sustainability

MDPI

Article

A Copula-Based Approach for Accommodating the Underreporting Effect in Wildlife-Vehicle Crash Analysis

Yajie Zou [1], Xinzhi Zhong [1], Jinjun Tang [2,*], Xin Ye [1,*], Lingtao Wu [3], Muhammad Ijaz [1] and Yinhai Wang [4]

1 Key Laboratory of Road and Traffic Engineering, Ministry of Education, Tongji University, Shanghai 201804, China; yajiezou@hotmail.com (Y.Z.); xinzhizhong@outlook.com (X.Z.); m.ijaz58@yahoo.com (M.I.)
2 School of Traffic and Transportation Engineering, Key Laboratory of Smart Transport in Hunan Province, Central South University, Changsha 410075, China
3 Texas A&M Transportation Institute, Texas A&M University System, 3135 TAMU College Station, TX 77843-3135, USA; wulingtao@gmail.com
4 Department of Civil and Environmental Engineering, University of Washington, More Hall 133B, Seattle, WA 98195, USA; yinhai@uw.edu
* Correspondence: jinjuntang@csu.edu.cn (J.T.); xye@tongji.edu.cn (X.Y.)

Received: 29 December 2018; Accepted: 9 January 2019; Published: 15 January 2019

Abstract: Wildlife-vehicle collision (WVC) data usually contain two types: the reported WVC data and carcass removal data. Previous studies often found a discrepancy between the number of reported WVC and carcass removal data, and the quality of both datasets is affected by underreporting. Underreporting means the number of WVCs is not fully recorded in the database; neglecting the underreporting in WVC data may result in biased parameter estimation results. In this study, a copula regression model linking wildlife-vehicle collisions and the underreporting outcome was proposed to consider the underreporting in WVC data. The WVC data collected from 10 highways in Washington State were analyzed using the copula regression model and the Negative Binomial (NB) model. The main findings from this study are as follows: (1) the Gaussian copula model can provide different modeling results when compared with the conventional modeling approach; (2) the hotspot identification results indicate that the Gaussian copula-based Empirical Bayes (EB) method can more accurately identify hotspots than the NB-based EB method. Thus, the proposed copula model may be a better alternative to the conventional NB model for modeling underreported WVC data.

Keywords: wildlife-vehicle collisions; transportation; statistical methods; maximum likelihood estimation; mathematical and statistical techniques

1. Introduction

Wildlife-vehicle collisions (WVC) are a prominent road safety issue as highway expansion projects in natural areas endanger the safe sharing of highways between vehicles and wildlife, which is a great potential threat to humans and wildlife [1,2]. WVCs cause damage to wildlife populations, as well as serious human injuries and major property loss, especially in Western countries [3–6]. Recently, the number of WVCs has been approximately 5% of all motor vehicle collisions, and the proportion is continuously rising [7,8].

WVC data generally contain two types: reported WVC data and carcass removal data [9]. In order to mitigate the wildlife-vehicle collision risk and develop effective countermeasures, statistical regression models are frequently applied by transportation safety researchers to quantify the effect of

explanatory factors on WVCs [10]. Some recent studies [11,12] provided a comprehensive overview of vehicle collision analysis methodologies, including Poisson regression [13,14], Negative Binomial (NB) regression [15–17], Poisson–lognormal regression model [18], Gamma regression model [19,20], semi-nonparametric Poisson regression model [21], etc. Using carcass removal data, Gkritza et al. [22] applied the Poisson regression model and the NB regression model to estimate the effect of identified factors on the frequency and severity of WVCs. Using reported WVC data, a stepwise logistic regression model was applied to identify the significant factors at a landscape scale and recognize the points of high collision risk [23–25]. Seiler [26] developed a multiple logistic regression model to predict collisions in non-accident control sites through the WVC data reported in observed sites. Tappe [27] proposed a multivariate approach to estimate influential factors from the county level on WVC density. Lao et al. [28] applied a diagonal inflated bivariate Poisson regression to model reported WVCs and carcass data jointly and found a correlation between the two datasets. Neumann et al. [29] related the WVC risks to the probability of road-crossings of wildlife through generalized linear mixed model construction of reported WVC data. Murphy and Xia [30] found the positive effect of coverage degree of vegetation on WVCs by using a hierarchical Bayesian model.

So far, many existing studies have focused on analyzing either the reported WVC data or carcass removal data and neglected the possible underreporting issue [31]. Underreporting refers to the number of WVCs not being fully reported and recorded. This phenomenon can be observed in the discrepancy between the reported WVCs and carcass removal data [32]. In the United States, reported WVC data are typically collected by the transportation agency, while the carcass removal data are collected by the natural resource management agency [33]. Since the two datasets are commonly collected by different agencies using distinct equipment and methods, an inconsistency always exists between them. The quality of reported WVC and carcass removal data is affected by underreporting. Rowden et al. [34] and Yannis et al. [35] point out that it is difficult to estimate the level of underreporting since the combined effect of examined temporal and spatial factors is difficult to quantify. The underreporting of reported WVC data may be due to the reporting decision of travelers, the threshold of warrant report, communication techniques, and whether incidents are recorded by transportation agencies, etc. [9,36]. Carcass removal data may be underreported due to decomposition, difficulty in detecting the carcass, tardy removal, etc. [37]. As discussed by Huijser et al. [38], nearly two-thirds of WVC go unreported in the United States. Alsop and Langley [39] applied a multivariate stepwise logistic method to identify significant factors, including age, injury severity, etc., for underreporting. Correspondingly, Yamamoto et al. [40] suggested that not considering underreporting can lead to bias when estimating the significant factors, even when using sequential binary probit models for better performance than the ordered-response probit models. Patil et al. [41] replaced a multinomial logit model with a nested logit model to accommodate underreporting for more accurate crash severity level determination. However, Snow et al. [42] suggested that WVC studies are not sensitive to underreporting until the underreporting level becomes severe. Determining the underreporting level is important, and requires WVC data availability and reliability [43].

Based on the literature review, few approaches have been developed to analyze the underreported WVC data. Thus, the primary objective of this paper is to propose a copula-based approach to accommodate the underreporting issue and accurately quantify the impact of explanatory factors on WVCs when the additional underreporting information is available. To accomplish this objective, the WVC dataset collected in Washington State from 2002 to 2006 is considered and an underreporting indicator variable is generated to denote whether the wildlife-vehicle collisions are underreported or not. To demonstrate the advantages of the proposed copula-based approach, the hotspot identification results using the proposed method and the conventional NB model are also compared.

2. Data Description

The reported wildlife-vehicle collision and carcass removal data were collected on 10 highways (US2, SR8, US12, SR20, I90, US97, US101, US395, SR525, and SR970) in Washington State over a five-year period from 2002 to 2006. The dataset has been used in some previous studies [28] and the definition of variables is explained in the AASHTO [44]. The summary statistics of characteristics of explanatory variables in Washington data are provided in Table 1. Some variables (e.g., access control type, terrain type, animal habitats, etc.) are binary variables. As shown in Table 1, the reported wildlife-vehicle collision data range from 0 to 22, and the mean wildlife-vehicle collision frequency is 0.24, with a standard deviation of 0.81. The carcass removal data have a mean value of 0.94 and a standard deviation of 3.88.

Table 1. Description of variables for the wildlife-vehicle collision data.

Variables	Minimum	Maximum	Mean	S.D. [a]
Number of reported wildlife-vehicle collisions per road segment	0	22	0.24	0.81
Number of carcasses per road segment	0	95	0.94	3.88
Underreporting indicator (Underreporting: 1; otherwise: 0)	Underreporting: 16%; otherwise: 84%			
Annual average daily traffic (AADT) over year 2002 to 2006	0.31	148.8	13.85	19.76
Restrictive access control (yes: 1; no: 0) [b]	yes: 24%; otherwise: 76%			
Posted speed limit (mph)	20	70	52.76	10.79
Truck percentage (%)	0	52.28	14.05	8.29
Median width (feet)	0	60	7.9	15.62
Total number of lanes for both directions	1	9	2.79	1.24
Roadway length (mile)	0.01	6.99	0.22	0.4
Terrain type (rolling: 1; otherwise: 0)	rolling: 72%; otherwise: 28%			
Terrain type (mountainous: 1; otherwise: 0)	mountainous: 9.6%; otherwise: 90.4%			
Lane width (feet)	10	20	12.5	1.88
Left shoulder width (feet)	0	18	2.44	2.04
Right shoulder width (feet)	0	20	4.03	3.52
Rural or Urban (urban: 0; rural: 1)	urban: 75.8%; rural: 24.2%			
White-tailed deer habitat (yes: 1; no: 0)	yes: 31%; no: 69%			
Mule deer habitat (yes: 1; no: 0)	yes: 51%; no: 49%			
Elk habitat (yes: 1; no: 0)	yes: 31%; no: 69%			

Note: [a] S.D. means Standard Deviation; [b] Restrictive access control means that access to the roadways is fully controlled; [c] 6 out of 10475 road segments have one lane.

For the reported WVC data and carcass removal data, although both data sources are underreported for different reasons, a larger number of wildlife-vehicle collisions are usually recorded in carcass removal data [31], which is also found in the data source from Washington State. In other words, carcass removal data are less likely to suffer from underreporting. However, the spatial coverage of carcass removal datasets depends on the carcass removal strategies and funding availability [45]. Due to the restriction on finances, not every regional transportation agency collects carcass removal data [42]. In this study, in order to investigate the impact of influential factors contributing to the WVC data underreporting issue, a new variable (underreporting indicator) is generated to denote whether the number of reported wildlife-vehicle collisions per road segment is underreported or not. Specifically, if the number of carcasses is larger than the number of reported wildlife-vehicle collisions, it is assumed that the number of reported wildlife-vehicle collisions for that road segment is underreported. Otherwise, the number of wildlife-vehicle collisions reported for the road segment is assumed to be the actual number. Since the carcass removal data are not often collected, the following analysis mainly considers records of reported wildlife-vehicle collisions from data sources in Washington State.

3. Methodology

3.1. The Wildlife-Vehicle Collision Model

The NB [10] distribution is the model most frequently used by transportation safety researchers to analyze wildlife-vehicle collision data. Let Y be the number of wildlife-vehicle collisions during some time period, which is assumed to follow a Poisson distribution as defined below:

$$f(y_i|\lambda) = \frac{\lambda^{y_i} \exp(-\lambda)}{y_i!}, \tag{1}$$

where y_i is the number of wildlife-vehicle collisions during some time period at site i; λ is the mean parameter of the Poisson distribution; and i is the site index.

The NB distribution arises if we let λ take a gamma distribution. For the complete derivation of the NB, more information can be found in Lord and Mannering (2010) [11]. The probability density function of the NB is defined as follows:

$$f(y_i|\mu_i, \sigma) = \frac{\Gamma(y_i + \frac{1}{\sigma})}{\Gamma(\frac{1}{\sigma})\Gamma(y_i + 1)} \left(\frac{\sigma\mu_i}{1 + \sigma\mu_i}\right)^{y_i} \left(\frac{1}{1 + \sigma\mu_i}\right)^{1/\sigma}, \tag{2}$$

where $\mu_i = \exp(\mathbf{x}_i\boldsymbol{\beta})$ is the expected number of wildlife-vehicle collisions during some time period at site i; $\mathbf{x}_i$ is a vector of covariates, and $\boldsymbol{\beta}$ is a vector of the regression coefficients; and σ is the dispersion parameter.

3.2. The Underreporting Outcome Model

The underreporting indicator variable Z can be considered as a dichotomous variable (underreporting: 1; otherwise: 0). Thus, a logistic regression model is adopted here to analyze the impact of explanatory variables on the underreporting outcome, which is defined as follows:

$$z_i = \frac{\exp(\mathbf{x}_i\boldsymbol{\gamma})}{1 + \exp(\mathbf{x}_i\boldsymbol{\gamma})}, \tag{3}$$

where z_i is the probability that the number of wildlife-vehicle collisions during some time period at site i is underreported; $\mathbf{x}_i$ is a vector of covariates; and $\boldsymbol{\gamma}$ is a vector of the regression coefficients.

3.3. Linking the Wildlife-Vehicle Collision Model and the Underreporting Outcome Model

The occurrence of wildlife-vehicle collisions and the probability of underreporting are affected by the explanatory variables. To capture the link between the occurrence of wildlife-vehicle collisions and the underreporting probability, a bivariate copula approach is proposed to describe variables (Y, Z). The joint cumulative density function (cdf) can be defined as follows:

$$F(y, z) = C(F_1(y), F_2(z)), \tag{4}$$

where $C : [0, 1]^2 \rightarrow [0, 1]$ is the copula function; $F_1(y)$ is the cdf of NB distribution for the number of reported wildlife-vehicle collisions model; and $F_2(z)$ is the cdf of binomial distribution for the underreporting outcome.

One advantage of the copula model is that the marginal distributions for variables (Y, Z) can be different [46]. Some recent studies have comprehensively summarized different families of bivariate copula models [47–49], which are provided in Table 2. Some studies also applied different families of copula in transportation data analysis [50–53]. The commonly used families of bivariate copulas are summarized in Table 2. Note that some bivariate copulas only allow a moderate correlation between two variables.

Table 2. The characteristics of different families of bivariate copulas.

Name	Copula $C(u,v;\ \theta)$ [a]	Parameter Range of θ	Parameter Range of Kendall's tau
Gaussian	$\Phi_2\left(\Phi^{-1}(u),\Phi^{-1}(v),\theta\right)$ [b]	$\theta \in (-1,\ 1)$, $\theta = 0$ is independence	$\tau = (2/\pi)\sin^{-1}(\theta)$, $\tau \in (-1,\ 1)$
Farlie-Gumbel-Morgenstern	$uv[1+\theta(1-u)(1-v)]$	$\theta \in (-1,\ 1)$, $\theta = 0$ is independence	$\tau = \frac{2}{9}\theta$, $\tau \in \left(-\frac{2}{9}, \frac{2}{9}\right)$
Ali-Mikhail-Haq	$\frac{uv}{1-\theta(1-u)(1-v)}$	$\theta \in [-1,1)$, $\theta \to 0$ is independence	$\tau = \frac{3\theta-2}{3\theta} - \frac{2(1-\theta)^2}{3\theta^2}\ln(1-\theta)$, $-0.182 < \tau < 0.333$
Clayton	$\left(u^{-\theta}+v^{-\theta}-1\right)^{-1/\theta}$	$\theta \in (0,\infty)$, $\theta \to 0$ is independence	$\tau = \frac{\theta}{\theta+2}$, $0 < \tau < 1$
Frank	$-\frac{1}{\theta}\ln\left(1+\frac{\left(e^{-\theta u}-1\right)\left(e^{-\theta v}-1\right)}{e^{-\theta}-1}\right)$	$\theta \in (-\infty,\infty)\backslash\{0\}$, $\theta \to 0$ is independence	$\tau = 1-\frac{4}{\theta}[1-D_1(\theta)]$ [c], $\tau \in (-1,\ 1)$
Gumbel	$\exp\left(-\left[(-\ln u)^{\theta}+(-\ln v)^{\theta}\right]^{1/\theta}\right)$	$\theta \in [1,\infty)$, $\theta = 1$ is independence	$\tau = 1-\theta^{-1}$, $0 < \tau < 1$
Joe	$1-\left[(1-u)^{\theta}+(1-v)^{\theta}-(1-u)^{\theta}(1-v)^{\theta}\right]^{1/\theta}$	$\theta \in [1,\infty)$, $\theta = 1$ is independence	$\tau = 1+\frac{4}{\theta}D_J(\theta)$ [d], $0 < \tau < 1$

Note: [a] u and v represent the marginal cdfs of (Y,Z), respectively; [b] Φ_2 represents the standard cdf of bivariate normal distribution and Φ^{-1} denotes the inverse cdf of the standard univariate normal distribution; [c] $D_1(\theta)$ is the first order Debye function; and, [d] $D_J(\theta) = \int_{t=0}^{1} \frac{[\ln(1-t^{\theta})](1-t^{\theta})}{t^{\theta-1}}dt$.

Let $F(y,z)$ be the joint cdf of occurrence of wildlife-vehicle collisions and the underreporting probability, the copula regression is formulated as follows:

$$F(y,z) = C(F_1(y|\mathbf{x}_1, \boldsymbol{\beta}, \sigma), F_2(z|\mathbf{x}_2, \boldsymbol{\gamma}); \theta), \tag{5}$$

where $F_1(y|\mathbf{x}_1, \boldsymbol{\beta}, \sigma)$ and $F_2(z|\mathbf{x}_2, \boldsymbol{\gamma})$ are marginal cdfs of Y and Z; $\mathbf{x}_j$ is a vector of explanatory variables, $j = 1, 2$; $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$ are vectors of the regression coefficients; and, σ is the dispersion parameter of variable Y; θ is the dependence parameter of the copula. Note that if $\theta = 0$, then the formulation in Equation (5) corresponds to the multiplication of $F_1(y|\mathbf{x}_1, \boldsymbol{\beta}, \sigma)$ and $F_2(z|\mathbf{x}_2, \boldsymbol{\gamma})$. Under this circumstance, $F_1(y|\mathbf{x}_1, \boldsymbol{\beta}, \sigma)$ and $F_2(z|\mathbf{x}_2, \boldsymbol{\gamma})$ represent independent NB regression model and logistic regression model for describing the number of reported wildlife-vehicle collisions and underreporting probability, respectively.

The parameters of copula model are estimated using a penalized maximum likelihood method [54]. Specifically, the log-likelihood function is given as:

$$\ell(y, z|\boldsymbol{\beta}, \boldsymbol{\gamma}, \sigma, \theta) = \sum_{i=1}^{n} \ln f(y_i, z_i|\boldsymbol{\beta}, \boldsymbol{\gamma}, \sigma, \theta), \tag{6}$$

where $f(y,z)$ is the joint density function of the copula model. Note that variable Z is a dichotomous variable that takes on the value 1 or 0. The parameters of copula model can be estimated by maximizing the log-likelihood function: $\text{argmax}\ell(t_1, t_2|\mathbf{x}_1, \mathbf{x}_2, \boldsymbol{\Theta})$. Detailed explanation about the parameter estimation method can be found in [54].

4. Modeling Results

4.1. Variables Affecting the Number of Reported Wildlife-Vehicle Collisions and the Underreporting Outcome

In this section, the mean functional form for modeling the number of reported wildlife-vehicle collisions is described below:

$$\mu_i = \beta_0 L_i F_i^{\beta_1} \exp(\beta_2 x_{2i} + \ldots + \beta_m x_{mi}), \tag{7}$$

where μ_i is the expected number of wildlife-vehicle collisions at site i; L_i represents the length of roadway segment in miles for site i; F_i is the average daily traffic over five years traveling on site i; $x_{2i}, \ldots, x_{mi}$ are the explanatory variables included in the functional form at site i; $\boldsymbol{\beta} = (\beta_0, \beta_1, \beta_2, \beta_3, \ldots, \beta_m)'$ are the estimated coefficients; and m is the number of explanatory variables.

Similarly, the underreporting indicator is modeled using all explanatory variables:

$$z_i = \frac{\exp(\gamma_0 + \sum_{g=1}^{m} x_{gi}\gamma_g)}{1 + \exp(\gamma_0 + \sum_{g=1}^{m} x_{gi}\gamma_g)}, \tag{8}$$

where z_i is the expected probability of site i is underreported; $x_{1i}, \ldots, x_{mi}$ are the explanatory variables; and $\boldsymbol{\gamma} = (\gamma_0, \gamma_1, \ldots, \gamma_m)'$ are the estimated coefficients.

To consider the possible link between the number of reported wildlife-vehicle collisions and the underreporting outcome, different families of copula models are used. Interestingly, the goodness-of-fit statistics for all copula models differ slightly. Among the different bivariate copula models, the Gaussian copula model has a simple function and its dimension can be easily expanded to trivariate form. Thus, in the following analysis of the parameter estimation results, we mainly focus on comparing the effects of different explanatory variables on underreporting outcome and reported wildlife-vehicle collisions using the independence copula (the logistic regression model and NB regression model are estimated independently) and Gaussian copula.

The modeling results for underreporting outcome and reported wildlife-vehicle collisions using the Gaussian copula model and the independent copula model are shown in Table 3. From Table 3, the results for some estimates of identified explanatory variables contributing to the underreporting indicator in Gaussian copula are consistent with those of independent copula. Similarly, the modeling results for number of reported WVCs are also partially consistent between the Gaussian copula and the independent copula. Note that the positive or negative values of estimates indicate the increasing or decreasing effect on underreporting probability and number of reported WVCs. Highway segments with restrictive access control impose a decreasing effect on the underreported indicator and reported WVCs. Restrictive access control facilities can restrict wildlife's access to the road segments, which decreases the number of WVCs. The estimated values of total number of lanes for both directions are all negative for underreported indicator and reported WVCs in Gaussian copula and independent copula. Posted speed limit and left shoulder width are shown as positive variables on the underreported indicator and reported WVCs. The terrain type being rolling or mountainous is found to increase the underreporting probability, while it decreases the number of reported WVCs.

Table 3. Modeling results for underreporting outcome and reported wildlife-vehicle collisions using the Gaussian copula model and the independent copula model.

	Gaussian Copula Model	**Independent Copula Model**
Underreporting indicator variable	Estimate (Std. Error)	Estimate (Std. Error)
Intercept	−4.103 (0.355)	−4.221 (0.355)
Average daily traffic	-1.181×10^{-5} (4.77×10^{-6})	-*
Restrictive access control	−0.780 (0.160)	−0.874 (0.157)
Posted speed limit	0.034 (0.006)	0.039 (0.006)
Total number of lanes for both directions	−0.159 (0.063)	−0.249 (0.052)
Segment length	1.187 (0.091)	1.229 (0.085)
Terrain type: rolling	0.588 (0.115)	0.575 (0.115)
Terrain type: mountainous	0.304 (0.156)	0.315 (0.157)
Left shoulder width	0.085 (0.012)	0.080 (0.012)
White-tailed deer habitat	1.274 (0.082)	1.250 (0.082)
Elk habitat	0.491 (0.078)	0.503 (0.078)
Mule deer habitat	−0.288 (0.084)	−0.274 (0.083)
Number of reported wildlife-vehicle collisions variable	Estimate (Std. Error)	Estimate (Std. Error)
Intercept	−6.240 (0.811)	−8.718 (0.596)
Ln (Average daily traffic)	0.497 (0.057)	0.690 (0.052)
Restrictive access control	−1.050 (0.141)	−0.958 (0.127)
Posted speed limit	0.059 (0.007)	0.028 (0.006)
Truck percentage	−0.036 (0.005)	−0.036 (0.005)
Total number of lanes for both directions	−0.252 (0.048)	−0.177 (0.043)
Terrain type: rolling	−0.244 (0.094)	−0.213 (0.084)
Terrain type: mountainous	−0.742 (0.154)	−0.680 (0.140)
Lane width	−0.132 (0.045)	-*
Left shoulder width	0.057 (0.011)	0.057 (0.010)
White-tailed deer habitat	0.583 (0.075)	0.523 (0.067)
Elk habitat	0.654 (0.075)	0.705 (0.066)

Note: * Insignificant variables at the 0.05 level of significance.

There is also an estimation difference between the Gaussian copula and the independent copula for the underreported indicator and the number of reported WVCs. AADT is found to be a significant contributor that decreases the underreporting probability in a Gaussian copula, while, interestingly, AADT is not identified as a significant explanatory variable for the underreported indicator in an independent copula using the logistic regression model. AADT are identified with the positive effect on number of reported WVCs since more observers may call the police when WVCs occur on a highway segment with heavier traffic. Under such conditions, the number of reported WVCs is close

to the actual number of WVCs so that the underreporting probability is low. Therefore, the result of a Gaussian copula seems more reasonable for the underreported indicator to include AADT as a negative factor. Wider lanes result in a smaller number of reported WVCs for Gaussian copula, but are insignificant when using the independent copula.

4.2. Comparison of the Hotspot Identification Results Using the Gaussian Copula-Based EB Method and NB-Based EB Method

The Empirical Bayes (EB) method introduced by Hauer et al. [55] is adopted as a state-of-the-art hotspot identification (HSID) method and is recommended in the Highway Safety Manual for roadway safety management [56]. In this section, the safety performance function estimated from the Gaussian copula model and NB model is used to calculate the EB estimates. To compare the hotspot identification accuracy using the modeling results from the copula model and the NB model, it is important to know the true wildlife-vehicle collision risk at each site. As discussed in the data description section, the reported wildlife-vehicle collision data and carcass removal data are both underreported to some extent, and a larger value is generally observed in carcass removal data [31]. In this section, a new variable is defined as follows:

$$\eta_i = Max\{number\ of\ reported\ wildlife-vehicle\ collisions,\ number\ of\ carcasses\} \tag{9}$$

where for site i, η_i denotes the larger value between the number of reported wildlife-vehicle collision and the number of carcasses. When conducting the HSID analysis, variable η_i is adopted to possibly reflect the real wildlife-vehicle collision risk of each site. Hereinafter, the number of reported wildlife-vehicle collisions and the underreporting indicator is denoted as the training data. Variable η_i along with all explanatory variables are considered as the validation data. Similar to the three tests proposed by Cheng and Washington (2008), three performance measures are used to evaluate the hotspot identification results from the Gaussian copula-based EB method and the NB-based EB method.

Measure I

This measure is based on the idea that collision-prone sites will usually exhibit high collision frequencies, which means that a desirable HSID method should identify hotspot sites as those that can be expected to have poor safety performance. The measure considers the sites identified as hotspots by the copula-based EB method and NB-based EB method using the building data, and compares the methods based on the sum of η_i at collision-prone site i. The optimal HSID method as determined by this measure is the one with the largest number of wildlife-vehicle collision (η_i) occurring on the sites identified as high risk by that method. Out of n sites, a threshold is defined for each method such that $c \times n$ are designated as high risk by each method. Then, the following test statistic (Equation (10)) is computed for the two EB methods and the preferred method is denoted as the one which yields the highest value of $T_{I(j)}$ [57].

$$T_{I(j)} = \sum_{k=n-cn}^{n} \eta_{k,method=j}, \tag{10}$$

where $T_{I(j)}$ is the total number of wildlife-vehicle collisions from validation data; n is the total number of sites under analysis; η_k is the number of collisions for site k defined in Equation (9); c is the threshold for high-risk sites, defined as the fraction of all sites n that are designated high-risk; and j is the Gaussian copula-based EB method or the NB-based EB method.

Measure II

Measure II evaluates the performance of two EB methods by the extent to which the same sites are identified as hotspots using the building data and validation data. The better performing method

is evaluated by the larger number of identified hotspots that are consistent between the building data and validation data, which is defined in Equation (11):

$$T_{II(j)} = \{k_{n-cn}, k_{n-cn+1}, \ldots k_n\}_{j,\text{ buliding data}} \cap \{k_{n-cn}, k_{n-cn+1}, \ldots k_n\}_{j,\text{ testing data}}, \quad (11)$$

where $T_{II(j)}$ is the total number of the same sites identified using the building data and validation data; j is the Gaussian copula-based EB method or the NB-based EB method; and k is the site index.

Measure III

Measure III takes into account the safety performance ranking assigned by two EB methods, and estimates performance based on the ranking consistency between the building data and validation data. The performance of each method is calculated as the sum of differences between the rank assigned to all $c \times n$ high-risk sites for building data and the rank assigned to the same $c \times n$ sites for validation data. The test statistic for measure III is computed in Equation (12):

$$T_{III(j)} = \sum_{k=n-cn}^{n} (\mathfrak{R}(k_{j,\text{ buliding data}}) - \mathfrak{R}(k_{j,\text{ validation data}})), \quad (12)$$

where $T_{III(j)}$ is the total test statistic for method j; $\mathfrak{R}(k_{j,\text{ buliding data}})$ is the rank of site k obtained from the method j using building data; $\mathfrak{R}(k_{j,\text{ validation data}})$ is the rank of site k obtained from the method j using validation data; j is the Gaussian copula-based EB method or the NB-based EB method; and, k is the site index.

The Gaussian copula-based EB method and the NB-based EB method are evaluated using the three proposed measures. Specifically, for the building data, both the Gaussian copula-based EB method and the NB-based EB method are used to identify poor safety performance sites. For the validation data, the NB regression is applied to estimate the safety performance function and the NB-based EB method is considered to label the true high-risk sites. As all test procedures involve comparison across two wildlife-vehicle collision datasets, we consider three different scenarios in terms of the number of high-risk sites selected for consideration under each HSID method. These scenarios correspond to considering 1%, 5%, and 10% of all sites as high-risk (i.e., c = [0.01, 0.05, 0.10]). For example, in this study, when c = 0.05, a total of approximately 224 sites (i.e., about 5% of the 4474 sites) will be considered as high-risk.

Table 4 shows the results of the two HSID methods. For measure I, the underlying principle is that high-risk sites identified by two EB methods should also show high collision counts $T_{I(j)}$. And thus the larger the value of $T_{I(j)}$, the better performing the corresponding EB method is. From Table 4, for all cases (c = 0.01, 0.05 and 0.1 of sites are considered as high-risk), the copula-based EB method provides better accuracy than the NB-based EB method. Measure II is adopted to assess the consistent identification of the same high-risk sites using the building data and validation data. Similarly, higher value of test statistic $T_{II(j)}$ indicates the better performance of that HSID method. It can be observed that across all three scenarios, the copula-based EB method is preferred. For Measure III, the ranking of sites identified as high-risk using the building data are compared to the rankings of the same sites using the validation data. Thus, smaller value of the test statistic $T_{III(j)}$ suggests better performance of the HSID method. As shown in Table 4, the copula-based EB method yields smaller test statistic values than the NB-based EB method. In sum, the copula-based EB method consistently demonstrates better HSID performance than the NB-based EB method. The possible explanation for this is that since the copula model considers the underreporting issue associated with each site and the corresponding safety performance function estimated from the Gaussian copula model can better reflect the actual collision risk of the site.

Table 4. Test statistics of three measures using the Gaussian copula-based EB method and NB-based EB method.

Measures	Threshold Values		
Method I	**c = 0.01**	**c = 0.05**	**c = 0.10**
Copula model	1031	2842	4056
NB model	921	2624	3822
Method II	**c = 0.01**	**c = 0.05**	**c = 0.10**
Copula model	13	86	213
NB model	12	75	189
Method III	**c = 0.01**	**c = 0.05**	**c = 0.10**
Copula model	8337	90,657	236,760
NB model	11,490	114,093	294,874

5. Discussion and Conclusions

This research applied the copula regression model to examine the impact of underreporting on wildlife-vehicle collision data analysis. The proposed Gaussian copula model was compared with the conventional NB model for analyzing the effects of explanatory variables using the WVC data collected from Washington State. To evaluate the HSID results from the Gaussian copula-based EB method against the NB-based EB method, a new variable to reflect the actual WVCs risk of each site was proposed and three HSID performance measures were adopted. The major findings can be summarized as follows: (1) For some explanatory variables, the Gaussian copula model provided different modeling results compared with the independent model (logistic regression model and NB model). A further examination suggested that the estimates of parameters for some variables from the independent model were inappropriate (for example, AADT is not identified as the significant explanatory variables for affecting the probability of underreporting). Neglecting the underreporting of the WVC data may result in biased parameter estimation results. (2) For the considered Washington WVC dataset, the Gaussian copula-based EB method can more accurately identify the hotspots than the NB-based EB method. Since the Gaussian copula-based model can consider the underreporting of WVC data, the proposed approach can generally provide more accurate safety performance. Thus, the HSID accuracy can possibly be improved by properly considering the underreporting of WVC data. Although the proposed Gaussian copula-based EB method is not ready yet, transportation safety analysts may use this approach to calculate the EB estimates for underreported WVC data.

Since both reported WVC data and carcass removal data are underreported to some extent, it is hard to know the true number of WVCs. Thus, to further validate the findings from this study, the WVC data collected from other regions with different characteristics should be examined using the proposed Gaussian copula model. In addition, as discussed by Wu et al. [58], when using the simulated data, the true safety state of each road segment can be known and the true hotspots can be identified. Thus, in the future, a simulation study can be designed to examine the performance of the Gaussian copula model in identifying the hotspot. This study adopts the total crash count to identify hotspots. Crash severity (i.e., fatal, incapacitating injury, non-incapacitating injury, etc.) and collision type (i.e., rear-end, etc.) can be also considered for HSID.

Author Contributions: Conceptualization, Y.Z. and X.Z.; Methodology, Y.Z., J.T. and X.Y.; Software, Y.Z. and X.Z.; Validation, Y.Z., J.T. and X.Y.; Formal Analysis, Y.Z. and X.Z.; Investigation, Y.Z. and X.Z.; Resources, M.I.; Data Curation, Y.W.; Writing-Original Draft Preparation, Y.Z.; Writing-Review & Editing, Y.Z., J.T. and X.Y.; Visualization, M.I.; Supervision, Y.Z.; Project Administration, Y.Z.; Funding Acquisition, Y.Z.

Funding: This research was funded by the National Natural Science Foundation of China (Grant No. 51608386), the Shanghai Science and Technology Committee (Grant No. 18510745400), and the Shanghai Sailing Program (Grant No. 16YF1411900).

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Hughes, W.E.; Saremi, A.R.; Paniati, J.F. Vehicle-animal crashes: An increasing safety problem. *ITE J.* **1996**, *66*, 24–29.
2. Michael, C. *Resolving Human—Wildlife Conflict: The Science of Wildlife Damage Management*; CRC Press: Boca Raton, FL, USA, 2001.
3. Huijser, M.P.; Bergers, P.J.M. The effect of roads and traffic on hedgehog (*Erinaceus europaeus*) populations. *Biol. Conserv.* **2000**, *95*, 111–116. [CrossRef]
4. Proctor, M.F. Genetic analysis of movement, dispersal and population fragmentation of grizzly bears in southwestern Canada. In Proceedings of the International Conference on Ecology & Transportation, Lake Placid, NY, USA, 24–29 August 2003; pp. 186–193.
5. Zee, F.F.V.D.; Wiertz, J.; Braak, C.J.F.T.; Apeldoorn, R.C.V.; Vink, J. Landscape change as a possible cause of the badger *Meles meles* L. decline in The Netherlands. *Biol. Conserv.* **1992**, *61*, 17–22.
6. Russo, F.; Comi, A. From the analysis of European accident data to safety assessment for planning: The role of good vehicles in urban area. *Eur. Transport Res. Rev.* **2017**, *9*, 9. [CrossRef]
7. Huijser, M.P.; Mcgowen, P.T.; Fuller, J.; Hardy, A.; Kociolek, A. Wildlife-Vehicle Collision Reduction Study: Report to Congress. 2007. Available online: https://www.fhwa.dot.gov/publications/research/safety/08034/ (accessed on 12 January 2019).
8. Tamas, C.; Janos, F. Annual trends in the number of wildlife-vehicle collisions on the main linear transport corridors (highway and railway) of Hungary. *N.-West. J. Zool.* **2015**, *11*, 41–50.
9. Huijser, M.P.; Wagner, M.E.; Hardy, A.; Clevenger, A.P.; Fuller, J.A. Animal-Vehicle Collision Data Collection Throughout the United States and Canada. Road Ecology Center, 2007. Available online: https://escholarship.org/uc/item/573094wr (accessed on 12 January 2019).
10. Gunson, K.E.; Mountrakis, G.; Quackenbush, L.J. Spatial wildlife-vehicle collision models: A review of current work and its application to transportation mitigation projects. *J. Environ. Manag.* **2011**, *92*, 1074–1082. [CrossRef]
11. Lord, D.; Mannering, F. The statistical analysis of crash-frequency data: A review and assessment of methodological alternatives. *Transp. Res. Part A Policy Pract.* **2010**, *44*, 291–305. [CrossRef]
12. Mannering, F.L.; Bhat, C.R. Analytic methods in accident research: Methodological frontier and future directions. *Anal. Methods Accid. Res.* **2014**, *1*, 1–22. [CrossRef]
13. Jovanis, P.P.; Chang, H.L. Modeling the relationship of accidents to miles traveled. *Transport. Res. Rec.* **1986**, *1068*, 42–51.
14. Miaou, S.P.; Lum, H. Modeling vehicle accidents and highway geometric design relationships. *Accid. Anal. Prev.* **1993**, *25*, 689–709. [CrossRef]
15. Miaou, S.P. The relationship between truck accidents and geometric design of road sections: Poisson versus negative binomial regressions. *Accid. Anal. Prev.* **1994**, *26*, 471–482. [CrossRef]
16. El-Basyouny, K.; Sayed, T. Comparison of Two Negative Binomial Regression Techniques in Developing Accident Prediction Models. *Transp. Res. Rec. J. Transp. Res. Board* **2006**, *1950*. [CrossRef]
17. Malyshkina, N.V.; Mannering, F.L. Empirical assessment of the impact of highway design exceptions on the frequency and severity of vehicle accidents. *Accid. Anal. Prev.* **2010**, *42*, 131–139. [CrossRef] [PubMed]
18. Lord, D.; Miranda-Moreno, L.F. 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. *Saf. Sci.* **2008**, *46*, 751–770. [CrossRef]
19. Oh, J.; Washington, S.P.; Nam, D. Accident prediction model for railway-highway interfaces. *Accid. Anal. Prev.* **2006**, *38*, 346–356. [CrossRef] [PubMed]
20. Winkelmann, R.; Zimmermann, K.F. Recent developments in count data modelling: Theory and application. *J. Econ. Surv.* **2010**, *9*, 1–24. [CrossRef]
21. Ye, X.; Wang, K.; Zou, Y.; Lord, D. A semi-nonparametric Poisson regression model for analyzing motor vehicle crash data. *PLoS ONE* **2018**, *13*, e0197338. [CrossRef]
22. Gkritza, K.; Baird, M.; Hans, Z.N. Deer-vehicle collisions, deer density, and land use in Iowa's urban deer herd management zones. *Accid. Anal. Prev.* **2010**, *42*, 1916–1925. [CrossRef]
23. Malo, J.E.; Suarez, F.; Diez, A. Can we mitigate animal–vehicle accidents using predictive models? *J. Appl. Ecol.* **2004**, *41*, 701–710. [CrossRef]

24. Hubbard, M.W.; Danielson, B.J.; Schmitz, R.A. Factors influencing the location of deer-vehicle accidents in Iowa. *J. Wildl. Manag.* **2000**, *64*, 707–713. [CrossRef]
25. Rodriguez, K.E. Modeling Black Bear-Vehicle Collision Zones in Yosemite National Park. Master's Thesis, San Jose State University, San Jose, CA, USA, 2015.
26. Seiler, A. Predicting locations of moose–vehicle collisions in Sweden. *J. Appl. Ecol.* **2005**, *42*, 371–382. [CrossRef]
27. Tappe, P.A. County-Level Factors Contributing to Deer-Vehicle Collisions in Arkansas. *J. Wildl. Manag.* **2011**, *71*, 2727–2731.
28. Lao, Y.; Wu, Y.J.; Corey, J.; Wang, Y. Modeling animal-vehicle collisions using diagonal inflated bivariate Poisson regression. *Accid. Anal. Prev.* **2011**, *43*, 220–227. [CrossRef] [PubMed]
29. Neumann, W.; Ericsson, G.; Dettki, H.; Bunnefeld, N.; Keuler, N.S.; Helmers, D.P.; Radeloff, V.C. Difference in spatiotemporal patterns of wildlife road-crossings and wildlife-vehicle collisions. *Biol. Conserv.* **2012**, *145*, 70–78. [CrossRef]
30. Murphy, A.; Xia, J. Risk analysis of animal–vehicle crashes: A hierarchical Bayesian approach to spatial modelling. *Int. J. Crashworthiness* **2016**, *21*, 1–13. [CrossRef]
31. Donaldson, B.; Lafon, N. Personal Digital Assistants to Collect Data on Animal Carcass Removal from Roadways. *Transp. Res. Rec. J. Transp. Res. Board* **2010**, *2147*, 18–24. [CrossRef]
32. Romin, L.A.; Bissonette, J.A. Deer: Vehicle Collisions: Status of State Monitoring Activities and Mitigation Efforts. *Wildl. Soc. Bull.* **1996**, *24*, 276–283.
33. Gunson, K.E.; Chruszcz, B.; Clevenger, A.P. Large animal-vehicle collisions in the Central Canadian Rocky Mountains: Patterns and characteristics. *BMC Health Serv. Res.* **2003**, *10*, 1–8.
34. Rowden, P.; Steinhardt, D.; Sheehan, M. Road crashes involving animals in Australia. *Accid. Anal. Prev.* **2008**, *40*, 1865–1871. [CrossRef]
35. Yannis, G.; Papadimitriou, E.; Chaziris, A.; Broughton, J. Modeling road accident injury under-reporting in Europe. *Eur. Transport Res. Rev.* **2014**, *6*, 425–438. [CrossRef]
36. Lao, Y.; Zhang, G.; Wu, Y.-J.; Wang, Y. Modeling animal–vehicle collisions considering animal–vehicle interactions. *Accid. Anal. Prev.* **2011**, *43*, 1991–1998. [CrossRef] [PubMed]
37. Snow, N.P.; Andelt, W.F.; Stanley, T.R.; Resnik, J.R.; Munson, L. Effects of roads on survival of San Clemente Island foxes. *J. Wildl. Manag.* **2012**, *76*, 243–252. [CrossRef]
38. Huijser, M.P.; Mcgowen, P.T.; Clevenger, A.P.; Ament, R. *Wildlife-Vehicle Collision Reduction Study: Best Practices Manual*; The U.S. Federal Highway Administration (FHWA): Washington, DC, USA, 2008.
39. Alsop, J.; Langley, J. Under-reporting of motor vehicle traffic crash victims in New Zealand. *Accid. Anal. Prev.* **2001**, *33*, 353–359. [CrossRef]
40. Yamamoto, T.; Hashiji, J.; Shankar, V.N. Underreporting in traffic accident data, bias in parameters and the structure of injury severity models. *Accid. Anal. Prev.* **2008**, *40*, 1320–1329. [CrossRef] [PubMed]
41. Patil, S.; Geedipally, S.R.; Lord, D. Analysis of crash severities using nested logit model—Accounting for the underreporting of crashes. *Accid. Anal. Prev.* **2012**, *45*, 646–653. [CrossRef]
42. Snow, N.P.; Porter, W.F.; Williams, D.M. Underreporting of wildlife-vehicle collisions does not hinder predictive models for large ungulates. *Biol. Conserv.* **2015**, *181*, 44–53. [CrossRef]
43. Tavasszy, L.; de Jong, G. Data availability and model form. In *Modelling Freight Transport*; Elsevier: New York, NY, USA, 2014; pp. 229–244.
44. American Association of State Highway and Transportation Officials. *Policy on Geometric Design of Highways and Streets*; American Association of State Highway and Transportation Officials: Washington, DC, USA, 2001.
45. Knapp, K. Crash Reduction Factors for Deer-Vehicle Crash Countermeasures: State of the Knowledge and Suggested Safety Research Needs. *Transp. Res. Rec. J. Transp. Res. Board* **2005**, *1908*, 172–179. [CrossRef]
46. Genest, C.; Favre, A.-C. Everything you always wanted to know about copula modeling but were afraid to ask. *J. Hydrol. Eng.* **2007**, *12*, 347–368. [CrossRef]
47. Bhat, C.R.; Eluru, N. A copula-based approach to accommodate residential self-selection effects in travel behavior modeling. *Transp. Res. Part B Methodol.* **2009**, *43*, 749–765. [CrossRef]
48. Zou, Y.; Zhang, Y. A copula-based approach to accommodate the dependence among microscopic traffic variables. *Transp. Res. Part C Emerg. Technol.* **2016**, *70*, 53–68. [CrossRef]

49. Zou, Y.; Ye, X.; Henrickson, K.; Tang, J.; Wang, Y. Jointly analyzing freeway traffic incident clearance and response time using a copula-based approach. *Transp. Res. Part C Emerg. Technol.* **2018**, *86*, 171–182. [CrossRef]
50. Wang, K.; Yasmin, S.; Konduri, K.C.; Eluru, N.; Ivan, J.N. Copula-based joint model of injury severity and vehicle damage in two-vehicle crashes. *Transp. Res. Rec. J. Transp. Res. Board* **2015**, 158–166. [CrossRef]
51. Rana, T.; Sikder, S.; Pinjari, A. Copula-based method for addressing endogeneity in models of severity of traffic crash injuries: Application to two-vehicle crashes. *Transp. Res. Rec. J. Transp. Res. Board* **2010**, 75–87. [CrossRef]
52. Yasmin, S.; Eluru, N.; Pinjari, A.R.; Tay, R. Examining driver injury severity in two vehicle crashes–A copula based approach. *Accid. Anal. Prev.* **2014**, *66*, 120–135. [CrossRef] [PubMed]
53. Nashad, T.; Yasmin, S.; Eluru, N.; Lee, J.; Abdel-Aty, M.A. Joint modeling of pedestrian and bicycle crashes: Copula-based approach. *Transp. Res. Rec. J. Transp. Res. Board* **2016**, 119–127. [CrossRef]
54. Marra, G.; Wyszynski, K. Semi-parametric copula sample selection models for count responses. *Comput. Stat. Data Anal.* **2016**, *104*, 110–129. [CrossRef]
55. Hauer, E.; Harwood, D.W.; Griffith, M.S. Estimating Safety by the Empirical Bayes Method: A Tutorial. *Transp. Res. Rec.* **2002**, *1784*, 126–131. [CrossRef]
56. Zou, Y.; Ash, J.E.; Park, B.-J.; Lord, D.; Wu, L. Empirical Bayes estimates of finite mixture of negative binomial regression models and its application to highway safety. *J. Appl. Stat.* **2018**, *45*, 1652–1669. [CrossRef]
57. Cheng, W.; Washington, S. New Criteria for Evaluating Methods of Identifying Hot Spots. *Transp. Res. Rec. J. Transp. Res. Board* **2008**, *2083*, 76–85. [CrossRef]
58. Wu, L.; Zou, Y.; Lord, D. Comparison of Sichel and Negative Binomial Models in Hot Spot Identification. *Transp. Res. Rec. J. Transp. Res. Board* **2014**, *2460*, 107–116. [CrossRef]