Hindawi
Journal of Advanced Transportation
Volume 2018, Article ID 5604375, 15 pages
https://doi.org/10.1155/2018/5604375

Research Article

Characterizing Heterogeneity in Drivers' Merging Maneuvers Using Two-Step Cluster Analysis

Gen Li[1] **and Lu Sun**[2]

[1]*School of Transportation, Southeast University, Nanjing 210096, China*
[2]*Department of Operations Research and Financial Engineering, Princeton University, 229 Sherrerd Hall, Princeton, NJ 08544, USA*

Correspondence should be addressed to Gen Li; gilg4226307@aliyun.com

Received 13 October 2017; Accepted 15 April 2018; Published 17 May 2018

Academic Editor: Ludovic Leclercq

In order to investigate the heterogeneity in merging behaviors on freeways, a novel data mining tool, called two-step cluster analysis, is applied to the merging maneuvers (namely, initial speed, merging speed, and merging position). Merging maneuvers of 370 drivers collected from the NGSIM dataset are automatically and optimally segmented into four clusters (Early Merging Drivers at High Speed, Early Merging Drivers at Low Speed, Late Merging Drivers at Low Speed, and Late Merging Drivers at High Speed) by the two-step cluster analysis. Hypothesis test confirms the significant differences in merging maneuvers between different clusters. The clustered data are used to find the best corresponding fitting distributions. Seven distributions (Normal, Log-normal, Student's *t*, Logistic, Log-Logistic, Gamma, and Weibull) are considered for each cluster and the Kolmogorov-Smirnov test statics are used to select the best fitted distributions. It is found that merging drivers may merge either early or late, under congestion or uncongested traffic condition. Further analysis of merging durations shows that Late Merging Drivers use significantly shorter time than Early Merging Drivers to finish the merging maneuver, no matter if they are at high or at low speed. Hypothesis test of accepted lead gaps and lag gaps indicate that merging drivers are more sensitive to the lag gaps under congestion. The proposed method can automatically identify the heterogeneity in merging drivers and the results obtained in this paper can be used to enhance the accuracy of the merge behavior models in microscopic simulation software.

1. Introduction

Car following and lane changing are two fundamental tasks conducted by drivers and the car-following model and lane-changing model are two of the most significant submodels in microscopic traffic simulation models. However, lane-changing models have not been given as much attention as car-following models [1–4].

Lane changes can be divided into two types: discretionary lane changes and mandatory lane changes. Merging behavior is one kind of typical mandatory lane changes when vehicles have to move from an on-ramp to the main road. Merging behavior has been confirmed to have impact influence on the traffic operations, and it may trigger traffic congestions and breakdowns [5, 6].

Recently, several studies criticized the ability of lane-changing models to represent reality due to the ignorance of driver heterogeneity [7–10] (Hoogendoorn et al. 2006). Several studies have investigated the heterogeneity among drivers using macroscopic traffic models [11–15]. Other studies investigated the driver heterogeneity during car-following process [9, 10]. It is believed that accommodating the heterogeneity within the driver population is inevitable for producing a more precise car-following model (Kim et al. 2013). However, heterogeneity in lane-changing models has not received much emphasis in the literature.

To investigate the heterogeneity in merging behaviors, this study proposes a new data mining tool, called two-step cluster analysis, combining cluster analysis to capture the heterogeneity in merging behaviors by dividing merging into some homogeneous components. The paper is organized as follows. The next section will provide a literature review on the existing studies followed by Section 3, which describes the NGSIM data used in this paper. Section 4 gives the methodology. Results and discussions are presented in Section 5. Finally, the conclusions are presented in Section 6.

2. Literature Review

Gap acceptance was considered to be the most important part in existing lane-changing models [17–21] (Lee 2006). It is assumed that a driver makes a lane change when both the lead and lag gaps in the target lane are larger than the so-called critical gap. Different definitions of critical gap were used in different studies. Herman and Weiss [22] assumed an exponential distribution for critical gaps, Drew et al. [23] assumed Log-normal distribution, and Miller [24] assumed a normal distribution. Gipps [25] is believed to be the first to use gap acceptance theory to develop a comprehensive framework for lane-changing model. Hidas [26] and Wang [27] applied similar principle to build freeway mandatory lane-changing models. The framework of Gipps' model has also been use in several microscopic traffic simulation software [28–30]. In VISSIM, for a mandatory lane change, the critical gap depends on the acceptable maximum deceleration for the lane-changing driver and his presumed follower. In CORSIM, ten different driver types can be defined with variable gap acceptance values [28] and each gap acceptance decision is independent considering the current available gap and the personal gap acceptance value. In gap acceptance models using critical gaps, the most basic assumption is that a driver will accept the adjacent gap only if both the lead and lag gaps are larger than the critical gap [31] (Ma and Andreasson 2007; Hastie et al. 2001; Martin et al. 2012). However, this assumption is often criticized as it is often inconsistent with the real world observations that vehicles still take lane changes when only the lead or lag gap or even none of them are larger than the critical gap [2, 32, 33]. To overcome this deficiency, a binary logit model was built by Kita (1993). Weng and Meng [34] and Marczak et al. [32] used the same kind of model to predict merging decision in short-term work zone merging areas and to compare the gap acceptance of merging decision at two sites, respectively.

It was claimed that the existing lane-changing models could not reflect the real traffic behavior under congestion [26, 35] (Ahmed et al. 1996). Thus, "forced" and "cooperative" lane-changing models were proposed to describe distinctive behaviors of vehicles under congestion [18, 26] (Ahmed et al. 1996; Hidas 2005; Lee 2006; Rao 2006). A framework for merging behavior with latent plans was proposed by Choudhury et al. (2007). In this paper, three plans, normal merge, merge with courtesy and forced merge, were considered to be latent in the model. However, only accepted gaps were considered and rejected gaps were ignored, and some of the estimated coefficients in the model were not significant [32].

In recent years, driver heterogeneity has been considered as a key element in driver behavior and investigated in several macroscopic traffic studies [12, 14, 36–39] (Sun et al. 2001). In microscopic traffic models, heterogeneity in car-following models was investigated by deriving the joint distribution of model coefficients depending on an empirical basis [7–10] (Hoogendoorn et al. 2006). However, driver heterogeneity in lane-changing models has not drawn similar attention, and no studies try to investigate the heterogeneity among lane-changing drivers.

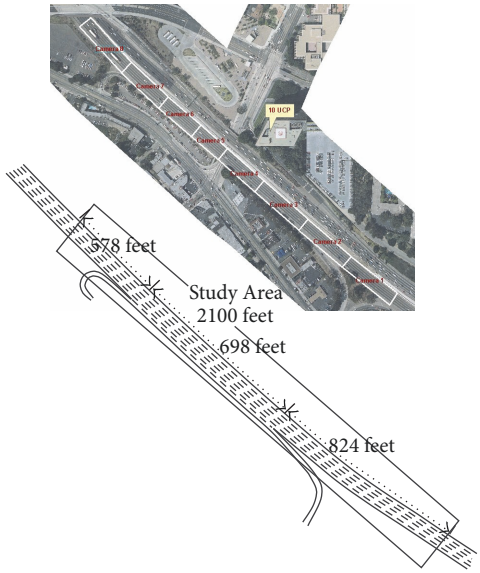


Figure 1: US Highway 101 study corridor from NGSIM [16].

Clustering is a widely used technique in data mining applications for identifying patterns in datasets. However, there are few studies that apply cluster analysis to identifying heterogeneity in drivers. The only related study was conducted by Hill et al. [40], in which 46 participant drivers were classified into 4 clusters based their behaviors in driving experiments using *k-means* clustering method. Two-step cluster analysis is a statistical method that can automatically identify similar clusters of people or objects within data sets [41]. Two-step cluster analysis is also considered more reliable and accurate than traditional clustering methods such as the *k*-means clustering algorithm [42].

A methodology framework than combines two-step cluster analysis; hypothesis test and probability distribution fitting are proposed in this paper. The two-step cluster analysis can naturally and automatically segment the merging drivers into optimal number of clusters. The hypothesis test can identify the significant differences between the clusters and capture the heterogeneity in merging behaviors. At last different probability distribution functions of related parameters can be identified through analyzing the clusters.

3. Data Preparation

In order to investigate the merging behaviors in the weaving sections, the NGSIM vehicle trajectory data collected on a segment of southbound US Highway 101 (Hollywood Freeway) in Los Angeles, CA, are chosen in this paper [43]. Figure 1 shows the site for US Highway 101. This US-101 section is 640 meters long and has five main lanes and one auxiliary lane. The vehicle trajectories were collected from 7:50 a.m. to 8:35 a.m. on June 15, 2005. The road section was covered by eight cameras and the dataset was updated at a resolution of 10 frames per second [16].

In our research, we focused on the merging behavior. The NGSIM database distinguishes the on-ramp, auxiliary lane,

and main lane clearly. Thus, it is easy for us to collect the trajectories of 399 merging vehicles. However, the previous study [31, 44] shows that the original trajectory data seem to contain some noise and errors. The velocities and acceleration provided by the NGSIM cannot be directly used. Thus, we applied the following data smoothing techniques proposed by Thiemann et al. [44]:

(1) The velocities and acceleration of vehicles are directly estimated from the longitudinal positions.

(2) The locations (both local lateral and longitudinal coordinates), velocities, and acceleration of vehicles are smoothed by the symmetric exponential moving average filter (sEMA) proposed by Thiemann et al. [44] to decrease measurement errors in the data. The smoothing times of sEMA method are set as the suggested values for the US Highway 101 data set in Thiemann et al. [44].

Although the random errors can be reduced by the smoothing process, there are still some errors in the data. Thus, the following heuristic rules are applied to filter the data sets:

(1) Filter out the data with motorcycle or heavy vehicle to focus on merge behaviors of cars.

(2) Filter out the data if the merging vehicle cross the lane line before entering the auxiliary lane.

(3) Filter out the data if the merging vehicle cross the lane line at the position further than the length of auxiliary lane

After filtering and smoothing the trajectory data, we conduct a further consistency checking to insure the consistency and accuracy of all the data subsets from US-101 dataset.

First, we check the frame ID, number of frames, vehicle length, vehicle width, and vehicle type and find that the frame IDs, the total numbers of the frames, global time, and the vehicle length and width of each vehicle in US-101 dataset are found to be consistent. However, through a searching process, we found that the local coordinates and global coordinates of the three sub-datasets are inconsistent. We search several data points that have the same global coordinates among the three data subsets of US-101 datasets and found that the three data subsets of US-101 dataset are consistent in local x, but inconsistent in local y. As this paper is focused on merging behavior, we use the beginning point of the auxiliary lane, for example. The local y for the beginning points of the auxiliary lane for three data sets are 615.402 ft, 655.675 ft, and 650.172 ft, respectively. It means the upstream edge in data set 1 is at 40.273 ft in data set 2 and 34.770 ft in data set 3. Thus, the three datasets are unified by using the local coordinates of data set 2.

After smoothing, filtering, and consistency checking, 370 trajectories are kept. After getting the trajectories, the parameters of merge maneuvers and gap acceptance can be interpolated. In this paper, merge maneuvers include merging speed, merging position, and merging duration. The parameters of gap acceptance include the accepted gaps, lead gaps, and lag gaps.

For each identified merging vehicle, the merging speed, merging position, accepted gap, accepted lead gap, and lag gap are collected when the front center of the merging vehicle crossed the lane markers.

To extract the merging durations, the method used to determine the start and end point of discrepancy lane-changing behavior in Wang et al. [1] is used in this paper. For each merging vehicle, the start time is taken as the first time when the merging vehicle has the lateral velocity larger than 0.2 m/s and the end time is taken as the first time when the merging vehicle has the lateral velocity less than 0.2 m/s after crossing the lane line.

4. Methodology

4.1. Two-Step Cluster Analysis. The two-step cluster analysis is a scalable cluster analysis algorithm that was designed to manage large datasets. The analysis has two steps: preclustering and clustering.

In the preclustering step, all the cases in the data are scanned and the log-likelihood distance between them is measured to determine whether they are going to form preclusters based on some threshold distance criterion [45].

The log-likelihood distance is a probability based distance. The distance between two clusters is related to the decrease in log-likelihood as they are combined into one cluster. The distance between clusters j and s is defined as follows:

$$d(i, j) = \varepsilon_i + \varepsilon_j - \varepsilon_{\langle i,j\rangle} \tag{1}$$

$$\varepsilon_i = -n_i \left(\sum_{k=1}^{K} \frac{1}{2} \log\left(\widehat{\sigma}_{ik}^2 + \widehat{\sigma}_k^2\right) \right) \tag{2}$$

$$\varepsilon_j = -n_j \left(\sum_{k=1}^{K} \frac{1}{2} \log\left(\widehat{\sigma}_{jk}^2 + \widehat{\sigma}_k^2\right) \right) \tag{3}$$

$$\varepsilon_{\langle i,j\rangle} = -n_{\langle i,j\rangle} \left(\sum_{k=1}^{K} \frac{1}{2} \log\left(\widehat{\sigma}_{i,jk}^2 + \widehat{\sigma}_k^2\right) \right), \tag{4}$$

where $d(i, j)$ is the log-likelihood distance between clusters i and j; $\langle i, j\rangle$ represents the cluster formed by merging clusters i and j; K is the total number of continuous variables; $\widehat{\sigma}_k^2$ is the estimated variance of the continuous variable k for the entire dataset; $\widehat{\sigma}_{ik}^2$, $\widehat{\sigma}_{jk}^2$, and $\widehat{\sigma}_{i,jk}^2$ are the estimated variances of the continuous variable k in clusters i, j and $\langle i, j\rangle$.

In the second step, the subclusters resulting from the preclustering step are clustered into the optimal number of clusters using an agglomerative clustering algorithm. The subclusters are regarded as single cases and merged into one cluster by satisfying the minimum distance in (1) until all data are placed within a single cluster.

If a desired number of clusters is not predetermined, the algorithm can automatically determine the optimal number of cluster using Schwarz's Bayesian Criterion (BIC) or the Akaike Information Criterion (AIC) [45]:

$$\begin{aligned} \mathrm{AIC}(J) &= -2\sum_{j=1}^{J} \varepsilon_j + 2m_J \\ \mathrm{BIC}(J) &= -2\sum_{j=1}^{J} \varepsilon_j + m_J \log(N), \end{aligned} \tag{5}$$

where J is the number of clusters; K is the total number of continuous variables clusters; N is the total number of observations; $m_J = 2KJ$.

The two-step analysis is conducted using SPSS 22.0 in this paper.

4.2. Hypothesis Test for Difference between Two Medians. To statistically investigate if there are significant differences between different clusters, the Mann–Whitney U test is conducted in this paper.

Normally, Student's t-test is used to determine if the means of two sets of data are different from each other. However Student's t-test is most commonly applied when the test statistic would follow a Normal distribution if the value of a scaling term in the test statistic was known [46]. It has been shown in Section 4 that the distributions of merging maneuver parameters do not all follow Normal distribution. Thus, Mann–Whitney U test, which is a distribution-free test and has been used in previous study [47], is performed in this paper. It is assumed in Mann–Whitney U test that the populations must be continuous and their probability density functions must have the same shape and spread, which are somewhat less restrictive than the assumptions needed for Student's t-test [47].

4.3. Best Probability Distribution. The clustered data are used to find the best corresponding fitting distributions using EasyFit 5.4 [48, 49]. To best investigate the distributions of the merging parameters, 7 distributions (Normal, Log-normal, Student's t, Logistic, Log-Logistic, Gamma, and Weibull) are considered for each cluster and the Kolmogorov-Smirnov test is used to select the best fitted distribution. The Kolmogorov-Smirnov test statistic is defined by the greatest vertical distance at any point between the two cumulative probability distributions. The smaller the Kolmogorov-Smirnov static is and the bigger the p value is, the better the distribution can describe the sample.

4.3.1. Normal Distribution. A random variable X has Normal distribution with the location parameter μ and scale parameter σ^2, if its density reads the following:

$$f\left(x \mid \mu, \sigma^2\right) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right) \tag{6}$$

The Normal distribution is usually denoted as $X \sim N(\mu, \sigma^2)$.

4.3.2. Log-Normal. A random variable X has Log-normal distribution if its logarithm is normally distributed with location parameter μ and scale parameter σ^2, which is usually denoted as $\ln(X) \sim N(\mu, \sigma^2)$

4.3.3. Student's t. A random variable Y follows a skew-t distribution with location the parameter μ and scale parameter σ^2 and degrees of freedom ν if it has the following representation:

$$\begin{aligned} Y &= \mu + \sigma \frac{X}{W} \\ X &\sim N\left(\mu, \sigma^2\right), \\ W &\sim \Gamma\left(\frac{\nu}{2}, \frac{\nu}{2}\right), \end{aligned} \tag{7}$$

where X is a standard Normal distribution and Γ is the Gamma function.

4.3.4. Logistic. A random variable X follows a logistic distribution if its cumulative distribution function is the logistic function. It resembles the Normal distribution in shape but has heavier tails (higher kurtosis). The probability density function of the logistic distribution is given by the following:

$$f\left(x \mid \mu, s\right) = \frac{e^{-(x-\mu)/s}}{s\left(1 + e^{-(x-\mu)/s}\right)^2}, \tag{8}$$

where μ is the location parameter and s is the scale parameter. The logistic distribution is usually denoted as $X \sim \text{Logistic}(\mu, s)$.

4.3.5. Log-Logistic. A random variable X has log-logistic distribution if its logarithm has logistic distribution with the location parameter μ and scale parameter s, which is usually denoted as $\ln(X) \sim \text{Logistic}(\mu, s)$

4.3.6. Gamma. A random variable X follows Gamma distribution with the shape parameter μ and scale parameter k if its density reads the following:

$$f\left(x \mid k, \theta\right) = \frac{x^{k-1} e^{-x/\theta}}{\theta^k \Gamma\left(k\right)} \tag{9}$$

Gamma distribution is usually denoted as $X \sim \Gamma(\theta, k)$.

4.3.7. Weibull. The probability density function of a Weibull random variable is

$$f\left(x \mid \lambda, k\right) = \begin{cases} \frac{k}{\lambda}\left(\frac{x}{\lambda}\right)^{k-1} e^{-(x/\lambda)^k} & x \geq 0 \\ 0 & x < 0, \end{cases} \tag{10}$$

where $k > 0$ is the shape parameter and $\lambda > 0$ is the scale parameter. Weibull distribution is usually denoted as $X \sim \text{Weibull}(k, \lambda)$.

5. Cluster Results and Discussions

5.1. Clustering Results. To investigate the heterogeneity among merging drivers, the two-step cluster analysis is applied to segment the data of merging maneuvers. We choose the initial speed, merging speed, and merging position for cluster analysis. The initials speed is defined as the speed when the merging vehicle enters the auxiliary lane.

Table 1: Results of cluster analysis.

Parameters	Variables	Cluster 1	Cluster 2	Cluster 3	Cluster 4
proportion		33.6%	26%	24%	16.4%
mean	Initial Speed	14.547	10.481	13.427	16.768
	Merging Speed	14.186	9.478	9.950	15.613
	Merging Position	45.81	37.09	147.52	194.89
median	Initial Speed	14.167	10.936	13.576	16.663
	Merging Speed	14.082	10.104	10.525	14.921
	Merging Position	44.00	32.11	145.48	190.48
standard deviation	Initial Speed	1.566	2.035	2.069	1.763
	Merging Speed	1.630	2.366	2.511	2.943
	Merging Position	21.46	25.95	49.19	47.89
Skewness	Initial Speed	0.91	−1.12	−0.05	0.30
	Merging Speed	0.52	−0.95	−0.69	0.08
	Merging Position	0.68	1.34	0.28	0.79

It can more or less reflect the expected speeds of merging drivers. The merging speed is defined as the speed when the front center of merging vehicle crosses the lane line. It can be linked to the speed adjustment behavior on the auxiliary lane. And the merging position is defined as the position of the vehicle when the front center of merging vehicle crosses the lane line. All the three maneuvers are indicators of aggressiveness and can reflect drivers' personal characteristics.

The optimal number of clusters is automatically determined by the two-step cluster analysis. The optimal number is both 4 by using AIC and BIC criteria. Table 1 shows the results of the cluster analysis. Figure 2 shows the 3D scatter plots and 2D scatter plots of the clustering results.

For comparison, a K-means clustering approach was also performed and the cluster number was also set at 4. Figure 3 shows the 3D scatter plots and 2D scatter plots of the K-means clustering results. One can find that the K-means clustering cannot well separate the initial speed and merging speed and the results of the two-step cluster analysis show better separation results in all three variables.

From Table 1, one can observe that merging drivers in each cluster is 33.6%, 26%, 24%, and 16.4%, respectively. It means no cluster dominated the most prevailing merging drivers.

From Table 1, the mean merging positions of clusters 1, 2, 3, and 4 are 45.81, 37.09, 147.52, and 194.89 m, respectively. Both the initial speed and merging speed have the lowest values in cluster 2 and largest values in cluster 4. The merging speeds are smaller than the initial speed in all clusters. It is interesting to find that mean values of merging speeds of clusters 2 and 3 are similar. However, the mean values of initial speeds of clusters 2 and 3 are 10.481 m/s and 13.427 m/s, respectively, the difference of which is much larger than that of merging speeds. It indicates that the merging drivers in cluster 2 have lower expecting speed than cluster 3 and they might have prepared for merging before they get on the auxiliary lane.

One can further observe that drivers of cluster 1 and cluster 2 both merge very early (i.e., 45.81 for cluster 1 and 37.09 m for state 2), corresponding to the so-called Early Merging Drivers. However, there are essential differences between the two clusters in terms of mean values of initial speed and merging speed. The mean values of initial speed and merging speed in cluster 2 are much smaller than those cluster 1, indicating that the traffic condition of cluster 2 is better than cluster 1. Clearly, most merging drivers choose to merge as soon as possible at both low and high speeds. Drivers in clusters 1 and 2 are called Early Merging Drivers at High Speed and Early Merging Drivers at Low Speed.

It is interesting to find that cluster 3 consists of merging drivers that have the largest speed reduction during the merging process (i.e., from 13.427 to 9.950 m/s). It means that drivers in cluster 3 have much higher speeds than the mainline traffic speeds when they enter the auxiliary lane and they use the auxiliary lane to reduce the speed difference between their speeds and mainline traffic speeds. Thus, the mean merging position of cluster 3 is much bigger than that of clusters 1 and 2. Drivers in cluster 3 are called Late Merging Drivers at Low Speed.

Cluster 4 contains the merging drivers that merge latest in the auxiliary lane. Previous studies have stated that drivers merge late because they cannot find a gap larger than their critical gap under congestion. However, it is interesting to find that the mean values of initial speeds and merging speeds are also the largest among the four clusters, indicating that cluster 4 has the best traffic condition. It means that drivers in cluster 4 merge late probably because they use the auxiliary lane to get further downstream of the mainline rather than that they cannot find gaps larger than critical gaps. Drivers in cluster 4 are called Late Merging Drivers at High Speed.

5.2. Hypothesis Tests between Different Clusters. To statistically investigate if there are significant differences in the median values of merging durations and gap acceptance between different clusters, two-sample Mann–Whitney U Tests are carried out. The obtained p values are given in Table 2 and the significance level is set as 0.1.

From Table 2, one can find that most p values are 0.000 except for the merging speed when testing clusters 2 and 3. It means that the two-step cluster analysis can well distinguish

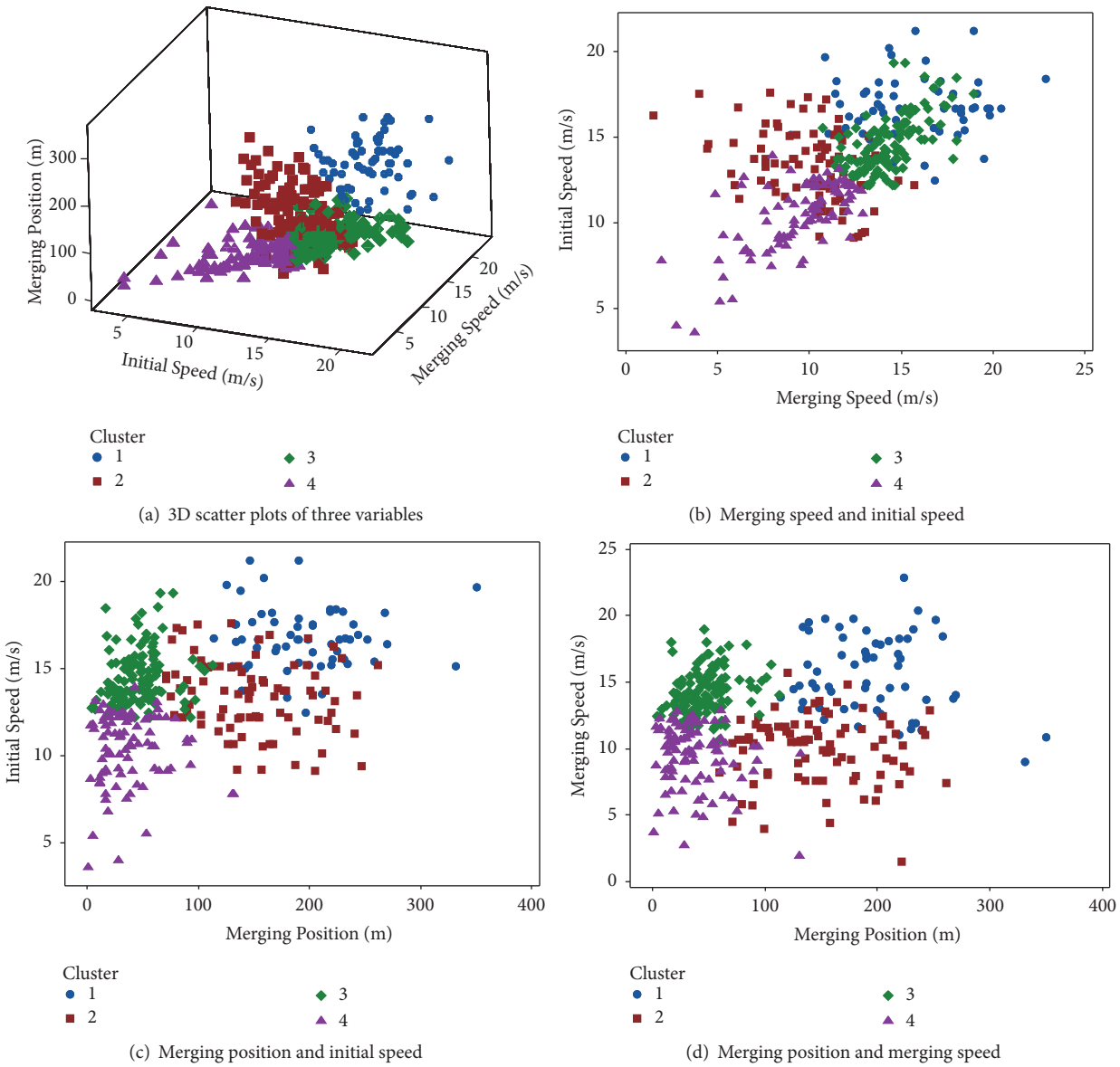


Figure 2: 3D scatter plots and 2D scatter plots of two-step clustering result.

Table 2: Mann–Whitney U test results of merging maneuvers.

Cluster	Initial Speed	Merging Speed	Merging Position
Clusters 1 & 2	0.000	0.000	0.000
Clusters 1 & 3	0.000	0.000	0.000
Clusters 1 & 4	0.000	0.000	0.000
Clusters 2 & 3	0.000	0.197	0.000
Clusters 2 & 4	0.000	0.000	0.000
Clusters 3 & 4	0.000	0.000	0.000

different merging drivers and identify the heterogeneity among merging drivers based on the merging maneuver parameters. Clusters 2 and 3 have similar merging speed probably because the traffic condition are similar. However the differences in initial speed and merging position indicates that different drivers do behave differently under the same traffic condition.

5.3. Probability Distribution Fitting Based on Clustering Results. To best investigate the heterogeneity in merging behaviors, the clustering results are used to develop the mixture models of merging maneuvers. 7 distributions (Normal, Log-normal, Student's t, Logistic, Log-Logistic, Gamma, and Weibull) are considered for each cluster and the Kolmogorov-Smirnov test is used to select the best fitted distributions. Table 3 shows best fitted distributions and corresponding K-S Test statistics, critical values, and the corresponding p values. The probability distributions of each merge maneuver parameters for different traffic densities are provided in Figure 2.

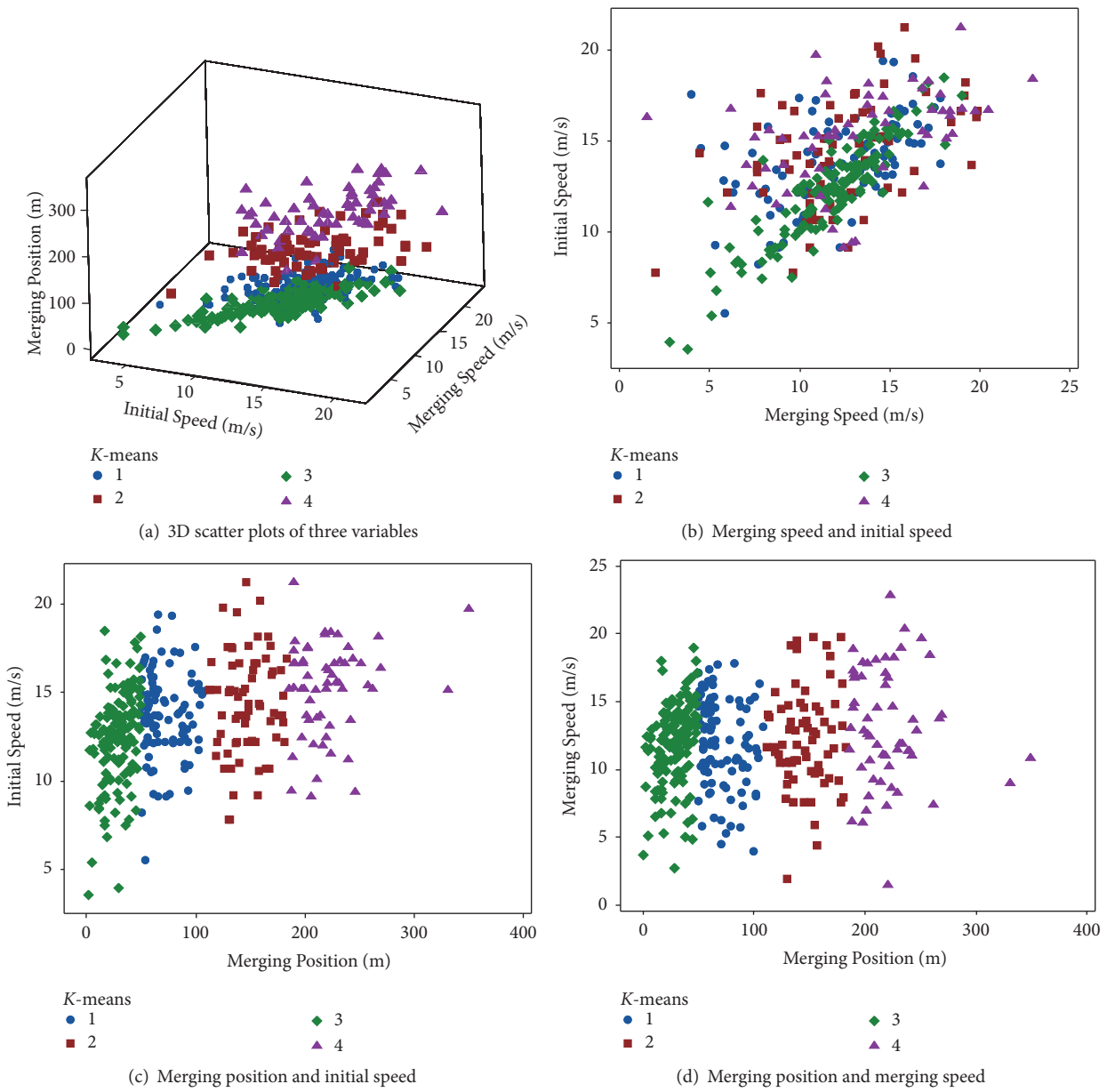
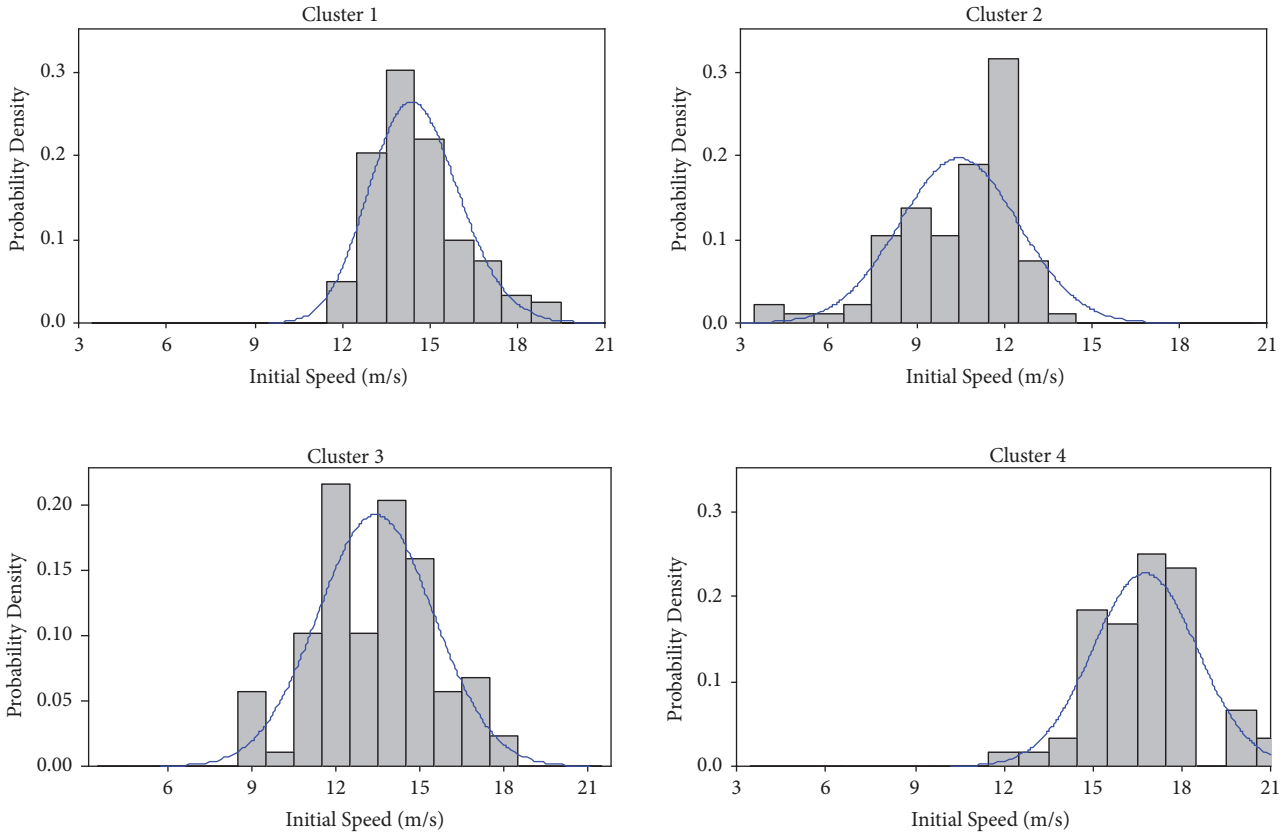


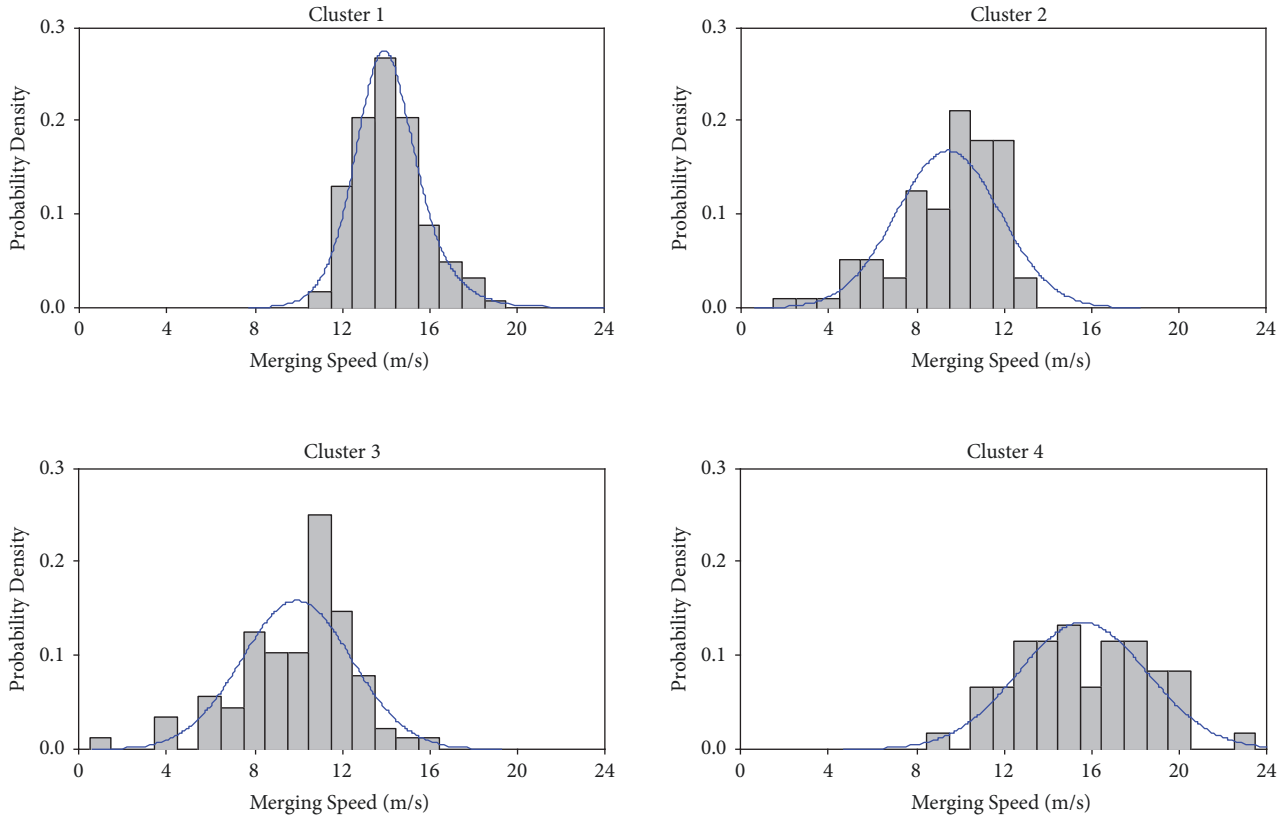
Figure 3: 3D scatter plots and 2D scatter plots of K-means clustering result.

One can observe from Table 3 and Figure 4 that the best fitted distributions of initial speed and merging speed are all Normal distributions except cluster 1. The best fitted distributions of initial speed and merging speed for cluster 1 are Gamma and Log-Logistic, respectively. The best distributions of merging position for 4 clusters are Gamma, Gamma, Log-normal, and Normal, respectively. The different distribution functions and different parameters and shapes of the fitted distributions show significant heterogeneities among the merging drivers across different clusters.

5.4. Analysis of Merging Durations and Gap Acceptance Based on Clustering Results. To further investigate the heterogeneity in different clusters, a comprehensive analysis of the merging durations and gap acceptance is given in this section. Figure 3 provides the cumulative curves of merging durations and gap acceptance parameters. The basic statics are provided in Table 4. One can observe from Table 4 and Figure 5 that cluster 2 has the largest mean value of merging duration and cluster 1 has the second largest value. The merging duration seems to be similar for clusters 3 and 4, both of which are smaller than that of clusters 1 and 2. One can also observe that the accepted gaps, lead gaps, and lag gaps have similar trends in statics and shapes. Clusters 1 and 4 have bigger mean values of three parameters than clusters 2 and 3, indicating that the traffic conditions of clusters 1 and 4 are better than those of clusters 2 and 3. Besides, the mean and median values of accepted lag gaps are much larger than accepted lead gaps in clusters 2 and 4. However they are similar in clusters 1 and 4.

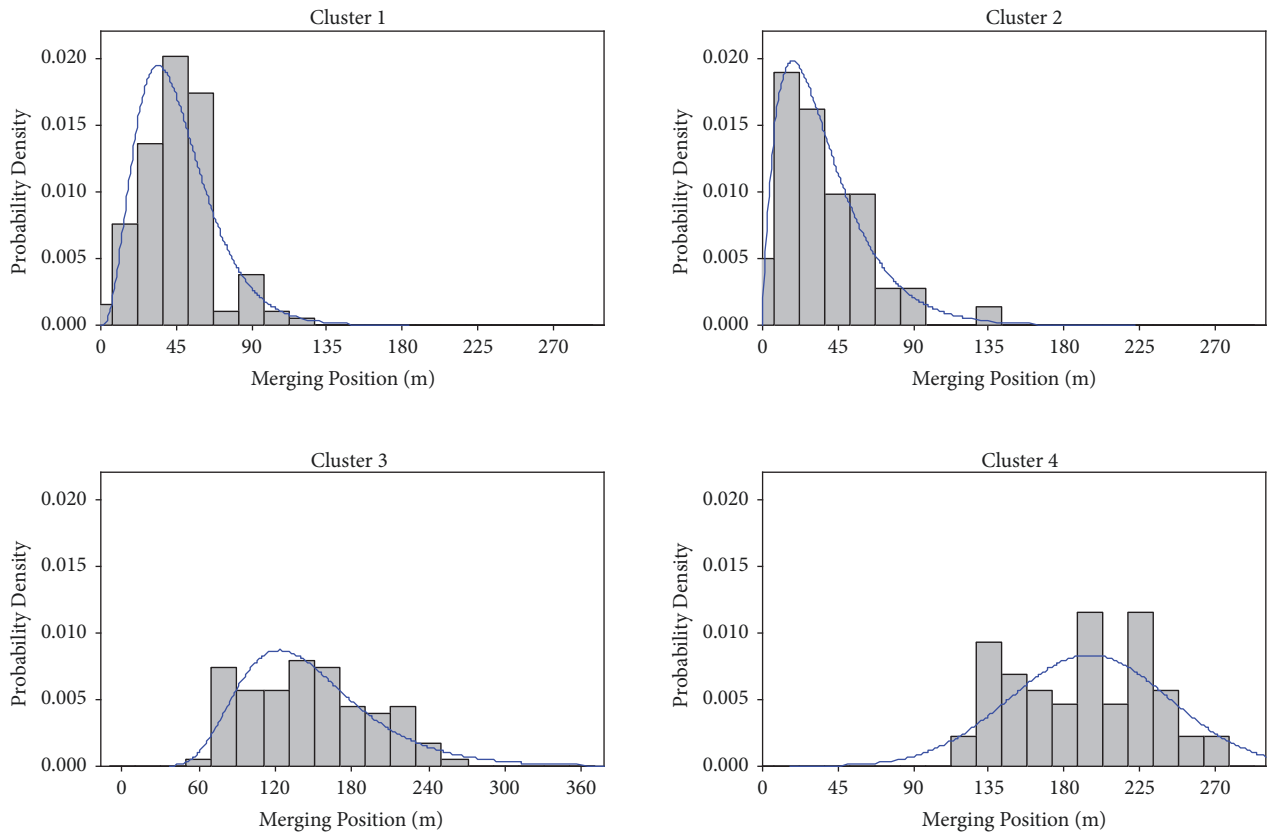


(a)



(b)

Figure 4: Continued.



(c)

Figure 4: Best fitted distributions for (a) initial speed, (b) merging speed, and (c) merging position for different clusters.

Table 3: Distribution fitting results of merging maneuver parameters for different clusters.

Merge Maneuver	Cluster	Best Fitted Distribution	Parameters		Goodness of Fit, K-S test	
			Location/Shape	Scale	K-S Static	p value
Initial Speed	1	Gamma	86.27	5.93	0.0755	0.463
	2	Normal	10.481	2.0346	0.108	0.202
	3	Normal	13.43	2.069	0.0647	0.831
	4	Normal	16.77	1.76	0.102	0.532
Merging Speed	1	Log-Logistic	3.66	0.474	0.0372	0.993
	2	Normal	3.18	0. 445	0.124	0.0996
	3	Normal	9.95	2.51	0.122	0.134
	4	Normal	15.61	2.94	0.109	0.438
Merging Position	1	Gamma	4.56	0.010	0.0671	0.614
	2	Gamma	2.04	0.055	0.05398	0.931
	3	Lognormal	4. 94	0.349	0.0660	0.814
	4	Normal	194.89	47.89	0.0613	0.968

To statistically investigate if there are significant differences in the median values of merging durations and gap acceptance between different clusters, two-sample Mann–Whitney U Tests are carried out. The obtained p values are given in Table 5 and the significance level is set as 0.1

From Table 5, one can find that the median values of merging durations of clusters 1 and 2 are significantly different from those of clusters 3 and 4. It indicates that the merging durations of drivers that merge late (cluster 3 and cluster 4) are significantly shorter than those of drivers merge early (cluster 1 and cluster 2). Previous study has claimed that the merging durations are decreasing with the increase of speed [1]. Although the merging speeds of cluster 2 and cluster 4 are much higher than those of cluster 1 and cluster 4, there are no significant differences between clusters 1 and 2 and between clusters 3 and 4. The Early Merging Drivers might

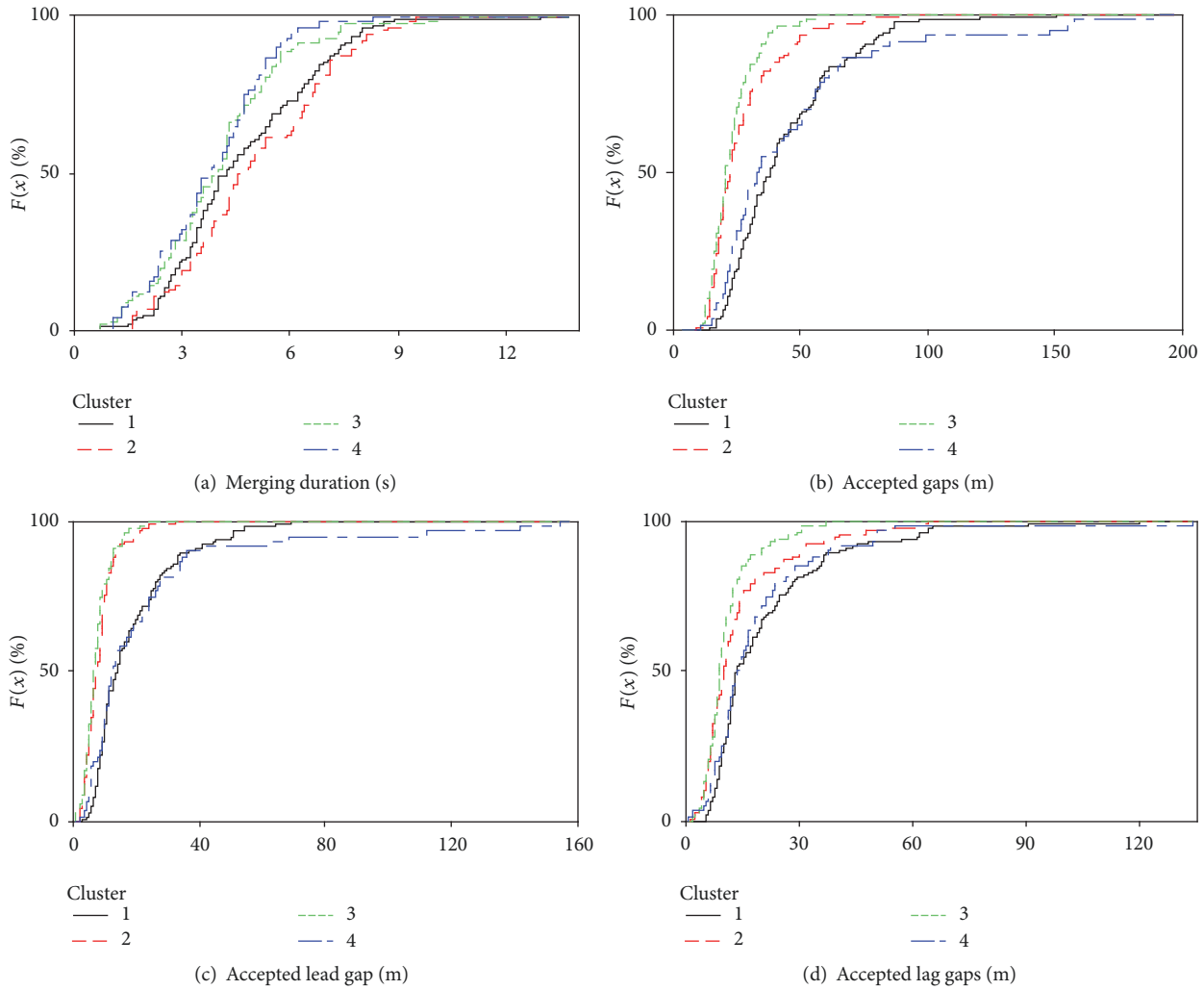


Figure 5: Cumulative distributions of (a) merging position (m), (b) merging speed (m/s), and (c) merging duration (s) for different clusters.

Table 4: Basic statistics of merging duration and gap acceptance for different clusters.

Parameters	Cluster	Mean	Standard Deviation	Median	Skewness
Merging Duration	1	4.67	2.00	4.2	0.78
	2	5.01	2.02	4.6	0.21
	3	4.00	1.80	3.95	0.81
	4	3.80	1.54	3.75	0.27
Accepted Gaps	1	43.63	22.64	38.15	1.56
	2	26.90	14.56	22.31	1.99
	3	22.85	9.07	20.53	1.42
	4	46.66	36.93	22.88	2.23
Accepted Lead Gaps	1	18.48	13.29	14.12	1.60
	2	8.376	5.09	6.992	1.95
	3	7.463	4.30	6.643	1.22
	4	22.95	29.51	12.63	3.22
Accepted Lag Gaps	1	20.77	17.66	13.25	2.63
	2	14.06	12.38	10.12	2.22
	3	10.70	6.60	8.856	1.87
	4	19.40	19.33	13.74	3.87

Table 5: Mann–Whitney U test results between different clusters.

Cluster	Merging Durations	Accepted Gaps	Accepted Lead Gaps	Accepted Lag Gaps
Clusters 1 & 2	0.297	0.000	0.000	0.000
Clusters 1 & 3	0.075	0.000	0.000	0.000
Clusters 1 & 4	0.043	0.399	0.459	0.926
Clusters 2 & 3	0.002	0.450	0.280	0.238
Clusters 2 & 4	0.001	0.000	0.000	0.002
Clusters 3 & 4	0.999	0.000	0.000	0.000

Table 6: Mann–Whitney U test results between accepted lead gaps and lag gaps.

Cluster	Accepted Lead Gaps versus Accepted Lag Gaps
Cluster 1	0.3789
Cluster 2	0.000
Cluster 3	0.000
Cluster 4	0.47588

be better prepared for merging before they enter the auxiliary lane, become more patient, and use more time to finish the merging maneuvers, while the long travelling distance might urge the Late Merging Drivers to finish the merging maneuvers as soon as possible even at low speeds.

The hypothesis tests of accepted gaps, accepted lead, and accepted lag gaps have the same results at 90% level. There are no significant differences in the gap acceptance parameters between clusters 1 and 4 and between clusters 2 and 3. However, the accepted gaps and accepted lead and lag gaps of clusters 1 and 4 are significantly larger than those of clusters 2 and 3, indicating that the traffic conditions of clusters 2 and 3 are more congested than those of clusters 1 and 4.

To check if the accepted lag gaps are larger than accepted lead gaps, a Mann–Whitney U Test between accepted lead gaps and lag gaps in different clusters is also carried out. Table 6 shows the obtained p values. One can find that the median values of accepted lag gaps are significantly larger than accepted lead gaps in clusters 2 and 3 and there are no significant differences between the median values of accepted lead gaps and lag gaps in clusters 1 and 4. It means merging drivers are more sensitive to lag gaps than lead gaps under congestion.

The data of merging durations and gap acceptance parameters for different clusters are also used to find the best corresponding fitting distributions using EasyFit 5.4 [48]. 7 distributions (Normal, Log-normal, Student's t, Logistic, Log-Logistic, Gamma, and Weibull) are considered for each sample and the Kolmogorov-Smirnov test is used to select the best fitted distribution. Table 7 shows best fitted distributions and corresponding K-S Test statistics, critical values, and the corresponding p values. The probability distributions of each parameters for different clusters are provided in Figure 6.

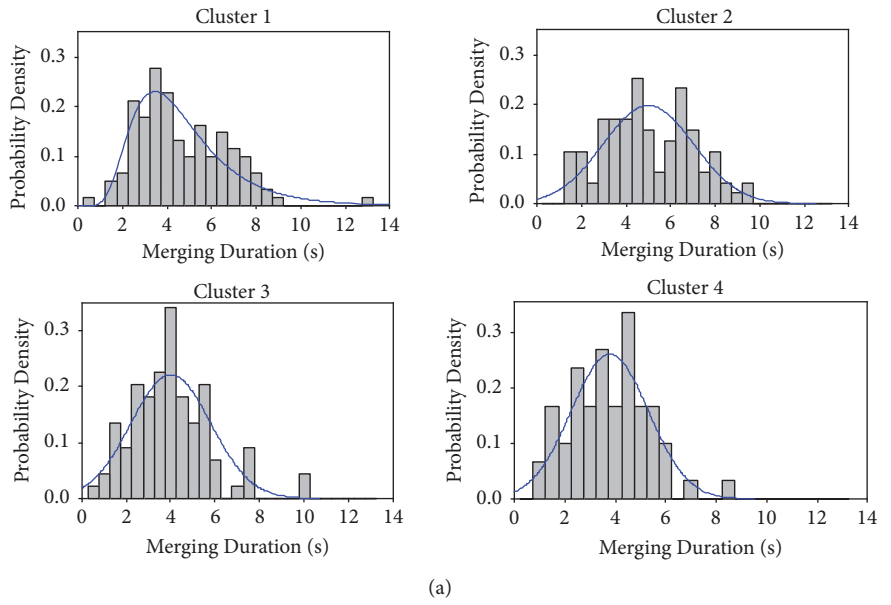
One can observe from Table 7 that the best fitted distribution of merging durations for are all Normal, except for the Log-normal distribution for cluster 1. One can also find that the best fitted distributions of gap acceptance parameters are all Log-normal, except for the Gamma distribution for accepted lead gap for cluster 3. There are considerable differences in the parameters and shape of fitted distribution between different classes, indicating the heterogeneity between different classes.

6. Concluding Remarks

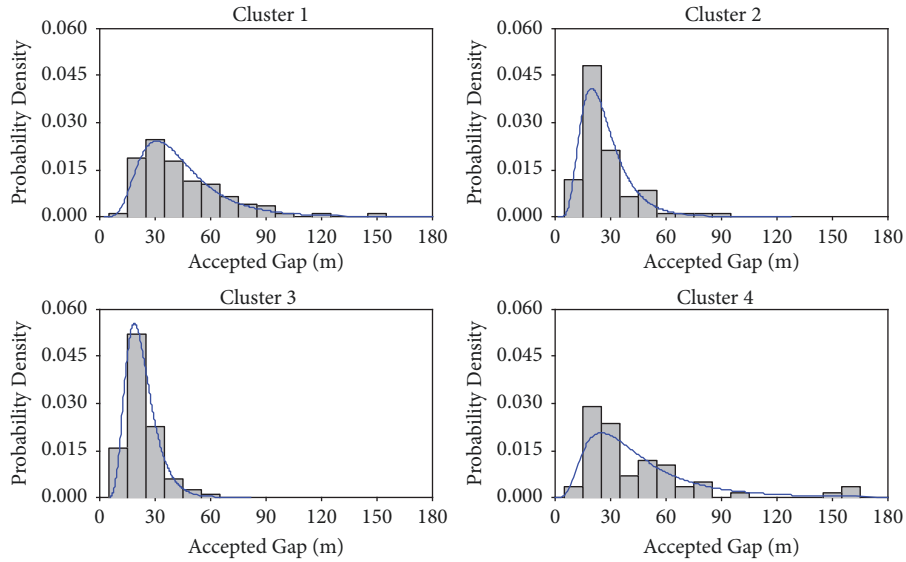
To investigate the heterogeneity in merging drivers, a framework that combines two-step cluster analysis, hypothesis, and probability distribution fitting is proposed in this paper. The Tto-step cluster analysis is applied to three merging maneuver parameters (namely, initial speed, merging speed, and merging position) to identify different clusters of merging drivers. Four clusters are automatically and optimally chosen as the best clustering results. Compared with K-means clustering approach, the two-step cluster analysis can not only produce the best clustering results automatically, but also better separate the data.

Mann–Whitney U Tests are carried out and prove that merging maneuvers are indeed different between different clusters. The identified four clusters are called Early Merging Drivers at High Speed, Early Merging Drivers at Low Speed, Late Merging Drivers at Low Speed, and Late Merging Drivers at High Speed. The clustered data are used to find the best corresponding fitting distributions using EasyFit 5.4 and 7 distributions (Normal, Log-normal, Student's t, Logistic, Log-Logistic, Gamma, and Weibull) are considered for each cluster. Kolmogorov-Smirnov test statics are used to select the best fitted distributions.

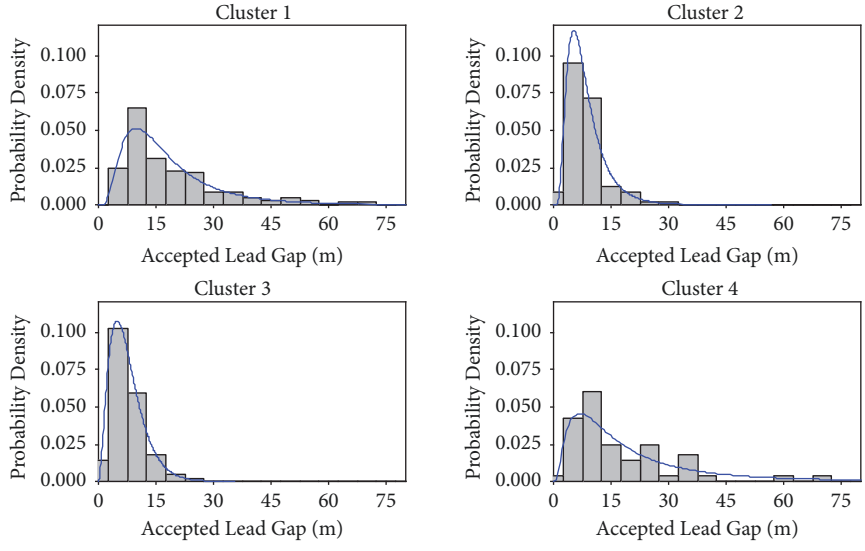
Merging durations and gap acceptance are analyzed based on the clustering results. It is discovered that merging durations of Late Merging Drivers are significantly shorter than Early Merging Drivers. However, there are no significant differences between Early Merging Drivers at High Speed and Low Speed and between Late Merging Drivers at High Speed and Low Speed. It indicates that drivers are losing patience and getting more aggressive with the increase of time on the auxiliary lane under either congestion or uncongested traffic condition. There are no significant differences in the gap acceptance parameters between clusters 1 and 4 and between clusters 2 and 3. However, the accepted gaps and accepted lead and lag gaps of clusters 1 and 4 are significantly larger than those of clusters 2 and 3, indicating that the traffic conditions of clusters 2 and 3 are more congested than those of clusters 1 and 4. The hypothesis tests between accepted lead gaps and accepted lags show that merging drivers the



(a)

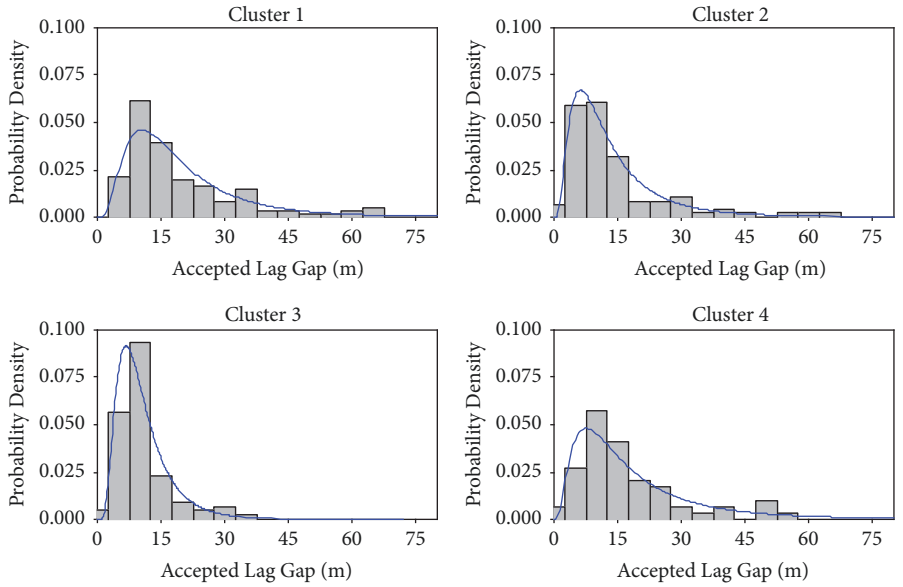


(b)



(c)

Figure 6: Continued.



(d)

Figure 6: Best fitted distributions for (a) merging duration, (b) accepted gaps, (c) accepted lead gaps, and (d) accepted lag gaps for different clusters.

Table 7: Results of the best fitted distributions for merging durations and gap acceptance parameters.

Parameters	Cluster	Best Fitted Distribution	Parameters		Goodness of Fit, K-S test	
			Scale	Shape	K-S Static	p value
Merging Durations	1	Lognormal	4.20	0.852	0.0671	0.612
	2	Normal	5.01	2.02	0.0944	0.344
	3	Normal	4.00	1.80	0.0920	0.421
	4	Normal	3.79	1.54	0.0670	0.934
Accepted Gaps	1	Lognormal	3.66	0.474	0.0676	0.603
	2	Lognormal	3.1823	0. 445	0.0891	0.413
	3	Lognormal	3.06	0.358	0.0618	0.868
	4	Lognormal	3.63	0.625	0.104	0.495
Accepted Lead Gaps	1	Lognormal	2.70	0.644	0.0907	0.248
	2	Lognormal	1.97	0.553	0.0568	0.9021
	3	Gamma	3.01	2.48	0.06678	0.803
	4	Lognormal	2.69	0.873	0.0914	0.664
Accepted Lag Gaps	1	Lognormal	2.79	0.660	0.0978	0.178
	2	Lognormal	2.36	0.728	0.07918	0.564
	3	Lognormal	2.22	0.551	0.0813	0.577
	4	Lognormal	2.66	0.789	0.0920	0.656

accepted lead gaps are significantly smaller than accepted lag gaps under congestion. It indicates that the merging drivers are more sensitive to the lag gap under congestion during merging process. The merging durations and gap acceptance parameters in different clusters are also used to find the best corresponding fitting distributions using EasyFit 5.4.

The proposed method can automatically identify the heterogeneity in merging drivers and the results obtained in this paper can be used to enhance the accuracy of the merge behavior models in microscopic simulation software.

In future research, more data will be collected to investigate the factors that may influence driver heterogeneity, such as the surrounding population and the time of data acquisition. On the other hand, driver heterogeneity will be incorporated into microscopic traffic simulation models to improve the accuracy and authenticity.

Conflicts of Interest

The authors declare that they have no conflicts of interest.

References

[1] Q. Wang, 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, pp. 246–253, 2014.

[2] W. Daamen, S. P. Hoogendoorn, and M. Loot, "Empirical analysis of merging behavior at freeway on-ramp," *Transportation Research Records: Journal of Transportation Research Board*, no. 2188, pp. 108–118, 2010.

[3] Z. Zheng, "Recent developments and research needs in modeling lane changing," *Transportation Research Part B: Methodological*, vol. 60, pp. 16–32, 2014.

[4] M. K. Ardakani, J. Yang, and L. Sun, "Stimulus response driving behavior: An improved general motor vehicle-following model," *Advances in Transportation Studies*, no. 39, pp. 23–36, 2016.

[5] L. Elefteriadou, R. P. Roess, and W. R. McShane, "Probabilistic nature of breakdown at freeway merge junctions," *Transportation Research Record: Journal of the Transportation Research Board*, vol. 1484, pp. 80–89, 1995.

[6] H. Yi and T. E. Mulinazzi, "Urban freeway on-ramp invasive influences on 3 mainline operations," in *Proceedings of the 86th Annual Meeting of the Transportation Research Board*, Washington, DC, USA, 2007.

[7] S. Ossen, S. P. Hoogendoorn, and B. G. H. Gorte, "Interdriver differences in car-following a vehicle trajectory-based study," *Transportation Research Record*, vol. 1965, no. 1, pp. 121–129, 2006.

[8] S. Ossen and S. P. Hoogendoorn, "Car-following behavior analysis from microscopic trajectory data," *Transportation Research Record: Journal of the Transportation Research Board*, vol. 1934, pp. 13–21, 2005.

[9] S. Ossen and S. P. Hoogendoorn, "Heterogeneity in car-following behavior: theory and empirics," *Transportation Research Part C: Emerging Technologies*, vol. 19, no. 2, pp. 182–195, 2011.

[10] J. W. Kim and H. S. Mahmassani, "Correlated parameters in driving behavior models: car-following example and implications for traffic microsimulation," in *Proceedings of the 90th Annual Meeting of the Transportation Research Board*, Washington, DC, 2011.

[11] L. Sun, Y. Pan, and W. Gu, "Data mining using regularized adaptive B-splines regression with penalization for multi-regime traffic stream models," *Journal of Advanced Transportation*, vol. 48, no. 7, pp. 876–890, 2014.

[12] L. Sun, "Spectral and time-frequency analyses of freeway traffic flow," *Journal of Advanced Transportation*, vol. 48, no. 7, pp. 821–857, 2014.

[13] L. Sun, "Stochastic projection-factoring method based on piecewise stationary renewal processes for mid- and long-term traffic flow modeling and forecasting," *Transportation Science*, vol. 50, no. 3, pp. 998–1015, 2016.

[14] Y. Pan and L. Sun, "Characterizing heterogeneity in vehicular traffic speed using two-step cluster analysis," *Journal of Southeast University*, vol. 28, no. 4, pp. 480–484, 2012.

[15] W. Xiong, L. Sun, and J. Zhou, "Spline-based multi-regime traffic stream models," *Journal of Southeast University*, vol. 26, no. 1, pp. 122–125, 2010.

[16] Cambridge Systematics Inc., *NGSIM U.S. 101 Data Analysis Summery Report*, Federal Highway Administration, Washington, DC, USA, 2005.

[17] Q. Yang and H. N. Koutsopoulos, "A microscopic traffic simulator for evaluation of dynamic traffic management systems," *Transportation Research Part C: Emerging Technologies*, vol. 4, no. 3, pp. 113–129, 1996.

[18] K. I. Ahmed, *Modeling Drivers Acceleration and Lane Changing Behavior [Ph.D. thesis]*, Department of Civil and Environmental Engineering, MIT, 1999.

[19] T. Toledo and D. Zohar, "Modeling duration of lane changes," *Transportation Research Record: Journal of the Transportation Research Board*, no. 1999, pp. 71–78, 2007.

[20] T. Toledo, H. N. Koutsopoulos, and M. Ben-Akiva, "Estimation of an integrated driving behavior model," *Transportation Research Part C: Emerging Technologies*, vol. 17, no. 4, pp. 365–380, 2009.

[21] M. A. Ahammed, Y. Hassan, and T. A. Sayed, "Modeling driver behavior and safety on freeway merging areas," *Journal of Transportation Engineering*, vol. 134, no. 9, pp. 370–377, 2008.

[22] R. Herman and G. H. Weiss, "Comments on the highway-crossing problem," *Operation Research*, vol. 9, pp. 828–840, 1961.

[23] D. R. Drew, L. R. Lamotte, J. A. Wattleworth, and J. H. Buhr, "Gap acceptance in the freeway merging process," *Highway Research Record*, vol. 208, 1967.

[24] A. J. Miller, "Nine Estimators of Gap Acceptance Parameters," in *Proceedings of the 5th International Symposium on the Theory of Traffic Flow and Transportation*, 1972.

[25] P. G. Gipps, "A model for the structure of lane-changing decisions," *Transportation Research Part B: Methodological*, vol. 20, no. 5, pp. 403–414, 1986.

[26] P. Hidas, "Modelling lane changing and merging in microscopic traffic simulation," *Transportation Research Part C: Emerging Technologies*, vol. 10, no. 5-6, pp. 351–371, 2002.

[27] J. Wang, "A simulation model for motorway merging behavior," *Transportation and Traffic Theory*, vol. 16, pp. 281–301, 2005.

[28] L. Bloomberg and J. Dale, "Comparison of VISSIM and CORSIM traffic simulation models on a congested network," *Transportation Research Record*, no. 1727, pp. 52–60, 2000.

[29] PTV, User Manual: VISSIM 4.0, 2004.

[30] SIAS., *S-Paramics 2005 Reference Manual*, SIAS Ltd, Edinburgh, UK, 2005.

[31] V. Punzo, D. J. Forminaso, and V. Torrieri, "Nonstationary kalman filter for estimation of accurate and consistent car-following data," in *Proceedings of the 84th Annual Meeting of the Transportation Research Board*, Washington, DC, USA, 2005.

[32] F. Marczak, W. Daamen, and C. Buisson, "Merging behaviour: Empirical comparison between two sites and new theory development," *Transportation Research Part C: Emerging Technologies*, vol. 36, pp. 530–546, 2013.

[33] T. D. Chu, T. Miwa, and T. Morikawa, "An analysis of merging maneuvers at urban expressway merging sections," *Procedia-Social and Behavioral Sciences*, vol. 138, pp. 105–115, 2014.

[34] J. Weng and Q. Meng, "Modeling speed-flow relationship and merging behavior in work zone merging areas," *Transportation Research Part C: Emerging Technologies*, vol. 19, no. 6, pp. 985–996, 2011.

[35] M. Sarvi and M. Kuwahara, "Microsimulation of freeway ramp merging processes under congested traffic conditions," *IEEE Transactions on Intelligent Transportation Systems*, vol. 8, no. 3, pp. 470–479, 2007.

[36] L. Sun, W. Gu, and H. Mahmassani, "Estimation of expected travel time using the method of moment," *Canadian Journal of Civil Engineering*, vol. 38, no. 2, pp. 154–165, 2010.

[37] L. Sun, J. Yang, H. Mahmassani, W. Gu, and B.-J. Kim, "Data mining-based adaptive regression for developing equilibrium speed-density relationships," *Canadian Journal of Civil Engineering*, vol. 37, no. 3, pp. 389–400, 2010.

[38] L. Sun, H. Zhang, R. Gao, W. Gu, B. Xu, and L. Chen, "Gaussian mixture models for clustering and classifying traffic flow in real-time for traffic operation and management," *Journal of Southeast University*, vol. 27, no. 2, pp. 174–179, 2011.

[39] L. Sun and J. Zhou, "Development of multiregime speed-density relationships by cluster analysis," *Transportation Research Record: Journal of the Transportation Research Board*, no. 1934, pp. 64–71, 2005.

[40] C. Hill, L. Elefteriadou, and A. Kondyli, "Exploratory analysis of lane changing on freeways based on driver behavior," *Journal of Transportation Engineering*, vol. 41, article 04014090, no. 4, 2014.

[41] M. J. Norusis, *IBM SPSS Statistics 19 Procedures Companion*, Addison-Wesley, Reading, Mass, USA, 2011.

[42] M. J. Norusis, *SPSS 15.0 Advanced Statistical Procedures Companion*, Prentice Hall, Chicago, Ill, USA, 2007.

[43] V. Alexiadis, J. Colyar, J. Halkias, R. Hranac, and G. McHale, "The next generation simulation program," *Institute of Transportation Engineers (ITE) Journal*, vol. 74, no. 8, pp. 22–26, 2004.

[44] C. Thiemann, M. Treiber, and A. Kesting, "Estimating acceleration and lane-changing dynamics from next generation simulation trajectory data," *Transportation Research Record: Journal of the Transportation Research Board*, vol. 2088, pp. 90–101, 2008.

[45] IBM., "Two-step cluster analysis," https://www.ibm.com/support/knowledgecenter/SSLVMB_20.0.0/com.ibm.spss.statistics.help/alg_twostep.htm, 2016.

[46] W. C. Navidi, *Statistics for Engineers and Scientists*, vol. 2, McGraw-Hill, New York, NY, USA, 2006.

[47] S. Gurupackiam and S. L. Jones, "Empirical study of lane changing in urban streets under varying traffic conditions," *Procedia-Social and Behavioral Sciences*, vol. 16, pp. 259–269, 2011.

[48] K. Schittkowski, "EASY-FIT: A software system for data fitting in dynamical systems," *Structural and Multidisciplinary Optimization*, vol. 23, no. 2, pp. 153–169, 2002.

[49] E. Balal, R. L. Cheu, T. Gyan-Sarkodie, and J. Miramontes, "Analysis of discretionary lane changing parameters on freeways," *International Journal of Transportation Science & Technology*, vol. 3, no. 3, pp. 277–296, 2014.