selected. The three cases correspond to considering 1%, 5%, and 10% of all sites as hotspots (i.e., c = [0.01, 0.05, 0.10]). For example, in this study, when c =0.01, a total of approximately 105 sites (i.e., about 1% of the 10,475 sites) will be considered as hotspots.

Table 8 shows the result of test I in EB estimate ranking results between the carcass removal and the reported AVC data. If the EB estimates from the carcass removal data and the reported AVC data yield similar HSID results, the number of hotspots will be equal to the threshold and the percentage will be concentrated to 100% (note that the number of road segments in the Washington data is 10,475). It can be easily seen from the table that when using the carcass removal data and the reported AVCs data to identify the hotspots, respectively, the results are different significantly.

4.1.2. Test II. Total Rank Differences Test. Taking the ranking difference into account, test II calculates the total ranking difference of the hotspots identified using the carcass removal data and the reported AVC data. Note that only c × n hotspots are considered. The test statistic for test II is shown in

$$T_{II} = \sum_{k=n-cn}^{n} \left(\Re \left(k_{carcass} \right) - \Re \left(k_{reported} \right) \right) \tag{9}$$

where

T_{II} is the total test statistic;

$\Re(k_{carcass})$ is the rank of site k obtained using the carcass removal data;

$\Re(k_{reported})$ is the rank of site k obtained using the reported AVC data; and,

k is the site ID.

Table 9: The total ranking difference of the hotspots identified using the carcass removal data and the reported AVCs data.

Threshold level c	Sum
c=1% using the carcass removal data	67,875
c=1% using the reported AVCs data	51,543
c=5% using the carcass removal data	406,430
c=5% using the reported AVCs data	407,258
c=10% using the carcass removal data	989,728
c=10% using the reported AVCs data	797,760

Note. There are 10,475 road segments in the Washington data.

The total ranking difference of the hotspots identified using the carcass removal data and the reported AVC data for different threshold levels c is provided in Table 9. For example, the sum of difference in ranks is up to 989,728 for threshold level 10% using the carcass removal data to identify the hotspots. The sum of difference in ranks using the carcass removal data is larger than that using the reported AVCs data when c= [0.01,0.1]. Moreover, when c =0.05, there is a slight difference between the two datasets. On the whole, the analysis in this part indicates that the result will be one-sided and inaccurate if we identify the hotspots only using the carcass removal data or the reported AVCs data. As a result, the type of the data will influence the identification of dangerous sites.

4.2. Comparison of the Carcass Removal and the Reported AVC by EB Method Based on the GNB Model. With the approach described in previous sections, the two datasets were analysed using GNB-based EB method with three models (i.e., (4)–(6)). Tables 10 and 11 show the modelling results with taking out insignificant variables for GNB model with the carcass removal data and the reported AVCs data, respectively. As can be seen from Tables 10 and 11, Model 3 outperforms the other two models, since Model 3 has the lowest Akaike information criterion (AIC) and Bayesian information criterion (BIC) values.

The GNB-based EB estimates for carcass removal and the reported AVCs are also compared. As observed in Figures 4(a) and 4(b), the two datasets have similar associations between the modelling values and the weight factor. In other words, the shape of the scatter distribution in the figure is approximately similar. When the modelling value $E(y_i)$ is fixed, the weight factor of the carcass removal is lower than that of the reported AVCs. We can find that a varying dispersion parameter will influence the weight factor for the crash prediction model. In addition, when adding a varying dispersion parameter to the NB model, there will be a similar influence on both the reported AVCs data and the carcass removal data.

Figure 5 presents the comparing results of hotspot identification between the reported AVC data and the carcass removal data by EB method based on the GNB model. The depth of the color in the figure represents the density of the point. As shown in Figure 5, the link between the two kinds of data based on the GNB model is more positive than that based

Table 10: Modelling results for the GNB model using the carcass removal data.

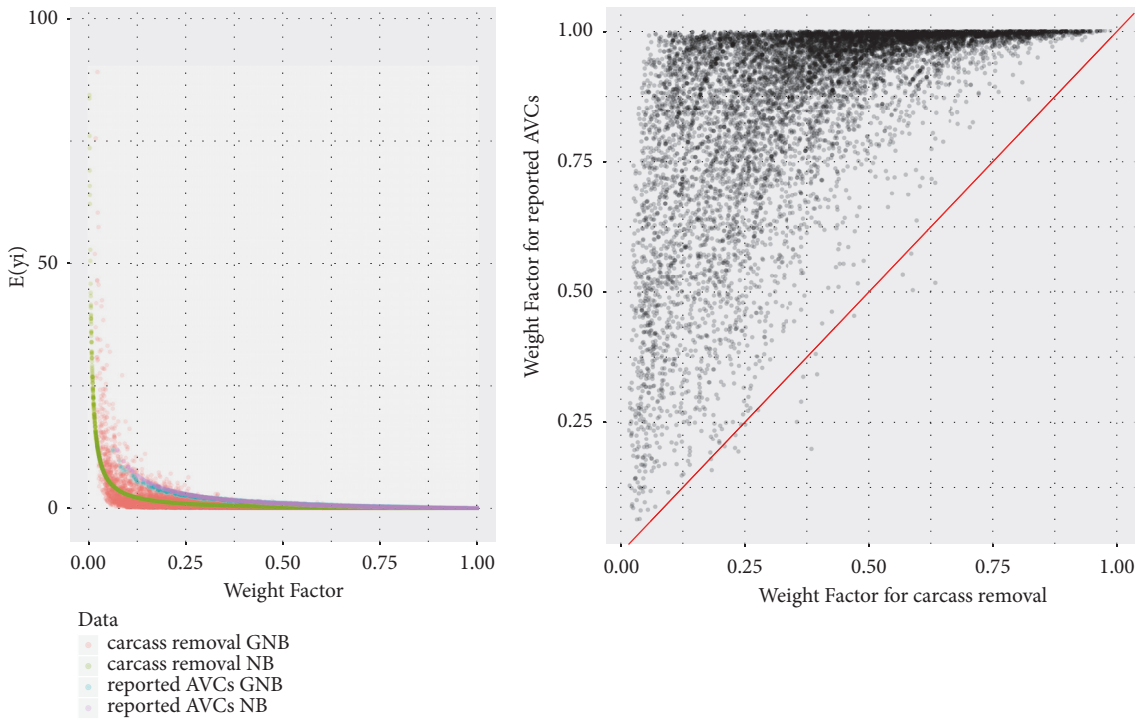
	Model 1		Model 2		Model 3	
Estimates	Value	SE	Value	SE	Value	SE
Intercept $\ln(\beta_0)$	-5.994	0.483	-8.842	0.538	-8.416	0.529
ln(Average daily traffic) β_1	0.299	0.042	0.652	0.041	0.587	0.044
Restrictive access control β_2	-0.837	0.097	-0.721	0.112	-0.702	0.112
Posted speed limit β_3	0.053	0.003	0.051	0.004	0.051	0.004
Truck percentage β_4	-0.009	0.003	-	-	-	-
Total number of lanes β_5	-0.413	0.036	-0.497	0.042	-0.491	0.042
Terrain type of rolling β_6	0.292	0.060	0.200	0.064	0.211	0.066
Lane width β_8	0.064	0.014	0.074	0.026	0.074	0.023
Left shoulder width β_9	0.087	0.014	0.104	0.019	0.109	0.018
White-tailed deer habitat β_{11}	1.252	0.053	1.173	0.056	1.212	0.059
Elk habitat β_{12}	0.305	0.053	0.320	0.057	0.341	0.060
Mule deer habitat β_{13}	-0.123	0.055	-	-	-	-
Rural or Urban β_{14}	-0.360	0.054	-0.745	0.060	-0.641	0.064
Median width β_{15}	-0.014	0.001	-0.010	0.001	-0.012	0.001
Intercept $\ln(\gamma_0)$	1.970	0.047	0.082	0.035	0.407	0.049
Segment Length γ_1	-	-	-	-	-0.702	0.030
AIC	17,334.400		15,343.670		15,250.650	
BIC	17,443.240		15,437.990		15,352.240	

Note. - = not applicable.

Table 11: Modelling results for the GNB model using the reported AVCs data.

	Model 1		Model 2		Model 3	
Estimates	Value	SE	Value	SE	Value	SE
Intercept $\ln(\beta_0)$	-7.710	0.625	-7.419	0.619	-7.795	0.641
ln(Average daily traffic) β_1	0.663	0.047	0.690	0.046	0.666	0.049
Restrictive access control β_2	-1.115	0.105	-1.041	0.106	-1.092	0.111
Posted speed limit β_3	0.031	0.004	0.028	0.004	0.031	0.004
Truck percentage β_4	-0.037	0.004	-0.038	0.004	-0.037	0.004
Total number of lanes β_5	-0.141	0.035	-0.167	0.036	-0.154	0.037
Terrain type of rolling β_6	-	-	-0.239	0.071	-	-
Terrain type of mountain β_7	-0.537	0.112	-0.651	0.122	-0.492	0.116
Lane width β_8	-0.086	0.028	-0.086	0.029	-0.081	0.028
Left shoulder width β_9	0.127	0.015	0.111	0.017	0.123	0.016
White-tailed deer habitat β_{11}	0.327	0.059	0.422	0.056	0.354	0.063
Elk habitat β_{12}	0.638	0.057	0.616	0.056	0.639	0.060
Mule deer habitat β_{13}	0.162	0.059	-	-	0.132	0.062
Rural or Urban β_{14}	-	-	-0.138	0.063	-	-
Median width β_{15}	-0.016	0.001	-0.011	0.001	-0.014	0.001
Intercept $\ln(\gamma_0)$	0.194	0.104	-1.135	0.093	0.088	0.095
Segment Length γ_1	-	-	-	-	0.125	0.096
AIC	8,507.420		8,621.486		8,454.546	
BIC	8,609.002		8,730.324		8,563.385	

Note. - = not applicable.



(a) Relationships between the model estimates and the weight factor

(b) Associations of weight factors between carcass removal and reported AVCs

Figure 4: Weight factors produced by GNB model.

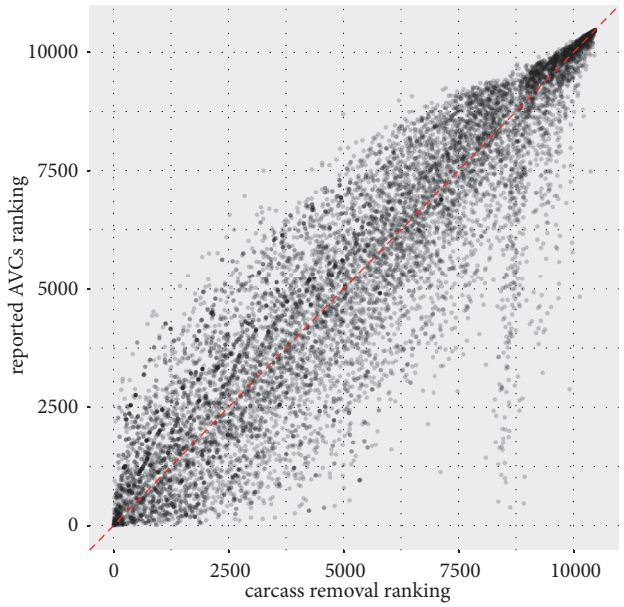


Figure 5: Comparison in HSID ranking by the carcass removal and the reported AVC using GNB model.

on the NB model. Moreover, the comparing results provided in Table 12 indicate that there is significant difference in the ranking between the reported AVCs and the carcass removal. Overall, the gap between the two kinds of data is narrowing when using the GNB model.

Similar to the NB model, two methods mentioned above are used to measure the differences between the carcass removal and the reported AVC data in identifying the hotspot. Table 13 presents the comparison results of hotspot identification between the reported AVC data and the carcass

Table 12: Differences in ranking between the reported AVC and the carcass removal using the GNB-based EB estimates.

Differences in ranking	Difference and percentage
Non-identical ranking	10,458 (99.83%)
Ranking difference beyond 100 positions	9,328 (89.05%)
Ranking difference beyond 500 positions	6,403 (61.12%)
Ranking difference beyond 1,000 positions	3,868 (36.92%)

Note. There are 10,475 road segments in the Washington data.

Table 13: The number of hotspots identified by both the carcass removal data and the reported AVCs data using the GNB-based EB estimates.

Threshold level α	Number and percentage
The top 1% of the hotspots (105)	82 (78.09%)
The top 5% of the hotspots (524)	438 (83.58%)
The top 10% of the hotspots (1,047)	863 (82.42%)

Note. There are 10,475 road segments in the Washington data.

Table 14: The sum of difference in ranks over all identified sites for threshold level α using the carcass removal data and the reported AVC data using the GNB-based EB estimates.

Threshold level c	Sum
c=1% using the carcass removal data	2,335
c=1% using the reported AVCs data	1,426
c=5% using the carcass removal data	45,029
c=5% using the reported AVCs data	20,914
c=10% using the carcass removal data	218,068
c=10% using the reported AVCs data	88,179

Note. There are 10,475 road segments in the Washington data.

removal data by the EB method based on the GNB model. As shown in Table 13, the percentage based on the GNB model is greater than that based on the NB model; that is, the link between the two kinds of data based on the GNB model is more positive. Moreover, the comparing results provided in Table 14 indicate that there is a significant difference in the ranking between the reported AVCs and the carcass removal. Furthermore, the sum of difference in ranks using the carcass removal data is different from the sum of difference in ranks using the reported AVCs data. Overall, the gap between the two kinds of data is narrowing when using the GNB model.

5. Conclusions

This paper has examined the difference between the reported AVCs data and the carcass removal data in identifying hotspots and the influence of explanatory variables. To accomplish the objectives of this study, the EB method based on the NB model and GNB model, separately, is used to model the animal crash data collected in Washington State. The important conclusions can be summarized as follows. (1) Some explanatory variables have different effects on the occurrence of carcass removal data and reported AVC data. (2) Based on the modelling results from NB and GNB models, the ranking results from EB estimates when using the carcass removal data and reported AVC data differ significantly. (3) The results of hotspot identification are significantly different between the carcass removal data and the reported AVC data. However, the ranking results with GNB models are relatively more consistent than that of NB models. Thus, transportation management agencies should be cautious when analysing the carcass removal data or reported AVC data to identify AVC-prone sites.

In this study, the EB method based on the NB model and GNB model is applied to compare the HSID results using the carcass removal and the reported AVCs data collected at ten highways in Washington State. In the future, the AVC datasets with more variables (i.e., road classification, etc.) from other sites will be collected to validate the findings from this study. In addition, spatial models should also be developed to analyse the carcass removal and the reported AVCs data [49].

Data Availability

The data used to support the findings of this study are available from the corresponding author upon request.

Disclosure

This manuscript was presented in the Transportation Research Board 97th Annual Meeting.

Conflicts of Interest

The authors declare that they have no conflicts of interest.

Acknowledgments

This research is sponsored jointly by the National Natural Science Foundation of China (Grant nos. 51608386 and 71601143), Shanghai Science and Technology Committee (Grant no. 18510745400), and Shanghai Sailing Program (Grant no. 16YF1411900).

References

[1] R. Van der Ree, R. D. J. Smith, and C. Grilo, *Handbook of Road Ecology*, John Wiley & Sons, 2015.

[2] M. R. Conover, W. C. Pitt, K. K. Kessler, T. J. Dubow, and W. A. Sanborn, "Review of human injuries, illnesses, and economic losses caused by wildlife in the United States," *Wildlife Society Bulletin*, vol. 23, no. 3, pp. 407–414, 1995.

[3] A. F. Williams and J. K. Wells, "Characteristics of vehicle-animal crashes in which vehicle occupants are killed," *Traffic Injury Prevention*, vol. 6, no. 1, pp. 56–59, 2005.

[4] R. L. Langley, S. A. Higgins, and K. B. Herrin, "Risk factors associated with fatal animal-vehicle collisions in the United States, 1995–2004," *Wilderness & Environmental Medicine*, vol. 17, no. 4, pp. 229–239, 2006.

[5] R. E. Allen and D. R. McCullough, "Deer-Car Accidents in Southern Michigan," *The Journal of Wildlife Management*, vol. 40, no. 20, pp. 317–325, 1976.

[6] F. F. van der Zee, J. Wiertz, C. J. F. Ter Braak, R. C. van Apeldoorn, and J. Vink, "Landscape change as a possible cause of the badger Meles meles L. decline in The Netherlands," *Biological Conservation*, vol. 61, no. 1, pp. 17–22, 1992.

[7] M. P. Huijser and P. J. M. Bergers, "The effect of roads and traffic on hedgehog (Erinaceus europaeus) populations," *Biological Conservation*, vol. 95, no. 1, pp. 111–116, 2000.

[8] M. Proctor, Genetic analysis of movement, dispersal and population fragmentation of grizzly bears in southwestern Canada, 2003.

[9] A. Seiler and J. O. Helldin, "Mortality in wildlife due to transportation," in *The ecology of transportation: managing mobility for the environment*, pp. 165–189, Springer, 2006.

[10] C. Ma, R. He, W. Zhang, and X. Ma, "Path optimization of taxi carpooling," *PLoS ONE*, vol. 13, no. 8, Article ID e0203221, pp. 1–15, 2018.

[11] Y. Peng, D. Lord, and Y. Zou, "Applying the Generalized Waring model for investigating sources of variance in motor vehicle crash analysis," *Accident Analysis & Prevention*, vol. 73, pp. 20–26, 2014.

[12] A. Montella, "A comparative analysis of hotspot identification methods," *Accident Analysis & Prevention*, vol. 42, no. 2, pp. 571–581, 2010.

[13] H. Yu, P. Liu, J. Chen, and H. Wang, "Comparative analysis of the spatial analysis methods for hotspot identification," *Accident Analysis Prevention*, vol. 66, pp. 80–88, 2014.

[14] Y. Zou, J. E. Ash, B.-J. Park, D. Lord, and L. Wu, "Empirical Bayes estimates of finite mixture of negative binomial regression models and its application to highway safety," *Journal of Applied Statistics*, vol. 45, no. 9, pp. 1652–1669, 2018.

[15] L. Wu, Y. Zou, and D. Lord, "Comparison of sichel and negative binomial models in hot spot identification," *Transportation Research Record: Journal of the Transportation Research Board*, vol. 2460, no. 1, pp. 107–116, 2014.

[16] M. P. Huijser, M. E. Wagner, A. Hardy, A. P. Clevenger, and J. A. Fuller, "Animal-vehicle collision data collection throughout the United States and Canada," in *proceedings of the International Conference on Ecology and Transportation*, 2007.

[17] Y. Zou, X. Zhong, J. Tang et al., "A copula-based approach for accommodating the underreporting effect in wildlife—vehicle crash analysis," *Sustainability*, vol. 11, no. 2, p. 418, 2019.

[18] K. E. Gunson, G. Mountrakis, and L. J. Quackenbush, "Spatial wildlife-vehicle collision models: A review of current work and its application to transportation mitigation projects," *Journal of Environmental Management*, vol. 92, no. 4, pp. 1074–1082, 2011.

[19] P. P. Jovanis and H.-L. Chang, "Modeling the relationship of accidents to miles traveled," *Transportation Research Record*, vol. 1068, pp. 42–51, 1986.

[20] S.-P. Miaou and H. Lum, "Modeling vehicle accidents and highway geometric design relationships," *Accident Analysis & Prevention*, vol. 25, no. 6, pp. 689–709, 1993.

[21] X. Ye, K. Wang, Y. Zou, and D. Lord, "A semi-nonparametric Poisson regression model for analyzing motor vehicle crash data," *PLoS ONE*, vol. 13, no. 5, 2018.

[22] Y. Wang, H. Ieda, and F. Mannering, "Estimating rear-end accident probabilities at signalized intersections: Occurrence-mechanisms approach," *Journal of Transportation Engineering*, vol. 129, no. 4, pp. 377–384, 2003.

[23] S.-P. Miaou, "The relationship between truck accidents and geometric design of road sections: poisson versus negative binomial regressions," *Accident Analysis & Prevention*, vol. 26, no. 4, pp. 471–482, 1994.

[24] N. V. Malyshkina and F. L. Mannering, "Empirical assessment of the impact of highway design exceptions on the frequency and severity of vehicle accidents," *Accident Analysis & Prevention*, vol. 42, no. 1, pp. 131–139, 2010.

[25] K. El-Basyouny and T. Sayed, "Comparison of two negative binomial regression techniques in developing accident prediction models," *Transportation Research Record: Journal of the Transportation Research Board*, no. 1950, pp. 9–16, 2006.

[26] Y. Zou, L. Wu, and D. Lord, "Modeling over-dispersed crash data with a long tail: Examining the accuracy of the dispersion parameter in Negative Binomial models," *Analytic Methods in Accident Research*, vol. 5-6, pp. 1–16, 2015.

[27] J. Tang, S. Zhang, X. Chen, F. Liu, and Y. Zou, "Taxi trips distribution modeling based on Entropy-Maximizing theory: A case study in Harbin city—China," *Physica A: Statistical Mechanics and its Applications*, vol. 493, pp. 430–443, 2018.

[28] D. Lord and L. F. Miranda-Moreno, "Effects of low sample mean values and small sample size on the estimation of the fixed dispersion parameter of Poisson-gamma models for modeling motor vehicle crashes: A Bayesian perspective," *Safety Science*, vol. 46, no. 5, pp. 751–770, 2008.

[29] J. Oh, S. P. Washington, and D. Nam, "Accident prediction model for railway-highway interfaces," *Accident Analysis & Prevention*, vol. 38, no. 2, pp. 346–356, 2006.

[30] R. Winkelmann and K. F. Zimmermann, "Recent developments in count data modelling: theory and application," *Journal of economic surveys*, vol. 9, no. 1, pp. 1–24, 1995.

[31] K. Gkritza, M. Baird, and Z. N. Hans, "Deer-vehicle collisions, deer density, and land use in Iowa's urban deer herd management zones," *Accident Analysis & Prevention*, vol. 42, no. 6, pp. 1916–1925, 2010.

[32] J. E. Malo, F. Suárez, and A. Díez, "Can we mitigate animal-vehicle accidents using predictive models?" *Journal of Applied Ecology*, vol. 41, no. 4, pp. 701–710, 2004.

[33] M. W. Hubbard, B. J. Danielson, and R. A. Schmitz, "Factors influencing the location of deer-vehicle accidents in Iowa," *The Journal of Wildlife Management*, vol. 64, no. 3, pp. 707–713, 2000.

[34] K. E. Rodriguez, *Modeling black bear-vehicle collision zones in*, San José State University, 2015.

[35] A. Seiler, "Predicting locations of moose-vehicle collisions in Sweden," *Journal of Applied Ecology*, vol. 42, no. 2, pp. 371–382, 2005.

[36] Y. Lao, Y.-J. Wu, Y. Wang, and K. McAllister, "Fuzzy logic-based mapping algorithm for improving animal-vehicle collision data," *Journal of Transportation Engineering*, vol. 138, no. 5, pp. 520–526, 2011.

[37] L. A. Romin and J. A. Bissonette, "Deer-vehicle collisions: Status of state monitoring activities and mitigation efforts," *Wildlife Society Bulletin*, vol. 24, no. 2, pp. 276–283, 1996.

[38] K. K. Knapp, C. Lyon, A. Witte, and C. Kienert, "Crash or carcass data: Critical definition and evaluation choice," *Transportation Research Record*, no. 2019, pp. 189–196, 2007.

[39] M. Huijser, J. Fuller, M. Wagner, A. Hardy, and A. Clevenger, *Animalvehicle collision data collection. A synthesis of highway practice. NCHRP Synthesis 370. Project 20-05/Topic 37-12*, Transportation Research Board of the National Academies, Washington, DC, USA, 2007, Online at http://www.trb.org/news/blurb_detail.asp.

[40] Y. Lao, Y.-J. Wu, J. Corey, and Y. Wang, "Modeling animal-vehicle collisions using diagonal inflated bivariate Poisson regression," *Accident Analysis & Prevention*, vol. 43, no. 1, pp. 220–227, 2011.

[41] C. Visintin, R. van der Ree, and M. A. McCarthy, "A simple framework for a complex problem? Predicting wildlife–vehicle collisions," *Ecology and Evolution*, vol. 6, no. 17, pp. 6409–6421, 2016.

[42] J. M. Hilbe, *Negative Binomial Regression*, Cambridge University Press, Cambridge, UK, 2011.

[43] E. Hauer, "Empirical bayes approach to the estimation of "unsafety": The multivariate regression method," *Accident Analysis & Prevention*, vol. 24, no. 5, pp. 457–477, 1992.

[44] S.-P. Miaou and D. Lord, "Modeling traffic crash-flow relationships for intersections: dispersion parameter, functional form, and bayes versus empirical bayes methods," *Transportation Research Record*, no. 1840, pp. 31–40, 2003.

[45] D. Lord and P. Y.-J. Park, "Investigating the effects of the fixed and varying dispersion parameters of Poisson-gamma models on empirical Bayes estimates," *Accident Analysis & Prevention*, vol. 40, no. 4, pp. 1441–1457, 2008.

[46] S. R. Geedipally, D. Lord, and B.-J. Park, "Analyzing different parameterizations of the varying dispersion parameter as a function of segment length," *Transportation Research Record*, no. 2103, pp. 108–118, 2009.

[47] Y. Wang, Y. Lao, Y.-J. Wu, and J. Corey, "Identifying high risk locations of animal-vehicle collisions on Washington state highways," *WSDOT Research Rep. WA-RD*, vol. 752, 2010.

[48] W. Cheng and S. Washington, "New criteria for evaluating methods of identifying hot spots," *Transportation Research Record*, vol. 2083, pp. 76–85, 2008.

[49] P. Xu and H. Huang, "Modeling crash spatial heterogeneity: Random parameter versus geographically weighting," *Accident Analysis & Prevention*, vol. 75, pp. 16–25, 2015.