

# Article Crash Risk Assessment of Off-Ramps, Based on the Gaussian Mixture Model Using Video Trajectories

Ting Xu <sup>1,2</sup>, Yanjun Hao <sup>1,\*</sup>, Shichao Cui <sup>1,\*</sup>, Xingqi Wu <sup>1</sup>, Zhishun Zhang <sup>1</sup>, Steven I-Jy Chien <sup>1,3</sup> and Yulong He <sup>2</sup>

- <sup>1</sup> College of Transportation Engineering, Chang'an University, Xi'an 710064, China
- <sup>2</sup> College of Metropolitan Transportation, Beijing University of Technology, Beijing 100124, China
- <sup>3</sup> Department of Civil and Environmental Engineering, New Jersey Institute of Technology, Newark, NJ 07102-1982, USA
- \* Correspondence: 2019122089@chd.edu.cn (Y.H.); 2019122088@chd.edu.cn (S.C.); Tel.: +86-183-0918-6200 (Y.H.); +86-135-7208-1739 (S.C.)

Received: 12 March 2020; Accepted: 9 April 2020; Published: 11 April 2020



MDP

**Abstract:** The focus of this paper is the crash risk assessment of off-ramps in Xi'an. The time-to-collision (TTC) is used for the measurement and cross-comparison of the crash risk of each location. Five sites from the urban expressway in Xi'an were selected to explore the TTC distribution. An unmanned aerial vehicle and a camera were used to collect traffic flow data for 20 min at each site. The parameters, including speed, deceleration rate, truck percentage, traffic volume, and vehicle trajectories, were extracted from video images. The TTCs were calculated for each vehicle. The Gaussian mixture model (GMM) was proposed to predict the TTC probability density functions (PDFs) and cumulative density functions (CDFs) for five sites. The Kolmogorov–Smirnov (K-S) test indicated that the samples followed the estimated GMM distribution. The relationship between the crash risk level and influencing factors was studied by an ordinal logistic regression model and a naive Bayesian model. The results showed that the naive Bayesian model had an accuracy of 86.71%, while the ordinal logistic regression model and a cumulative distribution is could be applied to the real-time collision warning system.

**Keywords:** time-to-collision; Gaussian mixture model; risk assessment; E-M algorithm; ordinal logistic regression model; naive Bayesian

# 1. Introduction

An urban expressway is a high-speed multi-lane highway with access ramps and median dividers. The expressway is an important component of the urban transportation system in China's major developed cities. The urban expressways have distinct features, such as short spacing between entrance or exit ramps, complex connections with frontage roads [1], and short deceleration lanes, which increase the probability of accidents. The off-ramps of urban expressways suffer from more crashes than other segments due to frequent, lane-changing behaviors. According to the crash report (Traffic Management Bureau, Ministry of Public Security 2017), there were 6116 crashes and 1472 fatalities on expressways in 2016. In addition, it was reported that urban expressway crashes make up 5.6% of total roadway crashes in Xi'an and resulted in 673,370 RMB of economic loss in 2011 [2]. In Shanghai, crashes at off-ramps accounted for 21.6% of total expressway crashes [3]. Actually, the crashes are likely to occur in the vicinity of ramps. Hence, it is very important to predict and evaluate the crash risk of ramps. However, the random circumstances, as well as various missing data increase the difficulty when it comes to the solid accuracy of safety evaluations.

Many factors play a role in the safety of off-ramps, including annual average daily traffic (AADT), ramp configurations, ramp lengths, deceleration lanes, acceleration lanes [4], and the number of lanes [1].

It is found that ramp AADT, the number of exit ramp lanes, ramp alignment [5], speed-changing lane length [6–8], vehicle characteristics (steering, acceleration, and speed) [9,10], the number of lanes of ramps [11,12], driving behaviors [13,14], etc., are closely related to ramp crashes.

The traditional models are used to predict the number of crashes including Poisson, negative binomial [5,15], and generalized linear models with the main effect variables [16,17]. The fact that the crash data are sometimes incomplete becomes a driving force for the studies of surrogate safety measures (SSMs) [18]. Such studies have become popular, which analyze the traffic conflict models [19], and several of these SSMs are also popular to evaluate the severity of a conflict [20,21]. The time-to-collision (TTC) is one of the most widely used indictors to detect dangerous situations [22–25]. If both vehicles continued traveling at the same constant speeds, a crash would occur after a time. Therefore, the probability of a crash is calculated by using the TTC. This indicator is inversely related to crash risks (smaller TTC values indicate higher crash risks) [26–28]. TTC is significantly affected by the road environment, traffic flow condition, driver characteristics, weather [21,29], and so on. It is found that the TTC decreases with an increase in traffic density [30].

Crash risk estimation is to calculate the crash likelihood before crashes, and the related models can be broadly classified into statistical methods, classification trees, and artificial intelligence. The statistical model is based on a certain probability distribution. Therefore, the traditional crash frequency modeling cannot be conducted. The case-control logistics model is employed to qualify the explanatory variables and risk level, such as the logistic regression and the Tobit regression [31,32]. The Bayesian network model emphasize the logical cause of crashes [33]. Moinul et al. [34] used a Bayesian belief network for a basic freeway segment, and the classified rate was only 66%. The non-parametric classification tree is usually applied to injury severity forecast. The well-known artificial neural network (ANN) paradigms were investigated for crash risk estimation. The reliability of the results relies on three features: network architecture, the model of the neuron, and learning algorithms [35]. Lee et al. [36] determined the collision warning in the car-following state by a multi-layer perceptron neural network. The output was the level of collision severity.

The previous studies were mainly focused on freeway segments rather than on off-ramps. Few research works studied the collision risks at ramps. The limitation of the logistic regression model is that there is no dependency assumption among the influencing risk factors. The critical values of the severity of risks for expressways are unclear [37–39]. The TTC distribution is uncertain in practice. The value of 15% TTC is considered as a threshold of traffic conflicts. In order to avoid bias, the severity prediction of the conflicts is ignored in most studies. The classification tree and neural network do not require any specific functional form to design a model of the risk factors. However, they cannot interpret the relationship between factors and crash severity. The Bayesian network model is time consuming when samples have a large dimension.

The crash risk level is measured by the TTC value. In order to qualify the relationship between contributing factors and the severity of crash risks, the crash risk is divided into different severity levels. However, the crash risk division in the previous studies was unclear. The real TTCs' distribution is explored to determine the transitions of the safety states. The crash risk levels are classified according to the TTCs' distributions. Since the risk level is an ordinal response variable, an ordinal method is employed to analyze crash risks. In this study, naive Bayesian is designed to overcome the time consumption problem.

The contributions of this research are as follows: (1) to explore the TTCs distributions with GMM at off-ramps; (2) to determine the critical values for the crash risk severity; (3) the naive Bayesian model is developed to explore the relationship between the severity of the TTC and explanatory variables. An ordinal logistic regression model is designed for comparison at the same time.

The remainder of the paper is organized as follows. Section 2 introduces the traffic flow analysis at off-ramps. Section 3 is the TTC distribution functions' estimation. The ordinal logistic regression model and the naive Bayesian model are developed and compared for crash risk estimation in Section 4. The discussion and conclusions will be given at the end.

# 2. Data Processing

### 2.1. Data Collection

In order to explore the crash risks upstream from expressway off-ramps, we selected five sites from the urban expressway in Xi'an, which is a six-lane expressway, and the lane width is 3.75 m. The five sites are shown in Figure 1. The speed limit is 100 km/h for passenger cars and 80 km/h for trucks. The off-ramp influence area is 150 m past the ramp and 200 m before, and there is no on-ramp in this influence area. The investigated off-ramps are at straight and flat segments, connecting parallel to accesses. The length of the deceleration lane is 140 m. The traffic signs are placed at 2 km, 1 km, and 500 m away from exits. The vehicles are divided into passenger cars and trucks, according to "Technical Standard for Highway Engineering 2016". The vehicles for which the axis distance is greater than 3.8 m are trucks. Otherwise, the vehicles are considered as passenger vehicles. An unmanned aerial vehicle was used to collect traffic flow for 20 min at each site, from 2 December to 3 December, during the morning peak hour. Images were used from a camera mounted on a nearby pedestrian overcrossing.



Figure 1. Five sites in Xi'an Ring Road.

## 2.2. Statistical Analysis of Traffic Flow

As mentioned earlier, the inner lane, middle lane, and outer lane are referred to as Lanes 1, 2, and 3, respectively. Vehicle trajectories were extracted from an unmanned aerial vehicle by Tracker at a frequency of 5 Hz. A one-second video was composed of 25 frames. The vehicle trajectories were extracted from the video images by the following steps:

Step 1: A coordinate reference system was set up in the video to calculate vehicle positions at different times.

Step 2: In this study, the 3.75 m width of the lane was used to calibrate the position of vehicles with the ground plane.

As shown in Figure 2, the pedestrian overcrossing and the upper-bound road constituted a coordinate reference system. Figure 3 is the vehicle trajectories extracted from the video. The red

trajectories represent the vehicles in the mainline. The blue trajectories indicate the vehicles out of the mainline. The green trajectories represent vehicles in the auxiliary lane.



Figure 2. The coordinate reference system in Tracker.



Figure 3. The vehicle trajectories extracted from the video.

Each vehicle was considered as a pixel, and the vehicle was tracked automatically. We extracted 1552 trajectories as samples at off-ramps. An eight second video was used for analysis in the observation area. Because we calculated the vehicle position at 5 Hz, there were 40 position data for each trajectory record.

In this study, the Grubbs outlier test was used to detect the outliers based on the spot speed, as Equation (1). The outliners were removed from the trace data to eliminate the error and improve the accuracy. The missing data were replaced by the mean of the adjacent points to obtain a full trace.

The correctness of each sample was calculated as Equation (1).

$$Q_c = 1 - \frac{\text{The number of outliers}}{50} \tag{1}$$

Therefore, the correctness of the subsample was calculated as follows:

$$Q_c = \frac{1}{50} \sum_{c=1}^{50} Q_c = 95.33\%$$

The correctness of the total sample was 95.33%. This showed that the quality of the extracted data was sufficient to meet the modeling requirements.

The average speed, 85% speed, speed standard deviation (speed S.D.), and traffic volume for each lane are statistically summarized in Table 1. The traffic composition was calculated for Lane 3.

No.	Sites	Lane No.	Average Speed (km/h)	85th Percentile Speed(km/h)	Volume (veh/h)	Truck Percentage (%)	Diverging Rate(%)	Speed S.D. (km/h)	N
		1	34.70	45.79	265	_		6.34	55
1	Western Zhangba	2	31.39	42.01	363	-	17.27%	7.42	121
	Exit	3	27.83	39.02	261	13.33%		10.36	87
	E ( CL'1''' '	1	40.75	49.50	287	_		6.37	83
2	Eastern Shilijinxiu	2	37.62	45.94	309	-	9.3%	7.89	123
	Exit	3	33.91	43.55	101	10.80%		8.14	67
		1	45.61	48.74	352	_		5.17	92
3	Eastern North Rd.	2	40.86	44.82	336	-	14.6%	6.15	112
	Beidian Exit	3	33.01	40.07	171	16.00%		7.14	57
		1	37.30	44.57	287	_		8.15	76
4	South King Kd.	2	33.16	41.22	358	-	19.27%	8.14	151
	Haidebao Exit	3	25.88	38.56	274	14.13%		10.78	88
	Eastern Ring Road	1	38.30	48.42	322	-		5.14	74
5	Engineering	2	34.74	44.86	462	-	6.71%	6.33	154
	University	3	31.72	40.54	68	6.25%		7.15	56

**Table 1.** The investigation information for each site.

Table 1 reveals that the average speed of off-ramps ranged from 25.88 km/h to 45.61 km/h, and the 85% speed ranged from 38.56 km/h to 49.50 km/h. The speed of Lane 1 was the highest of the three lanes, and the 85% speed was above 44.57 km/h. The speeds of Lane 2 and Lane 3 were similar. The speed S.D. of Lane 1 was higher than the other lanes, which illustrated larger speed gaps among vehicles. The 85% speed of Lane 3 was the lowest, usually below 40 km/h, due to the great impact that lane-changing had on it. The traffic volumes were between 697 veh/h and 919 veh/h. Traffic flow for Lane 2 was higher than the other lanes. The diverging rate described the proportion of exiting vehicles at the current off-ramp. Due to large residential communities, the diverging rates were higher at Site 1 and Site 4. The highest diverging rate as 19.27% at Site 4, and the lowest was 6.71% at Site 5. The truck percentage ranged from 6.25% to 23.8%. The speed decreased with the increased presence of trucks in the traffic flow.

The speed S.D. for the five sites of each lane is shown in Figure 4. It can be seen from Figure 4 that Lane 3 had the maximum speed S.D. and Lane 1 had the minimum speed S.D. at each location, which indicated that Lane 3 had a wider range of speed variation and a greater risk of crashes compared with other lanes.



Figure 4. The speed S.D. at each site.

# 3. Methodology

### 3.1. TTC Definition

TTC is defined as the time required for two vehicles to collide if they continue running at their present speed while on the same path [23]. The two vehicles are in a car-following state, as seen below in Figure 5. This indicator illustrates the crash risks in TTC.



Figure 5. Car-following state.

TTC of off-ramps in the mainline can be calculated by using Equation (2), as shown below.

$$TTC_f(t) = \begin{cases} \frac{L_l(t) - L_f(t) - l_l(t)}{V_f(t) - V_l(t)}, V_f(t) > V_l(t) \\ \infty, \text{ otherwise} \end{cases},$$
(2)

where  $TTC_f(t)$  denotes the TTC of the following vehicle.  $L_l(t)$  denotes the position of the leading vehicle at a certain time *t*.  $L_f(t)$  denotes the position of the following vehicle at a certain time *t*.  $l_l(t)$  denotes the leading vehicle.  $V_l(t)$  denotes the speed of the leading vehicle.  $V_f(t)$  denotes the speed of the following vehicle.

The distance between the rear of the leading vehicle and the front of the following vehicle can be represented as  $L_l(t) - L_f(t) - l_l(t)$ . The speed difference between the two vehicles can be represented as  $V_f(t) - V_l(t)$ . Therefore, the TTC of individual vehicles can be calculated by this equation.

# 3.2. TTC Calculation

The lane-change process at off-ramps is show in Figure 6.



Figure 6. The lane-change process at off-ramps.

After the observation, we found that the vehicle often made a lane-change from the middle lane or the outer lane to the auxiliary lane to diverge from the mainline.

TTC for each vehicle is calculated as Equation (2). The values of 15% TTC, 50% TTC, and 85% TTC are summarized in Table 2.

Site No.	Lane No.	15% TTC (s)	50% TTC (s)	85% TTC (s)	TTC S.D. (s)
	1	7.56	14.34	28.36	13.64
1	2	6.34	16.53	26.65	15.42
	3	5.03	10.35	14.050	9.34
	1	15.63	19.37	29.43	18.24
2	2	14.33	21.56	27.42	22.35
	3	9.03	16.35	22.84	18.56
	1	17.77	23.04	27.64	24.56
3	2	14.58	18.54	23.65	18.77
	3	8.06	14.76	18.31	13.24
	1	18.48	25.67	28.33	24.87
4	2	13.78	23.45	26.65	22.44
	3	3.05	11.25	16.052	11.67
	1	23.65	29.57	34.73	29.57
5	2	17.54	21.68	24.56	20.32
	3	7.06	11.34	19.118	7.35

Table 2. Times-to-collision (TTCs) of Xi'an urban expressway.

When comparing the TTCs of three lanes, the TTCs were distributed in a wider range in Lane 3. The 85% TTCs were smaller than 34.73 s for the expressway. The 50% TTCs ranged from 10.35 s to 29.67 s. The smaller TTCs indicated an increased probability of crash occurrences. The 15% TTCs of Lane 1 and Lane 2 were between 7.56 s and 23.65 s. The TTC of Lane 3 was significantly smaller than other lanes, since Lane 3 was disturbed by the interruption of vehicles merging from the other lane. The 15% TTCs ranged from 3.05 s to 9.03 s, which indicated that Lane 3 was more dangerous than the other lanes. TTC frequency distribution and TTC cumulative distribution are depicted in Figure 7a,b.



Figure 7. TTC distribution: (a) TTC frequency distribution; (b) TTC cumulative distribution.

Figure 7a reveals the TTC distributions for three lanes. Figure 7b is the cumulative frequency of TTCs for all three lanes. For Lane 3, the 15% TTC, 50% TTC, and 85% TTC were 3.4 s, 7.5 s, and 16.7 s, respectively. For Lane 1 and Lane 2, TTC almost overlapped. The 15% TTC, 50% TTC, and 85% TTC were 3.3 s, 9.7 s, and 16.7 s, respectively. The 50% TTCs of Lane 1 and Lane 2 were greater than Lane 3.

Cross-sectional comparisons for the five sites are shown above in Figure 8. The smaller TTC indicated a higher risk of crash. The 15% TTC, 50% TTC, and 85% TTC for Site 1 and Site 4 were smaller than the other sites, which indicated a higher crash risk level. Therefore, Site 1 and Site 4 were more dangerous than the other three sites.



**Figure 8.** Cross-sectional comparison: (**a**) 15% TTC for each lane at each location; (**b**) 50% TTC for each lane at each location; (**c**) 85% TTC for each lane at each location.

# 3.3. TTCs' Distribution Prediction with GMM

It is very important to understand the distribution of TTCs in danger determination. GMM is widely used in PDF estimation, which is a parametric PDF represented as a weighted sum of Gaussian component densities. The basic theoretical assumption is that an arbitrary distribution can be approximated by the weighted Gaussian models if there are enough Gaussian models. In this study, GMM was applied to explore the TTC distribution and capture the features for risk assessment without any distribution assumption. The PDF of the completed GMM is the sum of the sub-PDF described by Equation (3).

$$p\{TTC_{k}|\omega_{i},\mu_{i},\sigma_{i}^{2}\} = \sum_{i=1}^{N} \omega_{i}p(TTC_{k}|\mu_{i},\sigma_{i}^{2}), \sum_{k=1}^{N} \omega_{i} = 1, \ 0 \le \omega_{i} \le 1, \forall i = 1, \cdots, N,$$
(3)

where the parameter  $TTC_k$  is the TTC of the  $k^{\text{th}}$  vehicle, which can be calculated by choosing a mixture weight  $\omega_i$  and  $p(TTC_k | \mu_i, \sigma_i^2)$ .  $\omega_i$  is between zero and one, which represents the percentage of the TTC belonging to category *i*. The total sum of  $\omega_i$  is 1.  $\mu_i$  is the mean vector.  $\sigma_i^2$  is the variance vector. GMM was used to estimate the PDF of TTC samples, and the estimated model was the sum of several Gaussian components with different probabilities  $\omega_i$ .

Each component density function is as Equation (4).

$$p(TTC_k|\mu_i, \sigma_i^2) = \frac{1}{\sqrt{2\pi}} \exp\left\{-\frac{1}{2\sigma^2} (TTC_k - \mu_i)^2\right\},$$
(4)

The complete GMM was parameterized by means, variances, and mixture weights from all components of Gaussian densities. Each sub-model is represented as  $\lambda = (\omega_i, \mu_i, \sigma_i)$ . The number of sub-Gaussian models should be consistent with the crash risk levels.

The popular and well-known method for estimating the parameters of GMM is the maximum likelihood estimation (MLE). To train samples  $TTC_i$ , the MLE function of GMM can be written as Equation (5).

$$p(TTC|^{\Theta}) = \prod_{i=1}^{N} p(TTC_i|^{\Theta}) = L(^{\Theta}|TTC),$$
(5)

The function  $L(\Theta|TTC)$  is called the likelihood of parameters given the training data. The likelihood is a function of  $\Theta$  where the TTC value is fixed.

The expectation-maximization (E-M) algorithm iterates through two steps to obtain the estimation of parameters. The E-M algorithm is a general method of finding the maximum-likelihood estimation of parameters when the given data are incomplete or have missing values. The goal of the algorithm is to find the parameters that make  $L(\Theta|TTC)$  the largest. The two steps, the E-step and the M-step, are repeated until the maximum change in the estimation reaches convergence.

The condition of convergence is as Equation (6).

$$\sum_{i=1}^{K} \| \mu^{t+1} - \mu^t \| < \varepsilon,$$
(6)

where  $\varepsilon$  is a random constant.

In this study, the crash risks were divided into three levels: high, medium, and low. Therefore, three Gaussian sub-models were used to fit the TTCs' distribution. The iteration stop condition was  $1 \times 10^{-15}$ . The confidence level of estimation was 95%.

The modeling process is shown in Figure 9.

GMM was used to estimate the TTC distribution for five sites. The blue line represents the high risk level. The green line and red line represent the medium and low risk level.

The PDFs and CDFs for five sites are illustrated in Figures 10 and 11.



**Figure 10.** The three PDFs of TTC at the five sites: (a) Site No.1; (b) Site No.2; (c) Site No.3; (d) Site No.4; (e) Site No.5.



**Figure 11.** The three CDFs of TTC at the five sites: (**a**) Site No.1; (**b**) Site No.2; (**c**) Site No.3; (**d**) Site No.4; (**e**) Site No.5.

The PDF indicated the probability of crash risk at a certain TTC value. The intersections of crossing curves in Figure 10 show the transition from high to medium risk. The place with a smaller TTC than the intersection had a greater probability at a high crash risk level, otherwise it had a greater probability at the medium crash risk level. Therefore, the TTC of the transition point on the PDF curves was considered as the threshold to distinguish high crash risk from medium crash risk.

The K-S test is a nonparametric test of the equality of continuous, one-dimensional probability distributions that can be used to compare a sample with a reference probability distribution. In statistics, the K-S test is applied to compare the TTC samples with the proposed GMM probability distribution. The comparison was set at the 0.05 confidence level. The K-S sig. value shows that the null hypotheses is accepted, and the samples were drawn from the reference distribution as given in Table 3.

Site	ω	μ	$\sigma_i^2$	р	K-S Sig.
1	(0.36 0.31 0.33)	(2.71 5.02 16.95)	(0.78 9.41 65.58)	0.00	0.18
2	(0.26 0.33 0.31)	(4.52 11.36 38.42)	(0.87 21.4 33.99)	0.00	0.17
3	(0.34 0.34 0.32)	(4.34 11.42 33.76)	(1.92 4.45 21.79)	0.00	0.32
4	(0.26 0.29 0.35)	(4.09 9.13 32.10)	(1.67 3.96 21.39)	0.00	0.41
5	(0.36 0.29 0.35)	(5.60 8.48 26.84)	(1.89 3.70 21.70)	0.00	0.36
Total	(0.36 0.347 0.293)	(4.32 9.43 32.35)	(2.9 11.48 460.72)	0.00	0.57

Table 3. Estimation results of the sub-probability density function for GMM.

Table 4 is the theoretical estimation with GMM at a given percentage, and the severe crash thresholds are listed in the last column.

Site	15% TTC (s)	30% TTC (s)	50% TTC (s)	85% TTC (s)	Severe Crash Risk Thresholds (s)
1	2.34	4.89	7.57	12.79	2.23
2	8.92	13.00	17.79	26.95	6.71
3	6.48	11.029	15.64	24.75	2.61
4	4.33	8.3	12.37	20.49	2.29
5	5.59	9.8	14.13	22.66	2.60

 Table 4. The TTC distribution estimation.

The theoretical values of 15% TTC, 30% TTC, 50% TTC, and 85% TTC for the five sites were calculated. The severe crash risk thresholds obtained from the PDF curves for Sites 1 to 5 were 2.23 s, 6.71 s, 2.61 s, 2.29 s, and 2.60 s, respectively.

The 15% TTC, 50% TTC, and 85% TTC for Site 1 and Site 4 were smaller than the other sites. Comparing the crash risk thresholds with the 15% TTC, it was found that the severe crash risk thresholds obtained from the distribution density functions were smaller than the 15% TTC of the collected samples.

The small TTC indicated that the driver had a short time to take measures to avoid collisions. Therefore, crashes were likely to occur at Site 1 and Site 4 after TTC, and Site 2 was safer than other sites. The thresholds could also be used to warn about the collisions.

## 4. Crash Risk Modeling

The paper proposed an ordinal logical regression model and a naive Bayesian model with four variables including speed, speed S.D., traffic volume, and truck percentage.

TTC has been proven to be an effective indicator for rating the severity of crash risks. The crash risks were divided into three levels: high, medium, and low, according to the TTC thresholds in Table 4. The critical value for crash risk classification was determined by the derived PDFs. Combining with the results of previous research [40], the critical value for high risk was 2.7 s. When the TTC was between 0 and 2.7 s, the crash risk of off-ramps was high. The critical value for the medium risk was 4.7 s. When the TTC was between 2.7 and 4.7 s, the crash risk of off-ramps was medium; when the TTC was greater than 4.7 s, the crash risk of off-ramps was low.

Since not all vehicles on the mainline were required to enter the ramp, 158 samples were selected from 1552 trajectory samples to develop the models, including 69 samples with a low crash risk,

44 samples with a medium risk, and 45 samples with a high risk. The following two conditions were considered when screening samples. The first one was that the vehicle was about to run out of ramp, and the second condition was that a collision had occurred, which was represented by a small TTC.

In order to test the validity of the model, 120 samples were randomly selected as the training dataset, and all samples were selected as the validation dataset. Meanwhile, the training datasets included 51 samples with a low crash risk, 27 samples with a medium risk, and 42 samples with a high risk. Before establishing the model, the relationship between the crash risks of off-ramps and four independent variables was analyzed separately, and the relationship between the crash risks of off-ramps and speed, traffic volume, speed S.D., and truck percentage was obtained, as shown in Figures 12–15, respectively.



Figure 12. The relationship between crash risks of off-ramps and speed.



Figure 13. The relationship between crash risks of off-ramps and traffic volume.



Figure 14. The relationship between crash risks of off-ramps and speed S.D.



Figure 15. The relationship between crash risks of off-ramps and truck percentage.

It can be seen from Figure 13 that the impact of traffic volume on crash risk was relatively obvious. The lower crash risk corresponded to the smaller traffic volume, and the higher crash risk corresponded to the larger traffic volume. In addition, Figure 12, Figure 14, and Figure 15 illustrate that the impact of speed, speed S.D., and truck percentage on crash risk was not obvious.

## 4.1. The Ordinal Logistic Regression Model

In this study, the ordinal logistic regression model was used to describe the relationship between independent variables and ordinal response variables. It was based on the cumulative probability theory. The ordinal logistic regression model assumes that the dependent variable Y can be divided into ordinal g categories (Y = 1, 2, ..., g).  $X_1, X_2, ..., X_m$  are independent variables. The ordinal logistic regression model is as Equation (7).

$$In\left(\frac{P(Y \le j)}{1 - P(Y \le j)}\right) = \beta_{0j} + \beta_1 X_1 + \beta_2 X_2 + \dots + \beta_m X_m,$$
(7)

where  $\beta_{0j}$  is the constant term of the *j*<sup>th</sup> regression equation.  $\beta_m$  is the regression coefficient of independent variable  $X_m$ .

There are g - 1 equations for the different categories of *Y*. The concept of the ordinal logistic regression model is to assume that the independent variables have the same influence on the odds ratio of cumulative probability. The regression coefficients of each variable in all equations are the same, and the differences in the cumulative probability of the different categories are depicted by the constant terms.

The corresponding probability of event occurrence is  $P_1$ ,  $P_2$ ,  $P_g$  for each category, and  $P_1 + P_2 + P_g = 1$ . Hence, the probability when Y = j is as Equation (8).

$$=\frac{P_{j}=P(Y\leq j)-P(Y\leq j-1)}{\frac{1}{1+exp\left[-\left(\beta_{0j}+\beta_{1}X_{1}+\beta_{2}X_{2}+...+\beta_{m}X_{m}\right)\right]}-\frac{1}{1+exp\left[-\left(\beta_{0j-1}+\beta_{1}X_{1}+\beta_{2}X_{2}+...+\beta_{m}X_{m}\right)\right]}},$$
(8)

The value of *Y* means the crash risk level. Y = 1 indicates a low risk, and Y = 2 indicates a medium risk, while Y = 3 indicates a high risk.

The proportional odds assumption for the independent variables of the ordinal logistic regression is validated by the test of parallel lines. The test of parallel lines tests the proportional odds assumption for explanatory variables. The null hypothesis states that the slope coefficients in the model are the same across response categories.

The test results are shown in Table 5. In this case,  $\chi^2 = 4.351$ , P = 0.361, indicating that the regression equations were parallel to each other. Hence, the analysis could be carried out by the ordinal logistic regression model.

Model	–2 Log Likelihood	Chi-Squared	df	Sig.
Null Hypothesis	115.810			
General	111.459	4.351	4	0.361

Table 5. Test of parallel lines.

The ordinal logistic regression model was designed to explore the crash risk level and explanatory variables. The significance level of the model was set to the 95% confidence interval. The results showed the significance level of traffic volume and speed S.D. at 0.000 and 0.027, which were less than 0.05, indicating that these two independent variables could significantly affect the dependent variable. The significance level of speed and truck percentage at 0.062 and 0.502, which were greater than 0.05, indicated that the speed and truck percentage were less correlated with the dependent variable in this model. The ordinal logistic regression model results are shown in Table 6.

Table 6. The ordinal logistic regression model results.

		Estimate	Std. Error	Wald	df	Sig.	95% Confidence Interval
Risk level	Low risk	4.459	1.185	14.161	1	0.000	(2.137, 6.782)
	Medium risk	7.868	1.415	30.904	1	0.000	(5.094, 10.642)
<b>V</b> /	Traffic volume (per/h)	0.040	0.005	73.666	1	0.000	(0.031, 0.049)
Variables	Speed S.D. (m/s)	-0.218	0.099	4.878	1	0.027	(-0.412, -0.025)

Two variables including traffic volumes and speed S.D. were considered in the estimation model. The ordinal logistic regression showed that the proposed model could well evaluate the relationship between the crash risks of off-ramps and each independent variable. In addition, although the ordinal logistic regression model ignored the impact of speed and truck percentage on the crash risk, its accuracy was as high as 84.81%. The speed and truck percentage had a smaller impact on the crash risk than the other two independent variables, so they were ignored in the model.

## 4.2. The Naive Bayesian Model

The naive Bayesian model is a statistical method. Based on the Bayesian theorem and the independent hypothesis of characteristic conditions, it can predict the possibility between variables. The naive Bayesian model is characterized by high accuracy and high speed. It assumes that there is no interaction between any single attribute and classification result, which makes the calculation easier. In the study of traffic behavior, the Bayesian theorem can be applied to judge the probability of traffic behavior based on traffic flow.

According to the Bayesian rule, the conditional probability P(A|B) represents the possibility of event *A* occurring when *B* occurs. The calculation formula is as Equation (9).

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}$$
(9)

where P(A) is the prior probability when A occurs. P(B) is the prior probability when B occurs. P(A|B) is the conditional probability of A after the occurrence of B, which is called the posterior probability of A. P(B|A) is the conditional probability of B after the occurrence of A, which is also called the posterior probability of probability of B.

In the Bayesian rule, the denominator can be regarded as the normalized coefficient  $\eta$ , and Equation (10) can be obtained.

$$P(x|y) = \frac{P(y|x)P(x)}{P(y)} = \eta P(y|x)P(x), \eta = P(y)^{-1} = \frac{1}{\sum_{x} P(y|x)P(x)},$$
(10)

If there is various observation information in traffic flow detection, it can be calculated by Equation (11).

$$P(x|y,z) = \frac{P(x,y,z)}{P(y,z)} = \frac{P(y|x,z)P(x,z)}{P(y,z)} = \frac{P(y|x,z)P(x|z)P(z)}{P(y|z)P(z)} = \frac{P(y|x,z)P(x|z)}{P(y|z)},$$
(11)

Thus, the derivation formula of Bayesian filtering is as Equation (12).

$$P(x|z_{1}, \dots, z_{n}) = \frac{P(z_{n}|x, z_{1}, \dots, z_{n-1})P(x|z_{1}, \dots, z_{n-1})}{P(z_{n}|z_{1}, \dots, z_{n-1})} = \frac{P(z_{n}|x)P(x|z_{1}, \dots, z_{n-1})}{P(z_{n}|z_{1}, \dots, z_{n-1})} = \eta_{n}P(z_{n}|x)P(x|z_{1}, \dots, z_{n-1}) = \eta_{n}P(z_{n}|x)\eta_{n-1}P(x|z_{1}, \dots, z_{n-2}) = \eta_{1}\cdots\eta_{n} \prod_{i=1\cdots n} P(z_{i}|x)P(x),$$
(12)

The specific analysis of the naive Bayesian model is as follows: Each TTC sample had four corresponding attributions, including speed, speed S.D., traffic volume, and truck percentage, which can be expressed by  $Z = \{z_1, z_2, ..., z_n\}$ . The crash risks of ramp were divided into low risk, medium risk, and high risk, which are represented by  $A_1$ ,  $A_2$ , and  $A_3$ , respectively. The proportion of each group of samples was classified into three levels with four attributes. Finally, the probability of risk level was obtained.

The naive Bayesian model was established with variables such as speed, speed S.D., traffic volume, and truck percentage. The model was trained by 120 training datasets and validated by 158 datasets. The Bayesian naive model classification results are shown in Table 7.

Variables	Low	Risk	Mediu	m Risk	High Risk	
variables	Mean	Std.	Mean	Std.	Mean	Std.
speed (m/s)	11.3105	2.4384	10.0719	2.3245	8.9801	4.2843
speed S.D. (m/s)	6.3195	1.9577	5.4465	1.1900	5.3576	0.8400
traffic volume (per/h)	108	30	177	33	275	49
truck percentage	0.1038	0.0232	0.1109	0.0231	0.1112	0.0308

Table 7. The naive Bayesian model results.

It was found that the naive Bayesian model could well evaluate the relationship between the crash risks of off-ramps and each independent variable. The established model was tested by the validation datasets, and the prediction accuracy was 86.71%, which ensured the efficiency of the naive Bayesian model. Traffic volume was one of the most common exposure variables in the previous analyses, and there was a significant positive relationship between traffic volume and crash risks of off-ramps. It could be seen from the naive Bayesian model that the relationship between crash risk and traffic volume was the most obvious. The average traffic volume corresponding to low crash risk, medium crash risk, and high crash risk was 108, 177, and 275, respectively.

## 5. Discussion

The ordinal logistic regression model and the naive Bayesian model were designed to explore the relationship between the crash risks of off-ramps and the explanatory variables including speed, speed S.D., traffic volume, and truck percentage.

When evaluating the fit of a model, the Akaike information criterion (AIC) and Bayesian information criterion (BIC) are usually used to analyze the model. The calculation formulas are shown in Equations (13) and (14). The AIC and the BIC are proposed to prevent overfitting problems when training the model. As the accuracy of the model continues to increase, the number of unknown parameters in the model will increase, and the model will become more complex, which will lead to

overfitting problems. Therefore, an evaluation for model should consider both the accuracy of fitting and the number of unknown parameters. In general, smaller AIC and BIC show a good fitting.

$$AIC = 2k - 2In(L), \tag{13}$$

$$BIC = kIn(n) - 2In(L), \tag{14}$$

where *k* is the number of model parameters, *n* is the number of samples, and *L* is the likelihood function.

The MLE is used in the training model of ordinal logistic regression. After calculation, the AIC was 119.81, and BIC was 125.9352. The naive Bayesian model did not use the MLE in the training model, so AIC and BIC could not be calculated. Then, the prediction accuracy of the two models was compared and analyzed.

The ordinal logistic regression model considered speed S.D. and traffic volume, while the naive Bayesian model took four independent variables into account. The ordinal logistic regression model and the naive Bayesian model were tested with all 158 data, and the predicted results are shown in Tables 8 and 9.

The Crash Risks of Off-Ramps	Actual Frequency	Accurate Prediction of Frequency	Prediction Accuracy
Low	69	65	94.20%
Medium	44	31	70.45%
High	45	38	84.44%
Total	158	134	84.81%

Table 8. The predicted results of the ordinal logistic regression model.

The Crash Risks of Off-Ramps	Actual Frequency	Accurate Prediction of Frequency	Prediction Accuracy
Low	69	65	94.20%
Medium	44	31	70.45%
High	45	41	91.11%
Total	158	137	86.71%

Table 9. The predicted results of the naive Bayesian model.

It can be observed from Tables 8 and 9 that the prediction accuracy of the two models was relatively high. The prediction accuracy of the ordinal logistic regression model was 84.81%, and the prediction accuracy of the naive Bayesian model was 86.71%. The naive Bayesian outperformed the ordinal logistic regression model.

In addition, both models had relatively low prediction accuracy for medium crash risk, which may be due to the similar sample data of medium crash risk and low crash risk.

An advanced crash warning system at off-ramps was designed based on the crash risk probability to enhance safety. The system integrating the model could predict the real-time crash risk level with real-time traffic flow information from loops including traffic volume, truck percentage, speed, and speed S.D. The system sends the warning signals according to the crash risk level. The framework of the system is as Figure 16.



Figure 16. The framework of the collision risk warning system.

# 6. Conclusions

The paper studied the TTC distribution of an urban expressway. The trajectories from the Xi'an expressway were collected by an unmanned aerial vehicle and a camera. After the analysis, the following conclusions were obtained.

- 1. The traffic flows of all sites were above 780 veh/h during rush hour. The 85% speed was approximately 38–49 km/h. The truck percentage was between 6.25% and 23.8%.
- 2. TTC decreased with increasing traffic volume, and was positive in the diverging rate, traffic volume, and speed deviation.
- 3. The TTC probability distributions of Lane 1 and Lane 2 were similar. The TTC of Lane 3 was smaller than Lane 1 and Lane 2. Due to the interaction of merging vehicles, the safety level of Lane 3 was lower than the other lanes.
- 4. The TTC distribution could be represented by three sub-Gaussian models. The smaller TTC indicated a greater probability of crash. The theoretical thresholds were considered as critical values for crash risk division.
- 5. Two prediction models of crash risks of off-ramps were proposed with four independent variables of speed, speed S.D., traffic volume, and truck percentage. The ordinal logistic regression model considered two independent variables, while the naive Bayesian model considered all independent variables.
- 6. The prediction accuracy of the ordinal logistic regression model was 84.81%, and the prediction accuracy of naive Bayesian model was 86.71%. The AIC and BIC of the ordinal logistic regression model were 119.81 and 125.9352. The naive Bayesian outperformed the ordinal logistic regression model.

**Author Contributions:** Conceptualization, T.X. and S.I.-J.C.; formal analysis, Y.H. (Yanjun Hao) and S.C.; investigation, S.C., X.W. and Z.Z.; methodology, T.X. and S.I.-J.C.; software, Y.H. (Yanjun Hao) and S.C.; validation, T.X., Y.H. (Yulong He) and X.W.; writing, original draft, Y.H. (Yanjun Hao) and Z.Z.; writing, review and editing, T.X., Y.H. (Yanjun Hao) and Y.H. (Yulong He). All authors have read and agreed to the published version of the manuscript.

**Funding:** This research was funded by the National Natural Science Foundation of China under Grant U1664264 and Grant No.51878066, Funds for Central Universities and Colleges of Chang'an University (No.300102229201 and No.300102220204), the Major scientific and technological innovation projects of Shandong Province under Grant No.2019JZZY020904, and Xi'an scientific and technological projects under Grant No.2019218514GXRC021CG022-GXYD21.5.

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

## References

- 1. Chen, S.-K.; Mao, B.-H.; Liu, S.; Sun, Q.-X.; Wei, W.; Zhan, L.-X. Computer-aided analysis and evaluation on ramp spacing along urban expressways. *Transp. Res. Part C-Emerg. Technol.* **2013**, *36*, 381–393. [CrossRef]
- 2. Chen, Z.l. Study on the Traffic Accidents Black-Spots of Urban Expressway. Master's Thesis, Southwest Jiaotong University, Chengdu, China, 2013.
- 3. Wei, D.; Lu, J.; Lu, L.; Shen, C. Optimal Distance between On-off Ramps in Urban Freeway. J. Highw. Transp. Res. Dev. 2013, 30, 109–114.
- 4. Bauer, K.M.; Harwood, D.W. Statistical Models of Accidents on Interchange Ramps and Speed-Change Lanes. *Highw. Des* **1998**, 1–166. Available online: https://rosap.ntl.bts.gov/view/dot/38348 (accessed on 10 April 2020).
- 5. Choi, Y.H.; Park, S.H.; Ko, H.; Kim, K.H.; Yun, I. Development of safety performance functions and crash modification factors for expressway ramps. *Ksce J. Civ. Eng.* **2018**, *22*, 804–812. [CrossRef]
- Easa, S.M.; Mehmood, A. Establishing highway horizontal alignment to maximize design consistency. *Can. J. Civ. Eng.* 2007, 34, 1159–1168. [CrossRef]
- Hu, J.-B.; Ma, W.-Q.; Wang, M. Study on the Length of Speed-Change Lane in Freeway Trumpet Interchange. In *Materials, Transportation and Environmental Engineering, Pts 1 and 2*; Kao, J.C.M., Sung, W.P., Chen, R., Eds.; Trans Tech Publications Ltd: Zurich, Switzerland, 2013; Volume 779, pp. 946–953.
- 8. Zhang, C.; Yan, X.; An, M.; Zhao, H. Spatial Influence Analysis of Traffic Safety in Diverging Areas between Freeway Segments and Off Ramps. *Discret. Dyn. Nat. Soc.* **2015**. [CrossRef]
- Yunlong, T.; Hongfei, J. Vehicle interaction behaviors model based on drivers characteristics at expressway-ramp merging area. In Proceedings of the 2013 6th International Conference on Information Management, Innovation Management and Industrial Engineering, Xi'an, China, 23–24 November 2013; pp. 371–374.
- 10. Suh, J.; Chae, H.; Yi, K. Stochastic Model-Predictive Control for Lane Change Decision of Automated Driving Vehicles. *IEEE Trans. Veh. Technol.* **2018**, *67*, 4771–4782. [CrossRef]
- 11. Calvi, A.; Bella, F.; D'Amico, F. Evaluating the effects of the number of exit lanes on the diverging driver performance. *J. Transp. Saf. Secur.* **2018**, *10*, 105–123. [CrossRef]
- 12. Hongyun, C.; Huaguo, Z.; Jiguang, Z.; Peter, H. Safety performance evaluation of left-side off-ramps at freeway diverge areas. *Accid. Anal. Prev.* **2011**, *43*, 605–612.
- 13. Kim, S.; Song, T.-J.; Rouphail, N.M.; Aghdashi, S.; Amaro, A.; Goncalves, G. Exploring the association of rear-end crash propensity and micro-scale driver behavior. *Saf. Sci.* **2016**, *89*, 45–54. [CrossRef]
- 14. Ma, C.; Hao, W.; Xiang, W.; Yan, W. The Impact of Aggressive Driving Behavior on Driver-Injury Severity at Highway-Rail Grade Crossings Accidents. *J. Adv. Transp.* **2018**. [CrossRef]
- 15. Chen, H.; Liu, P.; Lu, J.J.; Behzadi, B. Evaluating the safety impacts of the number and arrangement of lanes on freeway exit ramps. *Accid. Anal. Prev.* **2009**, *41*, 543–551. [CrossRef] [PubMed]
- 16. Reddy, G.S.; Dominique, L.; Sekhar, D.S. The negative binomial-Lindley generalized linear model: Characteristics and application using crash data. *Accid. Anal. Prev.* **2012**, *45*, 258–265.
- 17. Marina Gonzalez, R.; Marrero, G.A. Induced road traffic in Spanish regions: A dynamic panel data model. *Transp. Res. Part A-Policy Pract.* **2012**, *46*, 435–445. [CrossRef]
- 18. De Oña, J.; Mujalli, R.O.; Calvo, F.J. Analysis of traffic accident injury severity on Spanish rural highways using Bayesian networks. *Accid. Anal. Prev.* **2011**, *43*, 402–411. [CrossRef]
- 19. Fei, H.; Pan, L.; Hao, Y.; Wei, W. Identifying if VISSIM simulation model and SSAM provide reasonable estimates for field measured traffic conflicts at signalized intersections. *Accid. Anal. Prev.* **2013**, *50*, 1014–1024.
- 20. Liu, K.; Jia, J.; Zuo, Z.; Ando, R. Heterogeneity in the effectiveness of cooperative crossing collision prevention systems. *Transp. Res. Part C-Emerg. Technol.* **2018**, *87*, 1–10. [CrossRef]
- 21. Xu, C.; Qu, Z. Empirical Analysis on Time to Collision at Urban Expressway. In *Advances in Transportation*, *Pts 1 and 2*; Bao, T., Ed.; Trans Tech Publications Ltd: Zurich, Switzerland, 2014; Volume 505, pp. 1127–1132.
- 22. Farah, H.; Bekhor, S.; Polus, A. Risk Evaluation by Modelling of Passing Behaviour on Two-Lane Rural Highways. *Accid. Anal. Prev.* **2009**, *41*, 887–894. [CrossRef]
- 23. Babu, S.S.; Vedagiri, P. Proactive safety evaluation of a multilane unsignalized intersection using surrogate measures. *Transp. Lett. Int. J. Transp. Res.* **2018**, *10*, 104–112. [CrossRef]

- 24. Guido, G.; Saccomanno, F.F.; Vitale, A.; Gallelli, V.; Rogano, D. A calibration framework of car following models for safety analysis based on vehicle tracking data from smartphone probes. *Int. J. Mob. Netw. Des. Innov.* **2014**, *5*, 205–212. [CrossRef]
- 25. Yan, K.; Xiaobo, Q.; Shuaian, W. A tree-structured crash surrogate measure for freeways. *Accid. Anal. Prev.* **2015**, *77*, 137–148.
- Laureshyn, A.; Ceunynck, T.D.; Karlsson, C.; Svensson, Å.; Daniels, S. In search of the severity dimension of traffic events: Extended Delta-V as a traffic conflict indicator. *Accid. Anal. Prev.* 2017, *98*, 46–56. [CrossRef] [PubMed]
- 27. Tsui, M.A.; Alfredo, G. Use of speed profile as surrogate measure: Effect of traffic calming devices on crosstown road safety performance. *Accid. Anal. Prev.* **2013**, *61*, 23–32.
- 28. Vogel, K. A comparison of headway and time to collision as safety indicators. *Accid. Anal. Prev.* **2003**, *35*, 427–433. [CrossRef]
- 29. Chengcheng, X.; Wei, W.; Pan, L. Identifying crash-prone traffic conditions under different weather on freeways. *J. Saf. Res.* **2013**, *46*, 135–144.
- 30. Chengcheng, X.; Pan, L.; Wei, W.; Zhibin, L. Evaluation of the impacts of traffic states on crash risks on freeways. *Accid. Anal. Prev.* **2012**, *47*, 162–171.
- 31. Moon, J.-P.; Hummer, J.E. Development of Safety Prediction Models for Influence Areas of Ramps in Freeways. *J. Transp. Saf. Secur.* **2009**, *1*, 1–17. [CrossRef]
- Abdel-Aty, M.; Uddin, N.; Pande, A.; Abdalla, M.F.; Hsia, L. Predicting freeway crashes from loop detector data by matched case-control logistic regression. In *Statistical Methods and Safety Data Analysis and Evaluation*; Transportation Reserach Board Natl Research Council: Washington, WA, USA, 2004; pp. 88–95.
- 33. Sun, J.; Sun, J. A dynamic Bayesian network model for real-time crash prediction using traffic speed conditions data. *Transp. Res. Part C-Emerg. Technol.* **2015**, *54*, 176–186. [CrossRef]
- 34. Moinul, H.; Yasunori, M. A Bayesian network based framework for real-time crash prediction on the basic freeway segments of urban expressways. *Accid. Anal. Prev.* **2012**, *45*, 373–381.
- 35. Abdelwahab, H.T.; Abdel-Aty, M.A. Artificial neural networks and logit models for traffic safety analysis of toll plazas. In *Statistical Methodology: Applications to Design, Data Analysis, and Evaluation: Safety and Human Performance;* Transportation Reserach Board Natl Research Council: Washington, WA, USA, 2002; pp. 115–125.
- 36. Lee, D.; Yeo, H. Real-Time Rear-End Collision-Warning System Using a Multilayer Perceptron Neural Network. *IEEE Trans. Intell. Transp. Syst.* **2016**, *17*, 3087–3097. [CrossRef]
- 37. Vangala, P.; Lord, D.; Geedipally, S.R. Exploring the application of the Negative Binomial-Generalized Exponential model for analyzing traffic crash data with excess zeros. *Anal. Methods Accid. Res.* **2015**, *7*, 29–36. [CrossRef]
- 38. Tageldin, A.; Zaki, M.H.; Sayed, T. Examining pedestrian evasive actions as a potential indicator for traffic conflicts. *Iet Intell. Transp. Syst.* **2017**, *11*, 282–289. [CrossRef]
- Ponte, G.; Ryan, G.A.; Anderson, R.W.G. An estimate of the effectiveness of an in-vehicle automatic collision notification system in reducing road crash fatalities in South Australia. *Traffic Inj. Prev.* 2016, 17, 258–263. [CrossRef] [PubMed]
- 40. Zhao, S.P. Interchange Traffic Safety Evaluation Model Based on Traffic Conflict Technology. Master's Thesis, School of Transportation Southeast University, Nanjing, China, 2016.



© 2020 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (http://creativecommons.org/licenses/by/4.0/).