Comprehensive survival analytics for lane-changing duration

Yang Li
Ph.D., Candidate
Key Laboratory of Road and Traffic Engineering of Ministry of Education,
Tongji University, China,
4800 Cao'an Road, Shanghai, 201804,
Email: cc960719@tongji.edu.cn

Linbo Li*(Corresponding Author)
Ph.D., Associate Professor
Key Laboratory of Road and Traffic Engineering of Ministry of Education,
Tongji University, China,
4800 Cao'an Road, Shanghai, 201804
Email: llinbo@tongji.edu.cn

Daiheng Ni
Ph.D., Professor
Civil and Environmental Engineering,
University of Massachusetts Amherst, Massachusetts 01003, USA
Email: ni@engin.umass.edu

Word Count: 6426 words + 4 table (250 words per table) = 7,426 words

Submitted [July 31, 2020]

ABSTRACT

Lane-changing duration (LCD) is an important indicator of lane-changing (LC) maneuver in favor of understanding the LC behavior and comprehending the mechanism of traffic phenomena. However, the conventional research efforts mainly focus on modeling the LC decision-making process and few efforts have been made on investigating the LCD. Therefore, this paper aims to present a complete univariant and regression analysis of LCD using the hazard-based duration model. Firstly, in order to investigate the difference in LCD under different LC behaviors, the Gaussian Mixture Model (GMM) is employed to classify the LC trajectories. Then the univariant non-parametric and parametric estimator are introduced to investigate this difference, which is set as the stratified variable for the subsequent regression model. Subsequently, in order to model the LCD as a continuous event, the time-dependent proportional hazard-based model is employed, which overcomes the shortcoming of modeling the LCD as an instantaneous event in the existing studies. Conclusions drawn by this research may help us better understand LC behaviors and may provide some useful insights to the (ADAS) Advanced Driving Assistance System.

Keywords: Lane-changing behavior, Lane-changing duration, Gaussian Mixture Model, Hazard-based duration model, Advanced driving assistance system

INTRODUCTION

Lane changing (LC) behavior describes the lateral movement of the vehicles from one lane to another lane while proceeding forward. Along with car following (CF) behavior, LC behavior is also an indispensable component of traffic flow theories (*1; 2*). However, the modeling efforts of LC behavior appear less than those of CF behavior, which only considers single interaction with the front vehicle. In contrast, LC modeling considers more factors such as LC purpose, vehicle types, impact on the surrounding vehicles (*1*). On the other hand, numerous research efforts indicate that LC behavior has significant impacts on traffic flow(*1; 3; 4*). Therefore, it is imperative for us to comprehensively understand the LC behavior.

Over the past decades, a considerable amount of works have been made to model LC behaviors, but the research efforts mainly focus on modeling LC decision-making process and the LC's impact on surrounding vehicles(*1*). There is a paucity of research on studying the LC action itself(*5*), especially in the area of modeling the lane-changing duration(LCD). LCD measures the gap time for the entire LC execution, which is one of the commonly used indicators to describe lane-changing behavior(*4; 6*).

The research theme for LCD is roughly divided into two parts: One is researching the distribution of LCD, and numerous studies demonstrated that LCD roughly ranges from 1s to 16s(*4-12*). The other research part is to explore the influencing factors of LCD(*4-7; 12; 13*). It is common knowledge that the execution of LC at every moment during the process of LC may be affected by the interactions with the surrounding vehicles or driver's characteristics. The total effects during the process of LC will inevitably lead to different lengths of LCD. Therefore, through analyzing and modeling the influencing factors of LCD, it is possible for us to better understand the mechanism of LC behavior (For example, what are the factors that may affect the execution of LC? What are the main influencing factors of the execution of LC?). Thus, the research on LCD is of great importance for us to understand the LC behavior. Without a good understanding of LCD, we may be unable to have a complete understanding of LC behavior, and thus fail to understand the essence and characteristics of traffic phenomena(*2; 5*). At the same time, as autonomous driving technology (*14; 15*) matures, it requires us to have a complete and deep understanding of LC behavior, including researching the LCD.

In this research regarded, many studies demonstrated that various factors may directly or indirectly influence LCD. These factors may include: traffic density(*5; 11*), vehicles types(*4; 5; 10; 13*), driver characteristics(*4; 6; 12*), the direction of LC(*5; 11; 16*), different time periods(*13*), road types (*17*), the interactions with surrounding vehicles(*4-7; 12*).

The methodology of modeling the influencing factors of LCD can be mainly summarized into two categories. One is using multiple linear regression model, assuming the LCD as the dependent variable and other variables as the independent variables. In this research direction, Toledo and Zohar (*5)* made the first attempt to model LCD. The dataset is the well-known NGSIM dataset, which contains a set of trajectory data at a fine time resolution. The results indicate that traffic density, by the direction of the change, and by other vehicles around the subject vehicle may influence the LCD. Using the same modeling method, Cao, Young and Sarvi (*7)* analyzed the LCD using the data collected from the video camera mounted on a high building adjacent to the road. The result reveals that the traffic conflict during a lane change event also influences LCD. Meanwhile, Wang, Li and Li (*11)* found that there is no significance difference between the left-to-right LCD, which is not consistent with the result in Toledo and Zohar (*5)*. They also conjectured that the duration times will reach a saturation value when the velocity becomes even higher. Recently, Yang, Wang and Quddus (*4)* developed a three-level mixed-effects linear regression model to explore the variables affecting lane change duration. The research results are consistent with results of Toledo and Zohar (*5)*.

Another method is using the proportional(cox) hazard-based duration model. Wu et al. (*13)* introduced the semi-parametric proportional hazard-based model to analyze the mandatory LCD using the data collected from an unmanned aerial vehicle in a freeway maintenance construction area. Results indicate that there is no significant evidence showing that different vehicle types have an effect on LCD, but there is a significant difference in LCD during different time periods. After that, Vlahogianni (*12)* analyzed the overtaking maneuver in two-lane highways using the parametric hazard-based model for the first time. The Weibull, Log-logistic and Log-normal distributions are introduced to model the hazard function. Results

demonstrate that speed difference relative to the lead vehicle, the speed of opposing traffic, the spacing from the lead and opposing traffic, and the driver's gender influence the LCD. However, this result is not consistent with the result in Olsen et al. (*9)*. Olsen et al. (*9)* found that there are no significant differences in LCD among the independent variable of gender, vehicle type and driver type. Recently, Ali et al. (*6)* introduced the parametric accelerated failure time hazard-based model to analyze the mandatory LCD during connected vehicle environment using the data collected in the CARRS-Q Advanced Driving Simulator. The model identified three classes of variables that affects the gap time for LC, including driving conditions, operational variables and driver demographic factors.

It is worth noting that the above two modeling methods can be considered as regression models. The major difference is that the hazard-based duration model has two dependent variables, one representing the state (whether complete LC or not) and the other representing the time (LCD), while there is only one dependent variable in the multiple linear regression model.

Although the existing literatures has achieved certain progress in modeling LCD, efforts are needed to bridge the following research gaps: One is whether LCD varies differently under different LC behaviors. The LC behaviors may vary differently due to various LC motivations, which may cause LCD to vary accordingly. Another significant research gap is that the existing studies model the LCD only as an instantaneous event(*4-7; 12; 13*). However, the modelling of LCD should be an entire-process modelling, which means that the impact of various factors on LCD should be modeled at every timestep during the LC. Actually, these influencing factors are time-dependent and time-varying rather than time-invariant.

Therefore, the objective of this paper is to address the existing research gaps and present a complete analysis of LCD in order to have a deep understanding of LC behavior. The motivations and contributions of this study are concluded as follows:

(1) Investigating the difference in LCD between different LC behaviors, the LC trajectories are classified into different groups using GMM (Gaussian Mixture Model) at first. Then, the univariate non-parametric survival estimator and parametric survival estimator are employed to investigate the difference in LCD between different LC behaviors.

(2) Modelling the continuous impacts of time-dependent factors on LCD, the time-dependent proportional hazard-based model is introduced for the first time in the literature for its capable of modeling the continuous impacts on the entire time axis (the entire LC process). The previous models can only model the impact at a certain moment and the conclusions may not convincing enough.

(3) Presenting a comprehensive survival analysis of LCD including the univariant analysis and regression analysis of LCD, which help us understand the LC behavior more deeply. The findings are interesting, and may be helpful and applicable to the ADAS (Advanced Driving Assistance System).

The remainder of this paper is organized as follows: The proposed survival analysis framework of LCD is given in Section 2. Section 3 presents the data description and LC behaviors classification. Section 4 presents the univariant analysis of LCD and Section 5 presents the regression analysis of LCD. Section 6 presents the conclusion and future work.

THE PROPOSED SURVIVAL ANALYSIS FRAMEWORK OF LCD

The Structure of The Framework

To further investigate the difference in LCD under different LC behaviors, the GMM (Gaussian Mixture Model) is employed to classify LC behaviors. Then, a complete survival analysis of LCD is presented including the univariant analysis and regression analysis. In the univariant analysis of LCD, the non-parametric and full-parametric are employed to model the LCD. In the regression analysis of LCD, the time-dependent proportional hazard-based model is introduced to model the continuous impacts of time-dependent factors on LCD. The proposed framework of modeling influencing factors of LCD is given in **Figure 1.**

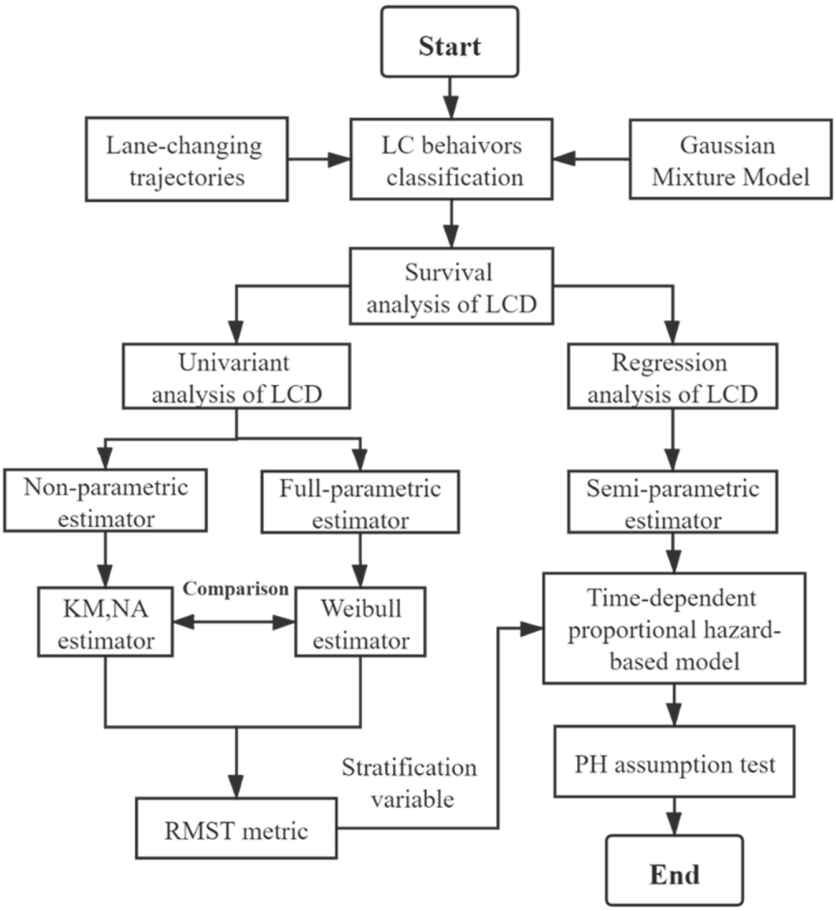

Figure 1 The proposed survival analysis framework of LCD

Gaussian Mixture Model

GMM (Gaussian Mixture Model) has been extensively used in the existing literatures (*18*). GMM clustering method is similar to K-means clustering method. K-means method assign each data point to a certain category, while GMM gives the probability that each point is assigned to each category, which can be regarded as soft assignment. GMM assumes the data is generated from multiple Gaussian distributions. Through increasing the number of Gaussian distributions, GMM is capable of approximating any continuous probability distribution. The formula of the GMM is given below:

$$p(x) = \sum_{i=1}^{M} \alpha_i p_i(x) = \sum_{i=1}^{M} \alpha_i \mathrm{N}(x; \mu_i, \Sigma_i) \qquad (1)$$

Where M is the number of mixtures, $\mathrm{N}(x; \mu_i, \Sigma_i)$ is the normal function with the mean value μ_i and the covariance matrix Σ_i . α_i is the mixture weight with the constraint $\sum \alpha_i = 1$.

Introduction of survival analysis

In the field of survival analysis, the hazard-based duration model can be divided into: non-parametric model, semi-parametric model or full-parametric model (*19*). The choice among the above three models should depend on how much information we know about the underlying distribution of the observation data. If we know little about the statistical distributions of survival or duration data, applying

the non-parametric model may be appropriate. Full-parametric model assumes the specific distribution of duration times (for example Weibull, Exponential, Log-logistic and so on) (*12; 19*), which has been widely used in the existing literature. Meanwhile, the most famous semi-parametric model is the proportional hazard-based model, which has been widely used in the existing literatures(*13; 20*).

On the other hand, the hazard-based duration model can also be divided into univariant model and regression model. Since this paper aims to present a complete survival analysis of LCD, the univariant analysis and regression analysis are fully considered in this paper. Therefore, this study begins with the univariant analysis of LCD after we classify the LC behaviors. In this part, both non-parametric and parametric estimators are introduced to analyze the LCD. Then, the time-dependent proportional hazard-based model is employed to model the continuous impacts of time-dependent factors on LCD.

LANE-CHANGING BEHAVIOR CLASSIFICATION

Definition of terms

A typical LC schematic involves four vehicles: the subject vehicle, the front vehicle on the current lane, the target front vehicle on the target lane, and the target rear vehicle on the target lane. Since this paper focuses on researching the LCD of the subject vehicle, we add the speed difference variables and distance variables for the subsequent research. The description of the LC scenario and the variables terms are shown in **Figure 2** and **Table 1**.

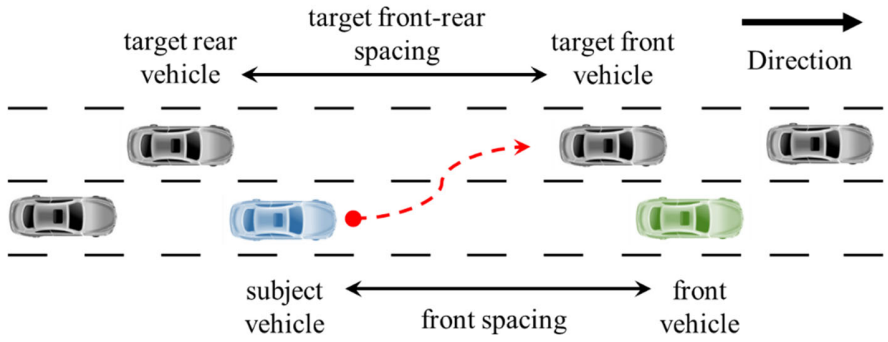

Figure 2 Description of the lane-changing scenario

Table 1 Variables describing the interaction with surrounding vehicles during the LC

Variable	Definition
T_lcd	Lane-changing duration time
$Acc\_s(t)$	The acceleration of subject vehicle at time t
$Spe\_s(t)$	The speed of subject vehicle at time t
$Acc\_f(t)$	The acceleration of front vehicle at time t
$Spe\_f(t)$	The speed of front vehicle at time t
$Acc\_tf(t)$	The acceleration of target front vehicle at time t
$Spe\_tf(t)$	The speed of target front vehicle at time t
$Acc\_tr(t)$	The acceleration of target rear vehicle at time t
$Spe\_tr(t)$	The speed of target rear vehicle at time t
$Dis\_f(t)$	Distance between subject vehicle and front vehicle at time t
$Dis\_tf(t)$	Distance between subject vehicle and target front vehicle at time t
$Dis\_tr(t)$	Distance between subject vehicle and target rear vehicle at time t

$Dis\_trf(t)$	Distance between target front vehicle and target rear vehicle at time t
$Spediff\_f(t)$	Speed difference between subject vehicle and front vehicle at time t
$Spediff\_tf(t)$	Speed difference between subject vehicle and target front vehicle at time t
$Spediff\_tr(t)$	Speed difference between subject vehicle and target rear vehicle at time t

Data processing and visualization

The data adopted in this study are collected from the Next Generation Simulation Program (NGSIM) supported by the Federal Highway Administration (FHWA) (*21*). This well-known project is designed to develop a new generation of microscopic traffic simulation system. This database provides a series of high-fidelity individual microscopic trajectories, and has been studied in a variety of ways over the past decade(*5; 11; 22; 23*). The dataset employed in this study is collected from I-80 and the US101 highways. The I-80 dataset is composed of three 15 minutes periods: 4:00 p.m. to 4:15 p.m., 5:00 p.m. to 5:15 p.m., and 5:15 p.m. to 5:30 p.m. The US101 highway also contains the 45 minutes periods from 7:50 a.m. to 8:35 a.m. The interval of the both datasets is 0.1 seconds (*21*).

The procedure of processing the data is given as below: (1) Due to the inevitable systematic and discretization errors in the dataset, it's necessary for us to perform noise reduction in the above dataset. Therefore, the UKF (Unscented Kalman Filter)(*24*) method is introduced to eliminate the abnormal fluctuations in the dataset before we extract LC trajectories. (2) Filtering out the data of motorcycles and heavy cars, and the data of multiple LC maneuver since we only focus on researching the LCD of private cars and single lane change maneuver. (3) Filtering out all the trajectories when there is no front vehicle or no target front vehicle or no target rear vehicle, because this study mainly focuses on researching the impacts of the interactions with surrounding vehicles on the LCD.

Finally, we extract 561 LC trajectories from the I-80 and the US101 highways dataset, and it is sufficient for us to analyze the LCD. **Figure 3** presents the denoising results, the distribution of LCD and the relative position of LC trajectories. It can be found that the mean value of LCD is around $6.6s$ with the minimum value of $1.8s$ and with the maximum value of $13.2s$.

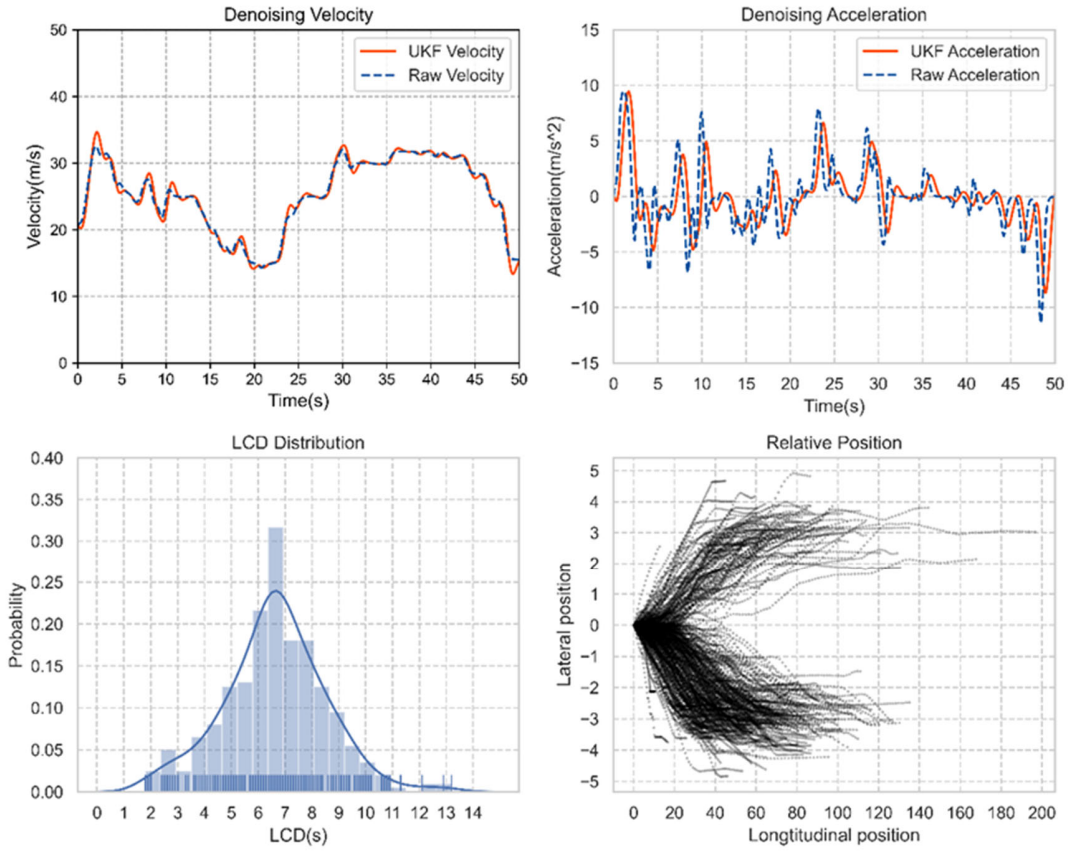

Figure 3 The denoising results, the distribution of LCD and the relative position of LC trajectories

Classification results

In this subsection, we present the detailed process of LC behaviors classification. Before we classify the various LC behaviors, the *Spearman Coefficient* is calculated to know whether there is a monotonic relationship between each two variables. The heatmap of *Spearman Coefficient* are given in **Figure 4.** In **Figure 4**, there is a strong correlation between the speed of the four vehicles, and their *Spearman Coefficient* values all exceed 0.8, which exhibit a strong correlation with each other. Therefore, in the subsequent studies, we eliminate the speed variables and retain the speed difference variables. Since the acceleration can also be obtained from vehicle speed, we exclude the acceleration variables. Considering the target front-rear spacing and the front spacing in the current lane may better reflect the driver's lane change behavior, so we retain these two variables. On the other hand, since the main objective of this section is to classify the LC behaviors to study the difference in LCD, we eliminate the T_lcd variable. Finally, there are five input variables considered in this paper for the classification of LC behaviors. These variables are: Dis_f, Dis_trf, $Spediff\_f$, $Spediff\_tf$, $Spediff\_tr$ respectively.

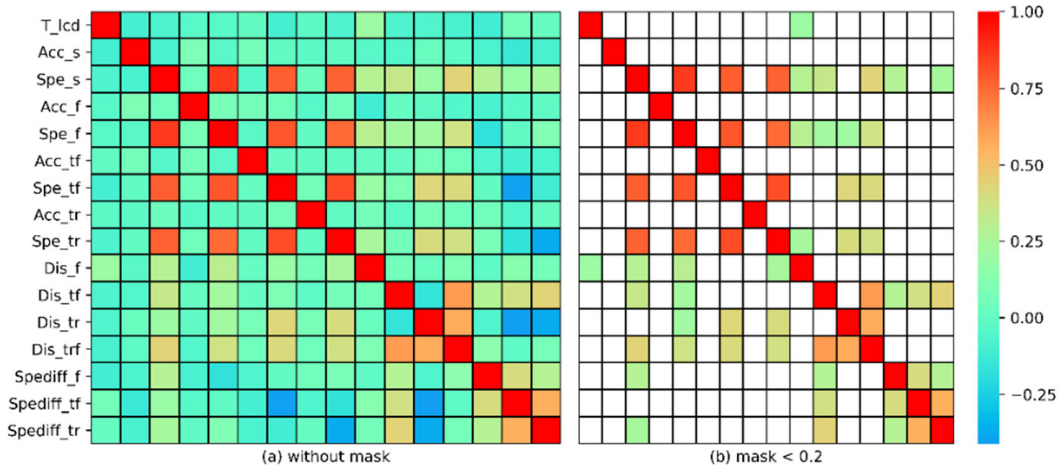

Figure 4 The heatmap of the Spearman Coefficient of the variables

Through observing the value of *Akaike Information Criterion* (AIC) and *Bayesian Information Criterion* (BIC) under different number of Gaussian components, the optimal number of clusters can be obtained. The AIC and BIC are both metrics of assessing model fit penalized for the number of estimated parameters. From the above equations, it can be found that the BIC penalizes model more for free parameters, and the AIC prefers a more complex over a simpler model. AIC presents the danger that it might overfit and BIC presents the danger that it might underfit. Therefore, it is recommended to use both AIC and BIC when choosing the best parameters.

The AIC and BIC values under different number of Gaussian models are given in **Figure 5**. It can be found that with the number of clusters gradually increases, the AIC value gradually decreases, which indicates that the model becomes more complicated. On the other hand, the BIC value decreases at first and then gradually increases, with the lowest value when the number of components is equal to 3. Therefore, the number of Gaussian components is determined as 3, which represents three different LC behaviors in this dataset. Then, the statistics summary of each classification is given in **Table 2** .

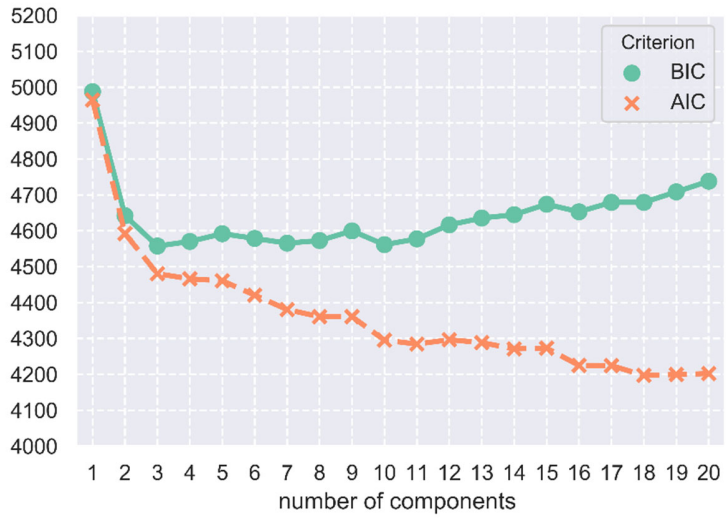

Figure 5 The AIC and BIC values under different numbers of Gaussian components

(1) Through analyzing the clustering results in **Table 2**, it is interesting to find that the speed of the subject vehicle is lower than the three surrounding vehicles in Group 1, and the speed of the subject vehicle is higher than the three surrounding vehicles in Group 2. Meanwhile, the value of variable $Dis\_trf(0)$ and $Dis\_f(0)$ are both very close to each other in Group 1 and Group 2. Therefore, it can be considered that Group1 represents the situation where vehicles move from the slow lane to a fast lane, and Group 2 represents the situation where vehicles move from the fast lane to a slow lane.

(2) Compared with Group 1 and Group 2, the speed of the subject vehicle is close to the surrounding vehicles in Group 3. Meanwhile, it is interesting to find that the distance of the front spacing and the target front-rear spacing are much higher than that in Group 1 and Group 2. Therefore, the vehicles in Group 3 have better comfortable LC conditions than other groups.

Table 2 The clustering results of different lane-changing behaviors

variable	group	min	max	25%	50%	75%	mean	std
$Dis\_f(0)$	1	7.68	37.3	14.65	18.13	22.63	19.09	6.01
	2	9.14	34.97	14.52	18.08	23.05	19.06	5.85
	3	9.3	76.06	20.66	29.77	43.26	32.33	15.62
$Dis\_trf(0)$	1	12.03	74.03	24.51	31.82	42.22	34.84	13.21
	2	14.05	75.74	26.43	32.55	43.77	35.36	12.16
	3	14.48	163.64	38.65	63.33	86.84	66.72	32.85
$Spediff\_f(0)$	1	-1.82	4.12	0.2	0.96	1.72	0.96	1.16
	2	-2.31	2.44	-1	-0.17	0.31	-0.27	1
	3	-6.37	8.18	-1.85	0.4	1.85	0.2	3.23
$Spediff\_tf(0)$	1	-2.94	3.66	-0.93	-0.01	1.03	0.11	1.34
	2	-6.1	0.77	-3.05	-2.07	-1.49	-2.31	1.29
	3	-11.54	6.88	-2.31	-0.9	1.76	-0.29	3.8
$Spediff\_tr(0)$	1	-3.84	5.37	0.39	1.23	2.28	1.28	1.55
	2	-4.73	1.16	-2.03	-1.18	-0.39	-1.27	1.2
	3	-13.02	10.76	-1.25	0.44	2.98	0.49	3.98

MODELING THE LCD USING UNIVARIANT SURVIVAL MODEL

Univariant survival model

In this subsection, we present a brief introduction of the KM (Kaplan–Meier) estimator, NA (Nelson Aalen) estimator and Weibull estimator. KM and NA estimator are the most used non-parametric univariant method to estimate the survival and hazard function of the data. As for the parametric model, we adopted the Weibull distribution model, which has often been used in the literature. This is because the Weibull distribution is a more generalized form of exponential, which is capable of allowing for positive, negative, or even no duration dependence(*19*).

$$\hat{S}(t) = \prod_{t_i<t} \frac{n_i - d_i}{n_i} \qquad (2)$$

$$\hat{\mathrm{H}}(t) = \sum_{t_i \le t} \frac{d_i}{n_i} \qquad (3)$$

Where d_i are the number of death events at time t, n_i is the number of subjects at risk of death just prior to time t.

The Weibull probability density function is shown below:

$$f(t) = \lambda P \times (\lambda t)^{\rho-1} \times \exp(-(\lambda t)^{\rho}) \qquad (4)$$

According to the formulation in the above section, the survival function and hazard function can be derived below:

$$S(t) = \exp(-\lambda t^{\rho}) \qquad (5)$$

$$h(t) = (\lambda P)(\lambda t)^{\rho-1} \qquad (6)$$

Where λ, ρ are the parameters of the Weibull distribution.

Univariant analysis results

In this subsection, the survival function and cumulative hazard function of LCD are calculated using the KM, NA and Weibull estimator. The result is shown in **Figure 6**. The blue curve represents the KM, NA estimator, the yellow curve represents the Weibull estimator, and the shaded area represents the confidence interval.

It can be found that the survival and cumulative hazard function of the two kind of estimator have a high degree of coincidence. However, there are still some differences that deserves our attention. The KM survival probability is higher than that of Weibull at the beginning half of time, and is lower than that of Weibull at the second half of time. The reason for this is because in the case of using the parametric estimator, the approximate form of survival and hazard function is determined in advance, while using the non-parametric estimator does not need to know the distributions form. Since we don't know the survival function of the LCD in advance, we believe that the non-parametric method is more appropriate for describing the survival function of the LCD than parametric method.

The median survival time is defined as the time where on average 50% of the LCD has expired. That means each vehicle has a 50% chance of completing its lane change maneuver. In this study, the value of median LCD of KM estimator is 6.7s, and the value of Weibull estimator is 6.6s. Meanwhile, the confidence interval of the median LCD can also be obtained. The lower confidence value of KM estimator and Weibull estimator are respectively 6.66s and 6.58s. The upper confidence value of KM estimate and Weibull estimate are respectively 6.75s and 6.72s. It can be found that the results of each estimator are very close to each other.

According to the above classification of LC behaviors, we segment the data on the above three regimes (different LC behaviors), and obtain the survival function of each group. The result is shown in **Figure 7**, and the at-risk percentage of counts of each group at each time point is also given in **Table 3**.

It can be found that the values of survival probability of three groups maintain at the level of 1 at the beginning 2 seconds, which indicates that most of the LCDs are longer than 2 seconds. The survival function declines rapidly in 4s~10s, which indicates that a large number of vehicles complete LC within 10s. From the below **Table 3**, it can be found that nearly 70% of vehicles complete lane change within 6s, nearly 80% of vehicles complete lane change within 8s, nearly 95% of vehicles complete lane change within 10s.

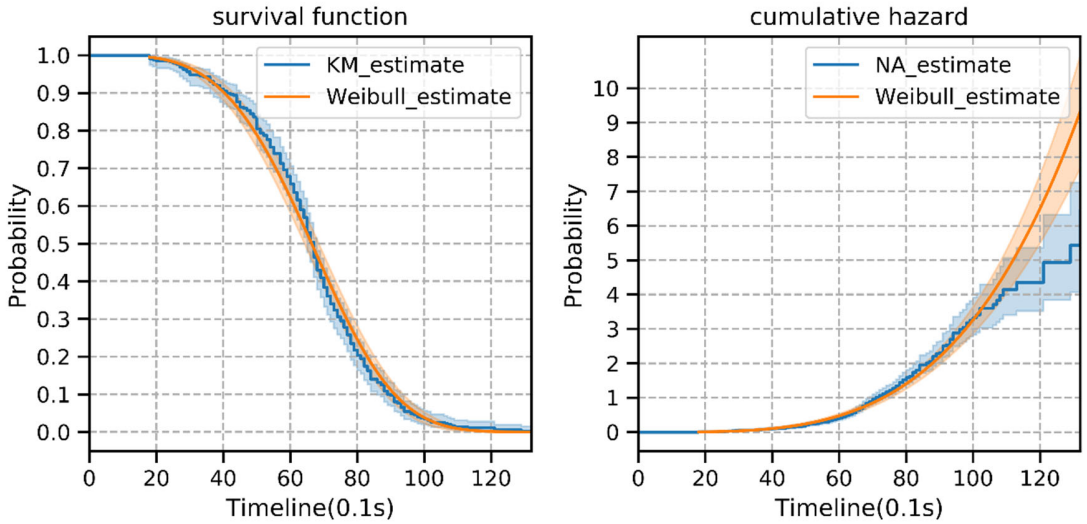

Figure 6 The survival and cumulative hazard function of LCD

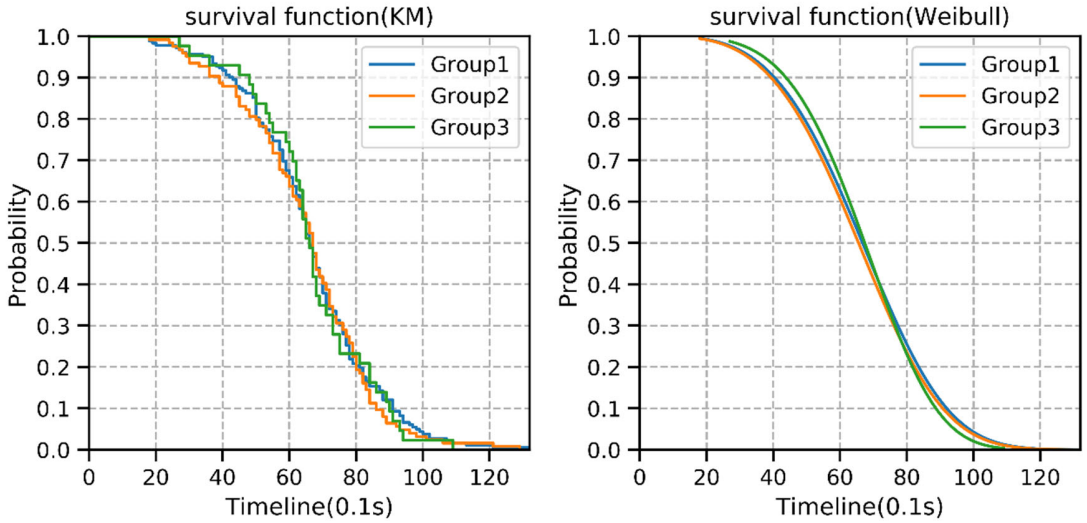

Figure 7 The survival function of different lane change behaviors

Table 3 The percentage of at-risk vehicles at each time points using the KM estimator

Group	Timeline(0.1s)						
	0	20	40	60	80	100	120
Group1	100.0%	98.4%	92.3%	67.6%	20.9%	4.4%	1.1%
Group2	100.0%	99.2%	88.7%	66.1%	22.6%	3.2%	1.6%
Group3	100.0%	100.0%	93.0%	74.4%	23.3%	2.3%	0.0%

It's worth noting that the above three curves have a high degree of coincidence. In order to further study the difference of the survival functions of the above three groups. We introduce the metric of RMST (Restricted mean survival times) to compare the survival functions. RMST is similar to ROC curve, which refers to the area under the survival curve in a certain period of time, that is the average survival time.

Comparison of survival difference via RMST integrates difference in survival probabilities from the beginning to the end of the follow-up time is a more meaningful metric.

$$RMST(t) = \int_0^t S(t)dt \quad (7)$$

In this study, the RMST means the cumulative survival probability of LCD at the given period time *t*. The larger the value of RMST of a group is, the longer the LCD of this group is. The RMST difference between two groups can be a measure of the relative size of LCD and the results of RMST comparison between each group is shown in **Figure 8.** It can be found that at the first 20 seconds, the difference in RMST between different groups remain at 0. With the increase of time, the difference gradually appears.

The RMST difference between Group 1 and Group 2 is always positive both using the KM and Weibull estimator. If we analyze the figure on the lower right, it is interesting to find that the red curve is always increasing. This means that the survival probability of Group 1 at each time point is almost higher than that of Group 2. This indicates that the average survival time of the vehicles in Group 1 is longer than the vehicles in Group 2. Therefore, it can be considered that the vehicles moving from the slow lane to a fast lane are more likely to have a longer LCD than the vehicles moving from the fast lane to a slow lane.

Meanwhile, the RMST difference between Group 1 and Group 3 is always negative, so as the value between Group 2 and Group 3. This may indicate that the vehicles with more comfortable LC conditions are more likely to have a longer LCD than the conditions in Group 1 and Group 2. If we further analyze the trends of the blue curve and green curve in the lower right figure, the lowest RMST difference value appears at 7.2s and 7.8s separately. This indicates that before the lowest value, the vehicles in Group 3 have higher probability of completing LC than Group 1 and Group 2 at each time step. After that, the vehicles in Group 1 and Group 2 have longer probability of completing LC than Group 3. However, in general, the vehicles in Group 3 have shorter LCD than that in Group 1 and Group 2 because the RMST difference curves are always below 0.

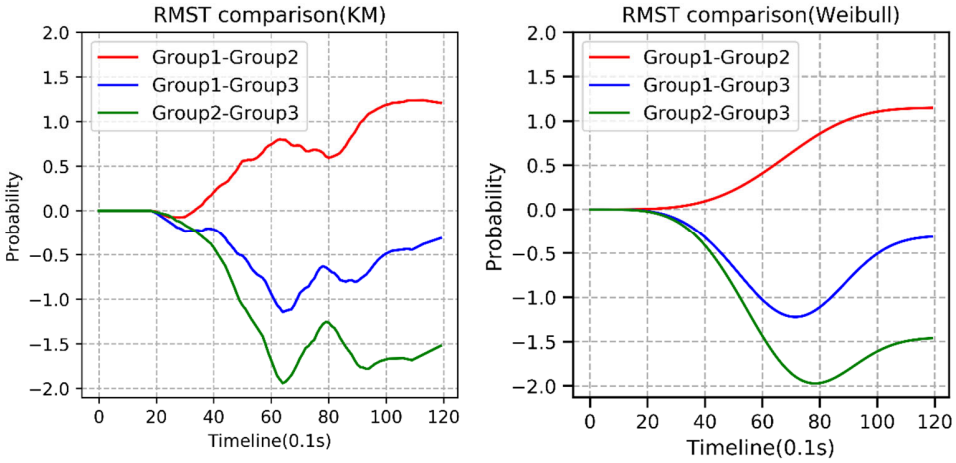

Figure 8 The RMST difference between different lane change behaviors

As for other variables, we also apply the RMST metric to explore the univariant analysis of LCD. We choose $Spe\_s(0)$, $Spediff\_f(0)$, $Spediff\_tf(0)$, $Spediff\_tr(0)$, $Dis\_f(0)$, and $Dis\_trf(0)$ as the research variables, and divide these variables into four clusters according to their corresponding minimum value, maximum value, 25th percentile value, 50th percentile value and 75th percentile value. Cluster 1 corresponds to the value between minimum value and 25th percentile value;

Cluster 2 corresponds to the value between 25th percentile value and 50th percentile value; Cluster 3 corresponds to the value between 50th percentile value and 75th percentile value; Cluster 4 corresponds to the value between 75th percentile value and maximum value. Then the RMST difference between each group of each variable is calculated and is shown in Figure 9.

It can be found that the RMST value of the group with the smaller $Spediff\_f(0)$ is always higher than the group with the larger $Spediff\_f(0)$. This means that the average survival time decreases with the increase of $Spediff\_f(0)$. Therefore, it can be deduced that if the speed of the subject vehicle is higher than the speed of the front vehicle, the average survival time of the subject vehicles will become smaller, which means that the subject vehicles are more likely to have a shorter LCD. Meanwhile, the RMST of the group with the smaller $Dis\_f(0)$ value is always lower than the group with the larger $Dis\_f(0)$ value. This means that with the increase of front spacing distance, the average survival time of the subject vehicle increase, which means that the subject vehicles are more likely to have a longer LCD. As for other variables, the RMST difference values are not all larger than 0 or less than 0, but this does not mean that these variables have no impacts on the LCD.

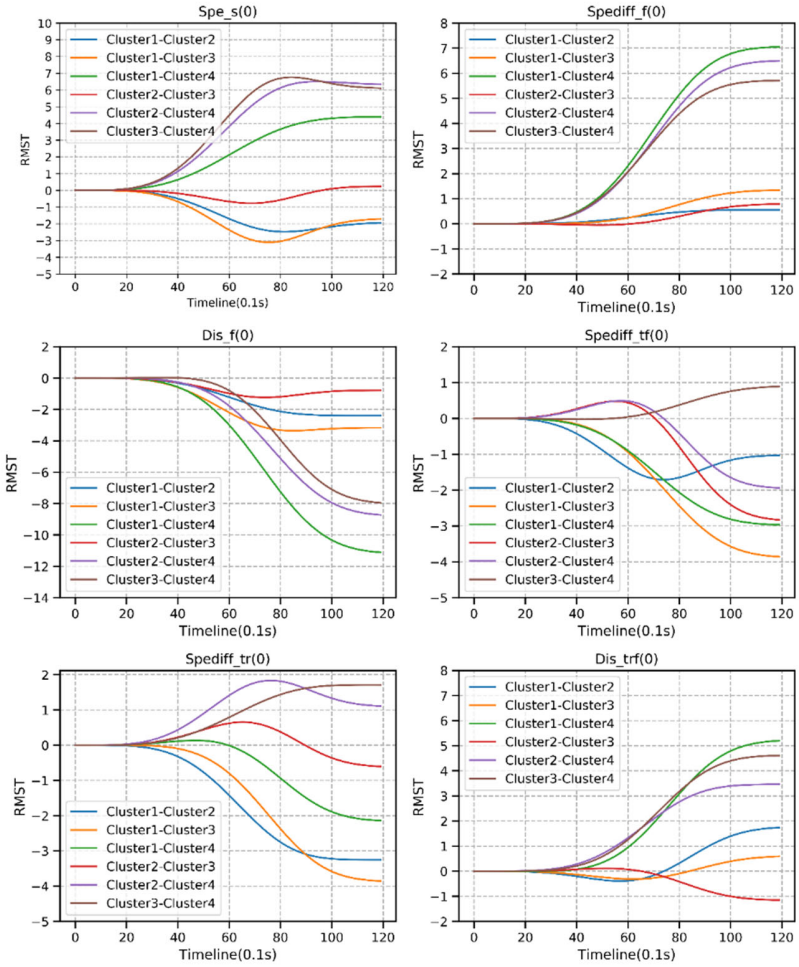

Figure 9 The RMST difference between different variables in each group

MODELING THE LCD USING THE REGRESSION SURVIVAL MODEL

Regression survival model

In this subsection, we present a brief introduction of the time-dependent proportional hazard-based model (*25; 26*), which can also be called as TDCM (time-dependent cox model). Unlike the time-invariant hazard-based model in the existing literatures(*6; 12; 13*), the explanatory variable change at every moment during the process of LC. The general mathematical description of TDCM is given below:

$$h(t \mid \overline{Z}) = h_0(t)\exp(\beta Z(t)) \quad (8)$$

$$S(t \mid \overline{Z}) = \exp(-\int_0^t h_0(t)\exp\left(\beta Z(t)\right)dt \quad (9)$$

Where Z denotes a set of possibly time-dependent covariates, $Z(t)$ denotes the value of Z at time t, and $\overline{Z}(t) = \{Z(s): 0 \le s \le t\}$ denotes the history of the covariates up to time t. $h_0(t)$ and $\exp(\beta Z(t))$ denote the baseline hazard and partial hazard respectively. β are a set of unknown regression parameters.

Regression analysis results

The reason why we employ the time-dependent proportional hazard-based model is that this model is capable of modeling the LCD as a continuous event rather than only as an instantaneous event (as detailed in the introduction). Therefore, in this subsection, we apply the time-dependent proportional hazard-based model to model the continuous impacts of influencing factors on LCD.

First, we temporarily choose nine variables to model the LCD. For each LC trajectory, there are a total of $T\_lcd / 0.1$ observation records. Meanwhile, each record not only contains the values of the nine variables, but also contains a binary variable which indicates whether the LC is over or not. Therefore, we can utilize the information of all variables at each timestep during the process of LC for each LC trajectory so as to model the continuous impacts of influencing factors on LCD. Meanwhile, we add the L1 and L2 penalty terms to the regression model so as to stabilize the coefficients and reduce the value to 0. Considering the target front-rear spacing distance may be no longer important during the LC for the subject vehicle, so we remove this time-varying variable and retain its value at the beginning time of LC. Moreover, considering the single trajectory have multiple observation records, we set the trajectory id variable as the cluster variable.

According to the above univariant analysis of LC behaviors, we introduce the stratification to the label variable. For each LC behavior category, it has its own baseline hazard, but they share the same regression parameters. This means that the differences in LC behavior are modeled in the regression model, and are reflected by the baseline hazard. Finally, we perform stepwise model selection by AIC to obtain the optimal model, and the model results are shown in **Table 4**.

Table 4 The improved result of the time-dependent proportional hazard-based model

variable	coef	exp(coef)	se(coef)	robust se	z	Pr(>|z|)
$Spe\_s(t)$	0.045	1.046	0.018	0.019	2.411	0.016
$Dis\_f(t)$	-0.026	0.974	0.006	0.006	-4.548	0.000
$Dis\_tf(t)$	-0.010	0.990	0.008	0.007	-1.507	0.132
$Dis\_trf(0)$	0.009	1.009	0.005	0.004	2.000	0.045
$Spediff\_tr(t)$	0.090	1.094	0.034	0.030	2.961	0.003
degree of freedom	5.000					
Concordance	0.624					

Likelihood ratio test	48.940	p-value	0.000
Wald test	48.690	p-value	0.000
Score (logrank) test	46.540	p-value	0.000
AIC	2710.502		

It can be found that the AIC value in **Table 4** is 2710.502. The result in **Table 4** indicates that there are four variables which simultaneously and significantly impact the LCD. There variables are $Spe\_s(t), Dis\_f(t), Dis\_trf(0), Spediff\_tr(t)$, and their regression coefficient are 0.045, -0.026, 0.009, and 0.090 respectively. Take the variable $Spe\_s(t)$ as example, without considering the influence of other factors, for every increase of 1m/s of $Spe\_s(t)$, the risk of completing the LC will increase by 1.046 times. Meanwhile, it can be found that the coefficients of $Spe\_s(t), Dis\_trf(0), Spediff\_tr(t)$ are all higher than 0. This indicates that with the increase of these variables, the risk of completing the LC will increase accordingly, and thus the vehicles are more likely to have a shorter LCD. While the coefficient of $Dis\_f(t)$ is less than 0, this indicates with the increase of front-spacing distance, the vehicles is more likely to have a longer LCD.

After determining the parameters of the model, it is necessary for us to test the model. When using the proportional hazard-based model, there is a very important prerequisite, namely the proportional hazard assumption, or PH assumption for short. PH assumes that the influence of covariates on survival rate does not change over time, that $h(t)/\lambda_0(t)$ is a fixed value. Only when the PH hypothesis is satisfied, the results of cox model are meaningful. The result of Schoenfeld residuals is shown in **Figure 10.** The horizontal axis in **Figure 10** represents time, the solid line represents the fitted smooth curve and the dotted line represents the standard deviation of 2 units. It can be found that the residuals of each variable are within standard deviation of 2 units and all the p-values are higher than 0.05. This means that each covariate satisfied the PH risk assumption.

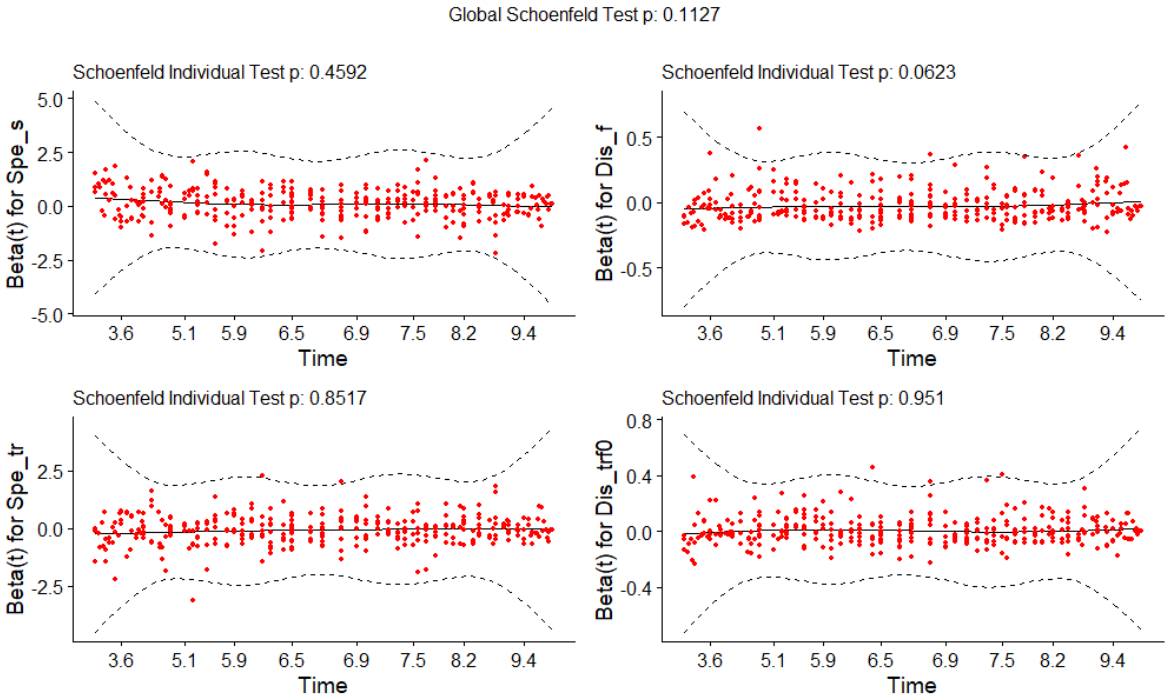

Figure 10 The Schoenfeld residuals of the TDCM model

Conclusion and future work

Lane-changing behavior is a very demanding and complex driving task in actual traffic flow. However, little efforts have been made on researching the LC action itself, especially researching the LCD. Therefore, this paper aims to present a complete survival analysis of LCD in order to have a deep understanding of LC behavior.

In short, both univariant survival model and regression survival model are introduced to model the LCD. Before modeling the LCD, the GMM is introduced to classify different LC behaviors so as to investigate the LCD difference under different LC behaviors, which aims to have a more thorough understanding of LC behavior. After that, in the univariate analysis of LCD, the non-parametric and full-parametric estimators are employed to model the survival function of LCD. Then, the differences in LCD for different LC behaviors have been investigated through introducing the RMST metric. In the regression analysis, the LCD is modeled as a continuous event using the time-dependent proportional hazard-based model, which revealed the significant factors influencing LC behavior.

Findings in the above results are interesting. (1) In the univariate survival analysis results, vehicles moving to a fast lane are more likely to have a longer LCD than those moving to a slow lane; the vehicles with comfortable LC conditions are more likely to have a longer LCD. Meanwhile, the interaction with the front vehicle significant impacts the LCD significantly more than other vehicles. (2) In the regression survival analysis results, the speeds of the subject vehicle, the speed of the target rear vehicle, the front spacing distance and the initial target front-rear spacing distance significantly impact the LCD during the process of LC. Furthermore, with the increase of $Spe\_s(t), Spe\_tr(t), Dis\_trf(0)$, the risk of completing the LC will increase, which are more likely to lead a shorter LCD. While with the increase of $Dis\_f(t)$, the subject vehicle is more likely to have a longer LCD.

In summary, efforts have been made to bridge the following three research gaps in LCD: (1) This paper presents a comprehensive survival analysis of LCD including both univariant survival model and regression survival model; (2) This paper investigates the difference in LCD under different LC behaviors, and the result indicates that there are indeed differences; (3) This paper models the LCD as a continuous event using the time-dependent proportional hazard-based model. Findings drawn from these three parts may further inspire our understanding of LC behaviors. Meanwhile, these results may be helpful and applicable to the design of the ADAS, thus making the LC action safer and more convenient.

It is worth noting that due to the limitations of the dataset used in this paper, this study only investigates the impacts of the interactions with the surrounding vehicles on LCD. Nevertheless, we believe that the analytical methods adopted in this paper are also applicable to other datasets. Future work of this study may be improved through collecting data about driver's characteristic, weather condition, different road types and so as to further have a deeper understanding of LC behaviors.

ACKNOWLEDGMENTS

This paper presents a part of the first author's PhD research. This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors. Authors thank the anonymous reviewers for their constructive comments and help to improve our paper.

AUTHOR CONTRIBUTIONS

The authors confirm contribution to the paper as follows: study conception and design: Yang Li. Author, Linbo Li. Author; analysis and interpretation of results: Yang Li. Author, Daiheng Ni. Author, Linbo Li. Author; draft manuscript preparation: Yang Li. Author, Daiheng Ni. Author. All authors reviewed the results and approved the final version of the manuscript.

REFERENCES

1 Zheng, Z. Recent Developments And Research Needs In Modeling Lane Changing. *Transportation Research Part B: Methodological,* Vol. 60, 2014, pp. 16-32.

2 Ni, D. Chapter 2 - Traffic Flow Characteristics I.In *Traffic Flow Theory*, Butterworth-Heinemann, 2016. pp. 19-35.

3 Laval, J. A., and C. F. Daganzo. Lane-Changing In Traffic Streams. *Transportation Research Part B: Methodological,* Vol. 40, No. 3, 2006, pp. 251-264.

4 Yang, M., X. Wang, and M. Quddus. Examining Lane Change Gap Acceptance, Duration And Impact Using Naturalistic Driving Data. *Transportation Research Part C: Emerging Technologies,* Vol. 104, 2019, pp. 317-331.

5 Toledo, T., and D. Zohar. Modeling Duration Of Lane Changes. *Transportation Research Record,* Vol. 1999, No. 1, 2007, pp. 71-78.

6 Ali, Y., M. M. Haque, Z. Zheng, S. Washington, and M. Yildirimoglu. A Hazard-Based Duration Model To Quantify The Impact Of Connected Driving Environment On Safety During Mandatory Lane-Changing. *Transportation Research Part C: Emerging Technologies,* Vol. 106, 2019, pp. 113-131.

7 Cao, X., W. Young, and M. Sarvi. Exploring Duration Of Lane Change Execution.In *Australasian Transport Research Forum*, Australia, 2013.

8 Aghabayk, K., S. Moridpour, W. Young, M. Sarvi, and Y.-B. Wang. Comparing Heavy Vehicle And Passenger Car Lane-Changing Maneuvers On Arterial Roads And Freeways. *Transportation Research Record,* Vol. 2260, No. 1, 2011, pp. 94-101.

9 Olsen, E. C., S. E. Lee, W. W. Wierwille, and M. J. Goodman. Analysis Of Distribution, Frequency, And Duration Of Naturalistic Lane Changes. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting,* Vol. 46, No. 22, 2002, pp. 1789-1793.

10 Moridpour, S., M. Sarvi, and G. Rose. Modeling The Lane-Changing Execution Of Multiclass Vehicles Under Heavy Traffic Conditions. *Transportation Research Record,* Vol. 2161, No. 1, 2010, pp. 11-19.

11 Wang, Q., Z. Li, and L. Li. Investigation Of Discretionary Lane-Change Characteristics Using Next-Generation Simulation Data Sets. *Journal of Intelligent Transportation Systems,* Vol. 18, No. 3, 2014, pp. 246-253.

12 Vlahogianni, E. I. Modeling Duration Of Overtaking In Two Lane Highways. *Transportation Research Part F: Traffic Psychology And Behaviour,* Vol. 20, 2013, pp. 135-146.

13 Wu, J., S. Zhang, A. K. Singh, and S. Qin. Hazard-Based Model Of Mandatory Lane Change Duration.In *17th COTA International Conference of Transportation*, China, 2008.

14 Bian, Y., J. Ding, M. Hu, Q. Xu, J. Wang, and K. Li. An Advanced Lane-Keeping Assistance System With Switchable Assistance Modes. *IEEE Transactions on Intelligent Transportation Systems,* Vol. 21, No. 1, 2020, pp. 385-396.

15 Zhu, B., S. Yan, J. Zhao, and W. Deng. Personalized Lane-Change Assistance System With Driver Behavior Identification. *IEEE Transactions on Vehicular Technology,* Vol. 67, No. 11, 2018, pp. 10293-10306.

16 Yang, L., X. Li, W. Guan, H. M. Zhang, and L. Fan. Effect Of Traffic Density On Drivers' Lane Change And Overtaking Maneuvers In Freeway Situation—A Driving Simulator–Based Study. *Traffic Injury Prevention,* Vol. 19, No. 6, 2018, pp. 594-600.

17 Gurupackiam, S., and S. L. Jones Jr. Empirical Study Of Accepted Gap And Lane Change Duration Within Arterial Traffic Under Recurrent And Non-Recurrent Congestion. *International Journal for Traffic & Transport Engineering,* Vol. 2, No. 4, 2012, pp. 306-322.

18 Zhou, J., and F. Zhu. Modeling The Fundamental Diagram Of Mixed Human-Driven And Connected Automated Vehicles. *Transportation Research Part C: Emerging Technologies,* Vol. 115, 2020, p. 102614.

19 Washington, S., M. G. Karlaftis, F. Mannering, and P. Anastasopoulos. *Statistical And Econometric Methods For Transportation Data Analysis*. CRC press, 2020.

20 Roque, C., and M. Jalayer. Improving Roadside Design Policies For Safety Enhancement Using Hazard-Based Duration Modeling. *Accident Analysis & Prevention,* Vol. 120, 2018, pp. 165-173.

21 Alexiadis, V., J. Colyar, J. Halkias, R. Hranac, and G. McHale. The Next Generation Simulation Program. *Institute of Transportation Engineers. ITE Journal,* Vol. 74, No. 8, 2004, p. 22.

22 Li, L., R. Jiang, Z. He, X. Chen, and X. Zhou. Trajectory Data-Based Traffic Flow Studies: A Revisit. *Transportation Research Part C: Emerging Technologies,* Vol. 114, 2020, pp. 225-240.

23 Xie, D.-F., Z.-Z. Fang, B. Jia, and Z. He. A Data-Driven Lane-Changing Model Based On Deep Learning. *Transportation Research Part C: Emerging Technologies,* Vol. 106, 2019, pp. 41-60.

24 Chu, K.-C., R. Saigal, and K. Saitou. Stochastic Lagrangian Traffic Flow Modeling And Real-Time Traffic Prediction.In *2016 IEEE International Conference on Automation Science and Engineering (CASE)*, IEEE, USA, 2016. pp. 213-218.

25 Fisher, L. D., and D. Y. Lin. Time-Dependent Covariates In The Cox Proportional-Hazards Regression Model. *Annual Review of Public Health,* Vol. 20, No. 1, 1999, pp. 145-157.

26 Cox, D. R. Regression Models And Life-Tables. *Journal of the Royal Statistical Society: Series B (Methodological),* Vol. 34, No. 2, 1972, pp. 187-202.