Hindawi
Journal of Advanced Transportation
Volume 2018, Article ID 1436521, 9 pages
https://doi.org/10.1155/2018/1436521

Research Article

Application of Finite Mixture of Logistic Regression for Heterogeneous Merging Behavior Analysis

Gen Li

School of Transportation, Southeast University, Nanjing 210096, China

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

Received 4 July 2018; Revised 15 October 2018; Accepted 4 November 2018; Published 21 November 2018

Guest Editor: Lele Zhang

A finite mixture of logistic regression model (FMLR) was applied to analyze the heterogeneity within the merging driver population. This model can automatically provide useful hidden information about the characteristics of the driver population. EM algorithm and Newton-Raphson algorithm were used to estimate the parameters. To accomplish the objective of this study, the FMLR model was applied to a trajectory dataset extracted from the NGSIM dataset and a 2-component FMLR model was identified. The important findings can be summarized as follows: The studied drivers can be classified into two components. One is called Risk-Rejecting Drivers. These drivers are consistent with previous studies and primarily merge in as soon as possible and have a distinct preference for the large gaps. The other is the Risk-Taking Drivers that are much less sensitive to the gap size and pay more attention to surrounding traffic conditions such as the speed of front vehicle in the auxiliary lane and lead space gap between the merging vehicle and its leading vehicles in the auxiliary lane. Risk-Taking Drivers use the auxiliary lane to get to the further downstream or less congested area of the main lane. The proposed model can also produce more precise predicting accuracy than logistic regression model.

1. Introduction

Congestion has become one of the most serious economic and social problems and has drawn great attention from the public, transportation research scientists, transportation managers, and so on. Understanding the causes and mechanism of traffic congestion can help traffic managers formulate targeted policies to make better use of the existing transportation infrastructures.

Merging areas are the bottleneck of freeway. Merging behavior is one of the typical mandatory lane changes when vehicles have to move from an on-ramp to the main road. It has been claimed in some studies that merging behavior at merging areas affects traffic operations and may trigger traffic congestions and breakdowns [1, 2]. Thus it is important to analyze the merging behaviors to help understand the mechanism of traffic jams to some extent from a microscopic viewpoint and build more accurate traffic simulation models.

Recently, driver heterogeneity has drawn great attention in microscopic traffic flow studies. Several studies investigated the driver heterogeneity during car-following process [3–6]. Accommodating heterogeneity within the driver population is important in building microscopic traffic models. To investigate the heterogeneity in merging behaviors, a finite mixture of logistic regression (FMLR) model was proposed in this paper. This model can incorporate the unobserved heterogeneity and automatically segments the merging drivers into different homogeneous populations. More specifically, this paper aims to achieve the following objectives:

(i) Prove the existence of heterogeneity among merging drivers.

(ii) Identify different driving styles and attitudes during merging process.

(iii) Model the merging behavior more accurately.

The present study is organized as follows. The next section will provide a critical review on the existing relevant literature followed by Section 3, which describes the NGSIM data used in this paper. Section 4 gives the methodology to build FMLR model. Results and discussions are presented in Section 5.

Finally, the conclusions and future work are presented in Section 6.

2. Literature Review

Several methods have been adopted to model merging behavior, among which gap acceptance theory was the most widely used method [8–13]. The most important assumption in gap acceptance theory was 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. The critical gap is determined by the characteristics of the drivers, traffic conditions, and so on [14]. Gap acceptance models were initially built to estimate the capacity of unsignalized intersections. Different distributions of critical gaps were assumed in various studies [15–17]. Gipps [18] first used the gap acceptance theory to propose a comprehensive framework of lane-changing model. Gipps's framework has been widely used in several merge models [19, 20] and microscopic traffic simulation software [21, 22]. Different definitions of critical gap were used in these models and software.

Gap acceptance theory was often criticized as its basic assumption is often inconsistent with the real world observation because some lane change behaviors occurred when only the lead or lag gap or even none of them are larger than the critical gap [14, 23, 24]. To overcome this deficiency, discrete choice models such as binary logit model were used by some researchers [14, 25–27]. Built upon a series of studies [9, 10, 28], a framework for merging behavior with latent plans was introduced by Choudhury *et al.* [29]. Normal merge, merge with courtesy, and forced merge were considered in this framework. However, Marczak *et al.* [14] pointed out that in this framework only accepted gaps were considered and rejected gaps were ignored; and some of the estimated coefficients in the model were not significant.

Traffic behaviors are always uncertain and variable and heterogeneity cannot be ignored in traffic studies. Some studies investigated the heterogeneity among the macroscopic traffic flow [30, 31]. Others studied the heterogeneity in car following behaviors from microscopic viewpoint by deriving the joint distribution of model coefficients depending on an empirical basis [4, 5, 32–34]. However, only a few studies were found to investigate driver heterogeneity in lane changing models. A two-step clustering analysis was proposed by Li and Sun [35] to analyze heterogeneity of the merging maneuvers. However, this study ignored the heterogeneity during gap selection and decision process. An empirical analysis conducted by Daamen *et al.* [23] showed that different merging strategies might be adopted by different drivers under different traffic conditions. It has been pointed out by Keyvan-Ekbatani *et al.* [36] that different strategies might be used during gap selection process; however the sample size was too small to perform statistically relevant tests and build merging model.

Thus, a FMLR model was introduced in this paper to model the gap selection behaviors during merging process and investigate the heterogeneity among merging drivers. The FMLR model takes the advantage of two techniques: clustering and regression analysis. The model naturally incorporates the unobserved heterogeneity into logistic regression model and automatically segments the drivers into different homogeneous populations. The proposed FMLR model can explain the different strategies in merging behaviors.

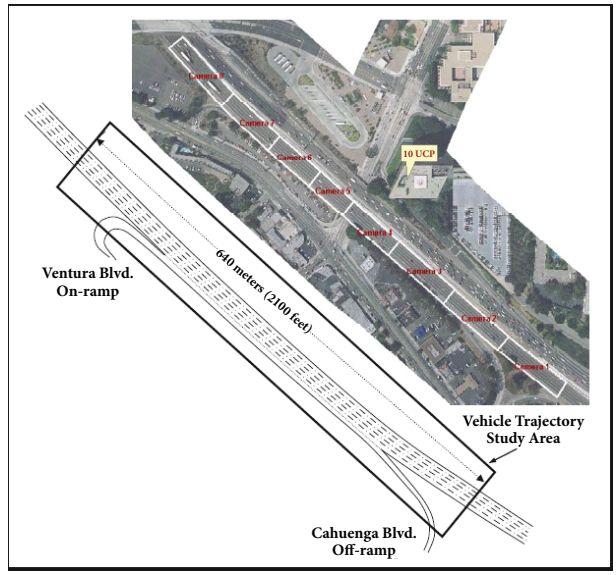

Figure 1: The section of US 101 [7].

3. Data Preparation

The NGISM dataset has been widely used for traffic flow and traffic simulation studies and proved to have high accuracy. Thus, in this paper, the vehicle trajectory data in NGSIM dataset collected on a segment of southbound U.S. Highway 101 (Hollywood Freeway) in Los Angeles, CA, are chosen [37]. Figure 1 shows the site for U.S. 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 [7]. The dataset has three data subsets, all of which were collected in 15 minutes.

In this study, we focus on the behavior of merging vehicles and only trajectory data in the weaving section were used. However, it has been pointed out that the original trajectory data contain some noise and errors, which are caused by the system errors and tracking errors [38–41]. Several methods have been proposed to filter the data [38–40] or re-extract the trajectory data [41]. Re-extracting can produce the most accurate data especially the acceleration data, which however would also make too much effort. In this paper, a smoothing method called sEMA developed by Thiemann *et al.* [38] is applied to reduce the noise and errors. The sEMA method is also adopted in other studies of merging behaviors and has been proved to be able to provide enough precision for lane change studies [42–44]. This data smoothing technique was applied as follows:

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

Table 1: Examples with the same global coordinates in the first and second subsets.

Data Point	Sub dataset 1				Sub dataset 2			
	Vehicle ID	Frame ID	Local x	Local y	Vehicle ID	Frame ID	Local x	Local y
1	33	424	54.612	1397.746	36	847	54.612	1438.019
2	33	429	54.687	1420.332	1070	4878	54.687	1460.518
2	63	290	67.936	514.811	1472	5857	67.936	550.085

Table 2: Examples with the same global coordinates in the first and third subsets.

Data Point	Sub dataset 1				Sub dataset 3			
	Vehicle ID	Frame ID	Local x	Local y	Vehicle ID	Frame ID	Local x	Local y
1	42	446	53.395	1449.048	1721	8609	53.395	1483.814
2	63	483	41.056	1494.004	1280	6744	41.056	1528.773
3	296	967	53.8340	1389.247	905	4719	53.834	1424.013

(2) The locations (both local lateral and longitudinal coordinates), velocities, and accelerations of vehicles are smoothed by the symmetric exponential moving average filter (sEMA) proposed by Thiemann *et al.* [38] to decrease measurement errors in the data. The smoothing times of sEMA method are set as the suggested values for the U.S. Highway 101 dataset in Thiemann *et al.* [38].

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 datasets:

(1) Filter out the trajectories when there are no putative leading vehicles or putative following vehicles on the adjacent main lane. Such trajectories are recorded at the beginning or ending of the video tape and cannot provide the interactions of merging vehicles with their surrounding vehicles.

(2) Filter out the trajectories when putative leading or putative following vehicle of a merging vehicle runs around the lane boundary (it keeps touching the lane boundary before lane change or turns back the original lane in about 1 second). These trajectories are always caused by the tracking errors.

After filtering, a searching process was conducted to check the consistency of the local coordinates and global coordinates. Linear regression was performed between local coordinates and global coordinates for each subdataset. Three linear relationships were obtained for each subset:

$$Localy_1 = 0.3209 globalx_1 - 1.1326 globaly_1 \quad (1)$$

$$Localy_2 = 0.3291 globalx_2 - 1.1334 globaly_2 \quad (2)$$

$$Localy_3 = 0.3209 globalx_3 - 1.1333 globaly_3 \quad (3)$$

R^2 of three linear relationships are 0.9996, 0.9997, and 0.9997, respectively. It means that the local y of three subsets in US-101 datasets are inconsistent with each other. We cannot find simple linear relationship between local x and global x in US-101 dataset. This could be caused by the specific coordinate system used and the special geometric shape of the road sections. It also could be caused by measuring errors.

To further verify the inconsistency of the US-101 dataset, several data points that have the same global coordinates among the three subsets were searched and obtained. By checking the local coordinates (local x and local y), it was found that the three subsets of US-101 dataset are consistent in local x, but inconsistent in local y. Tables 1 and 2 show the examples having the same global coordinates in the first and second subsets and in the first and third subsets.

One can find that, for the points with the same global coordinates, the three subdatasets have the same local x, but different local y. In the local longitudinal coordinate, the upstream edge (0 m) in datasets 1 is at 12.275m in dataset 2 and 10.598 m in dataset 3. Thus, the three datasets must be unified by using the local coordinates of one of the three subsets.

At every instant when offered a new gap, a merging vehicle driver assesses traffic conditions to decide whether to accept the offered gap or not. One merging vehicle could only accept one gap but could reject several gaps. After data processing, trajectories of 374 merging vehicles consisting of 925 observations were extracted from the dataset. The explanatory variables that may affect a driver's merging decision used as candidates for analyzing the merging behavior model are shown in Table 3.

4. Methodology

4.1. Finite Mixture of Logistic Regression. The FMLR model is based on the idea that the observed data come from a population with several subpopulations or components [45, 46]. The overall population is modeled as a mixture of the groups using finite mixture models.

Let $\mathbf{X}$ and $\mathbf{Y}$ denote random vectors with N samples and each sample has M_n observations $(\mathbf{x}_i, \mathbf{y}_i)$ $(i = 1, \ldots M_n, n = 1, \ldots N)$. Here, the response vector $\mathbf{Y}$ has values in $\mathbb{R}^d$ and the explanatory vector $\mathbf{X}$ has values in $\mathbb{R}^p$. Then, a FMLR with K components has the form

$$h(\mathbf{y} \mid \mathbf{x}, \psi) = \sum_{k=1}^{K} \pi_k f(\mathbf{y} \mid \mathbf{x}, \boldsymbol{\theta}_k) \quad (4)$$

Table 3: Descriptions of the explanatory variables.

Variable	Descriptions
D_n^i(m)	The size of the i^{th} offered gap of merging vehicle n
V_n^i(m/s)	The speed of merging vehicle n at i^{th} offered gap.
Y_n^i(m)	The longitudinal position of the merging vehicle n to the start of the auxiliary lane.
ΔV_{nPL}^i(m/s)	The speed difference between the putative leading vehicle and the merging vehicle n at offered gap i.
ΔV_{nPF}^i(m/s)	The speed difference between the putative following vehicle and the merging vehicle n at offered gap i.
δ	Existence of a lead vehicle in the merge lane. If there is a lead vehicle in the merge lane, $\delta = 1$; otherwise, $\delta = 0$.
$\Delta D_{n\lg}^i$(m)	Lead gap of merging vehicle n in the auxiliary lane at offered gap i.
V_{nLead}^i (m)	The speed of the leading vehicle in the auxiliary lane at offered gap i.
ΔV_{nLead}^i(m/s)	The speed difference between the leading vehicle in the auxiliary lane and the merging vehicle n at offered gap i.

$$\sum_{k=1}^{K}\pi_k = 1, \quad \pi_k > 0 \tag{5}$$

where $h(\mathbf{y} \mid \mathbf{x}, \Psi)$ is the conditional density of $\mathbf{y}$ given $\mathbf{x}$ and $\boldsymbol{\theta}_k$, π_k is the mixing proportion, $\boldsymbol{\theta}_k$ is the component-specific parameter vector for the density function f, and $\psi = (\pi_1, \ldots, \pi_K, \boldsymbol{\theta}_1, \ldots \boldsymbol{\theta}_K)$ is the vector of all parameters.

Several finite mixture models can be extended based on (4) and (5). For multivariate normal f and $\mathbf{x} \equiv \mathbf{1}$ we get a finite mixture of Gaussians without a regression part, also known as model-based clustering. If f is a univariate normal density with component-specific mean $\boldsymbol{\beta}_k'\mathbf{x}$ and variance $\boldsymbol{\sigma}_k^2$, we have $\boldsymbol{\theta}_k = (\boldsymbol{\beta}_k', \boldsymbol{\sigma}_k^2)$, and (4) describes a finite mixture of linear regression, also called latent class linear regression model or cluster-wise regression [47]. If f is a member of the exponential family, we get a FMLR models [48, 49].

The analyst does not observe directly which component, $k = 1, ..., K$, generated observation $\mathbf{y}_i$. The model assumes that individuals are distributed heterogeneously with a discrete distribution within the population. In order to impose the constraints in (2), the mixing proportions are parameterized with a multinomial logit form [50, 51]:

$$\pi_k = \frac{\exp(\alpha_k)}{\sum_{k=1}^{K}\exp(\alpha_k)}, \quad \alpha_K = 0 \tag{6}$$

The constraint on α_K is imposed because only $K - 1$ parameters are needed to specify. The last proportion is one minus the sum of the first $K - 1$.

If individual specific characteristics are provided, the mixing proportions are extended as [50, 51]

$$\pi_{ik} = \frac{\exp(\boldsymbol{\theta}_k\mathbf{z}_i)}{\sum_{k=1}^{K}\exp(\boldsymbol{\theta}_k\mathbf{z}_i)}, \quad \boldsymbol{\theta}_K = 0 \tag{7}$$

where $\boldsymbol{\theta}_K$ is the vector of component-specific parameters and $\mathbf{z}_i$ is an optional set of individual-specific characteristics for observation i.

For the observed random sample, $(\mathbf{x}_i, \mathbf{y}_i)$ $(i = 1, \ldots N)$, the log likelihood function for ψ is given by

$$\begin{aligned}\log L(\psi) &= \sum_{n=1}^{N}\sum_{i=1}^{M_n}\log h(\mathbf{y}_i \mid \mathbf{x}_n, \psi) \\ &= \sum_{i=1}^{N}\sum_{i=1}^{M_n}\log\left(\sum_{k=1}^{K}\pi_k h(\mathbf{y}_i \mid \mathbf{x}_n, \theta_k)\right)\end{aligned} \tag{8}$$

The maximum likelihood (ML) estimate of ψ is given by an appropriate root of the likelihood equation,

$$\frac{\partial \log L(\psi)}{\partial \psi} = 0 \tag{9}$$

The conditional probability that observation $(\mathbf{x}_i, \mathbf{y}_i)$ belongs to component j is given by

$$P(j \mid \mathbf{x}_i, \mathbf{y}_i, \psi) = \frac{\pi_j f(\mathbf{y}_i \mid \mathbf{x}_i, \theta_j)}{\sum_{k=1}^{K}\pi_k f(\mathbf{y}_i \mid \mathbf{x}_i, \theta_k)} \tag{10}$$

The conditional probabilities can be used to segment data by assigning each observation to the component with maximum conditional probability [50, 51]. A probabilistic segmentation of the data into K components can be obtained in terms of the fitted conditional probabilities. In the FMLR model we consider the latent component-indicator variables $\widehat{z}_n = \widehat{z}_{n1}, \ldots, \widehat{z}_{nK}, n = 1, \ldots, N$, to classify each single observation:

$$z_{nk} = \begin{cases} 1, & \text{if } y_n \text{ belongs to component } k \\ 0, & \text{otherwise} \end{cases} \tag{11}$$

The estimator of z_{jk}, $\widehat{z}_{nk}$ is

$$\widehat{z}_{nk} = \begin{cases} 1, & \text{if } \widehat{\pi}_k(\mathbf{y}_n; \widehat{\Psi}) \geq \widehat{\pi}_h(\mathbf{y}_n; \widehat{\Psi}), \ (h = 1, \ldots, K; h \neq k) \\ 0, & \text{otherwise} \end{cases} \tag{12}$$

4.2. Model Parameter Estimation. Parameters of FMLR models can be efficiently estimated through the EM algorithm [52].

(1) Initialization Step. Start with an initial seed (guess) for the parameter $\widehat{\psi}$ using the K-means clustering algorithm [53].

(2) E-Step. Estimate the conditional component probabilities, $\widehat{\pi}_{ik}$, for each observation using (7) and derive the mixing proportions as

$$\widehat{\pi}_k = \frac{1}{N}\sum_{i=1}^{N}\widehat{\pi}_{ik} \tag{13}$$

(3) M-Step. Maximize the log-likelihood for each component separately using the conditional probabilities as weights:

$$\max_{\theta_k}\sum_{i=1}^{N}\widehat{\pi}_{ik}\log\left(\mathbf{y}_i \mid \mathbf{x}_i, \theta_k\right) \tag{14}$$

The EM algorithm alternates between the expectation and the maximization steps until the likelihood improvement falls under a prespecified threshold or a maximum number of iterations are reached.

But the drawbacks of EM algorithm are its possible slow convergence rate and long processing time in computer. Thus, in this paper, Latent GOLD 5.0 is used to estimate the parameters. Latent GOLD 5.0 can take the advantages of both EM and Newton-Raphson algorithms. It first uses EM algorithm to get close to the final solution and then switches to Newton-Raphson to finish estimation [54].

The most important and difficult step in building FMLR model is to determine K, the number of components. Since K is not a parameter, hypotheses on K cannot be tested directly. BIC or AIC [50, 51, 55, 56] are generally used as criterion to determine K. In this study, we determined K based on BIC:

$$BIC_{\text{model}} = -2LL + \gamma\log(N) \tag{15}$$

where LL is the log-likelihood value, γ is the number of free parameters to be estimated, and N is the number of observations in the data. A lower BIC value indicates a better model.

5. Results and Discussion

5.1. Results. To select an optimal model, we apply the FMLR model having an increasing number of components from 1 to 4 to fit, and apply Bayesian Information Criterion (BIC) as the indicator to select the most appropriate number of components. Table 4 shows the BIC values of models for different number of components. It can be observed from Table 4 that the lowest BIC value occurs at $K = 2$. Hence, it is plausible to select $K = 2$ as a proper number of components.

To select the model variables, the forward-selection method is adopted in this paper. It starts with no variables in the model, tests the addition of each variable using Wald-statics, and adds the variable that gives the most statistically significant improvement of the fit. In this paper, variables will be added one by one until none produce a significant Wald-statistic in all components.

Table 5 shows the estimation results. For comparison, the result of logistic regression is also provided. In this paper, the component mixing proportions are a set of fixed constants (see (6)), as no sociodemographic characteristics of drivers are available in this dataset. The proportion of merging vehicle drivers in each component as indicated by H value in Table 5 is 67.2% and 32.8%, respectively.

Table 4: BIC value of FMLR model.

The Number of Components	BIC Value
$K = 1$	790.2955
$K = 2$	773.3871
$K = 3$	808.3826
$K = 4$	843.3061

By using (10)-(12), 374 drivers are classified into two components. One is the larger component, comprising 298 drivers and 612 observations, and the other is the smaller component, containing 75 drivers and 314 observations. To better understand the classification results, the mean values and standard deviations of related attribute variables are shown in Table 6.

5.2. Discussion. As seen from significance levels of parameters of Component 1 in Table 5, $\Delta D^i_{n\text{lg}}$ and ΔV^i_{nLead} fail to be significant at the 99% level. These suggest that front vehicles in auxiliary lane do not alter drivers' merge decisions in this component. Another impressing characteristic of this component is that the drivers have a distinct preference for the larger gaps. The negative sign of V^i_n indicates that drivers in this component tend to decrease their speeds during merging process. Consistent with previous studies, the decrease of speed difference between merging vehicle and putative leading vehicle and a gap located further towards the end of the auxiliary lane also increase the probability of accepting the current gap.

It is interesting to find that the parameter of D^i_n in Component 2 is much smaller than that in Component 1, which means the drivers in Component 2 do not pursue larger gaps as drivers in Component 1. In addition, speed difference between merging vehicle and putative leading vehicle is still important during merging process. Different from Component 1, $\Delta D^i_{n\text{lg}}$ and V^i_{nLead} are considered by drivers in Component 2. The sign of the parameter for $\Delta D^i_{n\text{lg}}$ is positive, suggesting that space in the auxiliary lane also affects the merging behaviors of drivers in Component 2 and the merging vehicle has a high probability of accepting a gap when there is an adequate space in front of the merging vehicle. One interesting finding from Table 5 is that the sign of the parameter for V^i_{nLead} is negative, suggesting that drivers in Component 2 are more likely to delay merge when the leading vehicle moves too fast. One possible reason for this result might be that when the leading vehicles move faster in the auxiliary lane, the drivers are provided more space in the auxiliary lane and they are using the auxiliary lane to reach further downstream in the main lane.

As illustrated in Table 6, the related variables show obvious differences across the two components. The average numbers of rejected gaps of the two components are 1.05

Table 5: Model estimation results of FMLR model.

Variables	Logistic Regression	FMLR(K = 2)	
		component 1(0.672)	Component 2(0.328)
	Parameter	Parameter	Parameter
V_n^i	-	-0.1810∗	0.1063∗
ΔV_{nPL}^i	-0.40848∗	-0.3903∗	-0.2557∗
D_n^i	.05490∗	0.1895∗	0.0158∗
Y_n^i	.01345∗	0.0109∗	0.0105∗
V_{nLead}^i	-.07111∗	-0.0400	-0.0568∗
ΔD_{nlg}^i	.01370∗	0.0037	0.0113∗
Constant	1.26281∗	0.8417∗	-1.7619∗

Note: ∗ means that the parameters are significant at 99% level.

Table 6: Mean values and standard deviations of related variables in each component.

Variables	Component 1		Component 2	
	Rejected Gaps (Standard Deviation)	Accepted Gaps (Standard Deviation)	Rejected Gaps (Standard Deviation)	Accepted Gaps (Standard Deviation)
V_n^i**(m/s)**	15.050 (3.196)	13.418 (3.107)	13.505 (2.852)	14.272 (3.466)
ΔV_{nPL}^i**(m/s)**	8.611 (3.825)	1.985 (2.766)	5.375 (3.698)	3.187 (2.908)
D_n^i**(m)**	10.068 (5.274)	33.14 (22.32)	17.468 (15.175)	27.09 (23.65)
V_{nLead}^i**(m/s)**	9.841 (7.114)	11.627 (6.671)	10.529 (6.420)	8.062 (8.100)
ΔD_{nlg}^i**(m)**	43.42 (47.04)	44.83 (42.99)	33.05 (35.06)	26.79 (35.46)
Merge Location**(m)**	41.66 (57.87)		108.58 (64.19)	
Number of Rejected Gaps	1.05		3.19	

and 3.19 and the average merge location is 41.66m and 108.58m, which indicates that drivers in Component 2 tend to choose gaps further downstream and rejected more gaps than drivers in Component 1. The average rejected gap of Component 2 (17.468 m) is much bigger than Component 1 (10.068m) while the average accepted gap of Component 2 (27.09 m) is much smaller than Component 1 (33.14 m), indicating the inconsistency of gap acceptance theories. One can also find that the drivers in Component 2 increase their speeds during their merging process from 13.505m/s to 14.272 m/s, while drivers in Component 1 decrease their speed from 15.050 m/s to 13.418m/s, and in Component 2, the speed difference between the putative leading vehicle and the merging vehicle for accepted gaps is 3.187m/s, which is much bigger than Component 1, both of which indicate that drivers in Component 2 are more aggressive than Component 1. It is interesting to find that the standard deviations of the speeds for rejected gaps and accepted gaps in Component 1 are similar, which is not the case in Component 2. And one can also find that the standard deviation of rejected gaps for Component 2 is much bigger than that in Component 1. These findings indicate that the merging process of drivers in Component 2 is much more complicated than drivers in Component 1.

Figure 2 shows the relation between the gap size and location for the rejected and accepted gaps in the two components. One can find that the accepted gaps of drivers in Component 1 are almost all located in the beginning

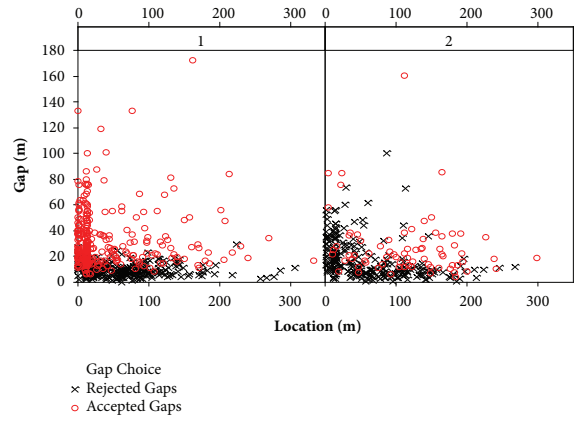

Figure 2: Relation between the gap size and location for the rejected and accepted gaps.

of the auxiliary lane while the accepted gaps of drivers in Component 2 are scattered along the lane. It is obvious that the rejected gaps of drivers in Component 2 are much larger than in Component 1 and are overlapped with the rejected gaps, while the overlapping area in Component 1 is much smaller.

Figure 3 shows the box plot of the reverse succession of offered gaps. The x-axis in Figure 3 is the reverse number of offered gaps before merging, in which 0 means the finally accepted gap and 1 means the last rejected gap before

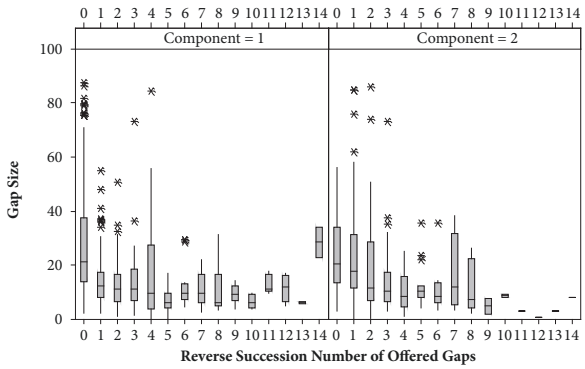

Figure 3: Box plot of the reverse succession of offered gaps.

Table 7: Comparison of estimated and observed values of logistic regression model.

	Estimated		
Observed	Reject	Accept	Total
Reject	**484.0**	68.0	552.0
Accept	92.0	**281.0**	373.0
Total	576.0	349.0	925.0

Table 8: Comparison of estimated and observed values of FMLR-2 model.

	Estimated		
Observed	Reject	Accept	Total
Reject	**510.0**	42.0	552.0
Accept	39.0	**334.0**	373.0
Total	547.0	378.0	925.0

merging. One can find that drivers in Component 2 might have several choices before merging, which indicates that drivers in Component 2 prefer to use the auxiliary lane to get further downstream.

Comparing the two components, drivers in Component 1 prefer larger gaps and lower speed difference, while drivers in Component 2 pay more attention to better surrounding traffic conditions and may sacrifice larger gaps to save travel time and get better traffic conditions. Thus, in this paper, Component 1 is named as Risk-Rejecting Drivers and Component 2 is named as Risk-Taking Drivers.

5.3. Accuracy of Developed Models. Tables 7 and 8 show the comparison of estimated and observed values of logistic regression model and 2-component mixture of logistic regression (FMLR-2) model. From theses tables, the proposed model improves the predicting accuracy from 82.70% to 91.24%. It can be concluded that the proposed model has better predictive power than logistic regression model.

6. Conclusions

To incorporate the unobserved heterogeneity into merge model, the present study builds a FMLR model which uses BIC to determine the proper number of mixing components and performs parameter estimation by using Latent GOLD 5.0.

Given U.S. Highway 101 data, the identified optimal model is a 2-component mixture of logistic regression model, which means the drivers can be divided into two components characterized by the driving behavior heterogeneity. One is the Risk-Rejecting Drivers whose drivers are consistent with previous studies and primarily merge in as soon as possible. Drivers in this component have a distinct preference for the larger gaps. The decrease of speed difference between merging vehicle and putative leading vehicle and a gap located further towards the end of the auxiliary lane also increase the probability of accepting the current gap. Contrast to Component 1, Component 2 is constituted with the drivers that are much less sensitive to the gap size and have more emphasis on surrounding traffic conditions such as the speed of front vehicle in the auxiliary lane and space gap between the merging vehicle and its leading vehicles in the auxiliary lane. These drivers are using the auxiliary lane to get to the further downstream or less congested area of the main lane. Thus they are called Risk-Taking Drivers.

In addition, the proposed model can produce more precise predicting accuracy than logistic regression model.

However, more empirical studies are needed to apply this method to datasets in other sites with different demographics, climate, and geometric parameters in order to fully assess the effect of the factors affecting merging behaviors as well as fully understand the strengths and weaknesses of the proposed model.

Data Availability

The NGISM data used to support the findings of this study have been deposited at the website: https://catalog.data.gov/dataset/next-generation-simulation-ngsim-vehicle-trajectories.

Conflicts of Interest

The author declares that there are no conflicts of interest regarding the publication of this paper.

References

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

[2] H. Yi and T. E. Mulinazzi, "Urban freeway on-ramps: Invasive influences on main-line operations," *Transportation Research Record*, no. 2023, pp. 112–119, 2007.

[3] S. P. Hoogendoorn and R. Hoogendoorn, "Generic calibration framework for joint estimation of car-following models by using microscopic data," *Transportation Research Record*, no. 2188, pp. 37–45, 2010.

[4] J. Kim and H. S. Mahmassani, "Correlated parameters in driving behavior models," *Transportation Research Record*, vol. 2249, pp. 62–77, 2011.

[5] 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.

[6] I. Kim, T. Kim, and K. Sohn, "Identifying driver heterogeneity in car-following based on a random coefficient model," *Transportation Research Part C: Emerging Technologies*, vol. 36, pp. 34–44, 2013.

[7] Cambridge Systematics, *NGSIM US 101 Data Analysis: Summary Report*, Prepared for Federal Highway Administration, 2005.

[8] 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.

[9] K. I. Ahmed, *Modeling drivers' acceleration and lane changing behavior*, Massachusetts Institute of Technology, 1999.

[10] G. Lee, *Modeling gap acceptance at freeway merges*, Massachusetts Institute of Technology, 2006.

[11] T. Toledo, H. N. Koutsopoulos, and M. Ben-Akiva, "Integrated driving behavior modeling," *Transportation Research Part C: Emerging Technologies*, vol. 15, no. 2, pp. 96–112, 2007.

[12] 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.

[13] 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.

[14] 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.

[15] D. C. Gazis, R. Herman, and R. W. Rothery, "Nonlinear follow-the-leader models of traffic flow," *Operations Research*, vol. 9, no. 4, pp. 545–567, 1961.

[16] A. J. Miller, *Nine estimators of gap-acceptance parameters*, Publication of: Traffic Flow and Transportation, 1971.

[17] H. Mahmassani and Y. Sheffi, "Using gap sequences to estimate gap acceptance functions," *Transportation Research Part B: Methodological*, vol. 15, no. 3, pp. 143–148, 1981.

[18] 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.

[19] 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.

[20] J. Wang, "A simulation model for motorway merging behaviour," *Transportation and traffic theory*, vol. 16, pp. 281–301, 2005.

[21] SiAS, *S-Paramics 2005 Reference Manual*, SIAS Ltd Edinburgh, 2005.

[22] PTV-Vision, *VISSIM 5.30-05 user manual*, 2011.

[23] 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.

[24] T. D. CHU, *A Study on Merging Behavior at Urban Expressway Merging Sections*, Nagoya University, 2014.

[25] H. Kita, "Effects of merging lane length on the merging behavior at expressway on-ramps," *Transportation and Traffic Theory*, pp. 37–51, 1993.

[26] H. Kita, "A merging-giveway interaction model of cars in a merging section: a game theoretic analysis," *Transportation Research Part A: Policy and Practice*, vol. 33, no. 3-4, pp. 305–312, 1999.

[27] 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.

[28] A. Rao, *Modeling Anticipatory Driving Behavior [MS Thesis]*, Department of Civil and Environmental Engineering, MIT, 2006.

[29] C. F. Choudhury, M. E. Ben-Akiva, T. Toledo, G. Lee, and A. Rao, "Modeling cooperative lane changing and forced merging behavior," in *Proceedings of the 86th Annual Meeting of the Transportation Research Board*, C., Ed., Washington, DC, 2007.

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

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

[32] S. Ossen and S. P. Hoogendoorn, "Car-following behavior analysis from microscopic trajectory data," *Transportation Research Record*, no. 1934, pp. 13–21, 2005.

[33] 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.

[34] X. Ma and I. Andréasson, "Statistical analysis of driver behavior data in different regimes of the car-following stage," *Transportation Research Record*, no. 2018, pp. 87–96, 2007.

[35] G. Li and L. Sun, "Characterizing Heterogeneity in Drivers Merging Maneuvers Using Two-Step Cluster Analysis," *Journal of Advanced Transportation*, 2018.

[36] M. Keyvan-Ekbatani, V. L. Knoop, and W. Daamen, "Categorization of the lane change decision process on freeways," *Transportation Research Part C: Emerging Technologies*, vol. 69, pp. 515–526, 2016.

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

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

[39] V. Punzo, M. T. Borzacchiello, and B. Ciuffo, "On the assessment of vehicle trajectory data accuracy and application to the Next Generation SIMulation (NGSIM) program data," *Transportation Research Part C: Emerging Technologies*, vol. 19, no. 6, pp. 1243–1262, 2011.

[40] M. Montanino and V. Punzo, "Trajectory data reconstruction and simulation-based validation against macroscopic traffic patterns," *Transportation Research Part B: Methodological*, vol. 80, pp. 82–106, 2015.

[41] B. Coifman and L. Li, "A critical evaluation of the Next Generation Simulation (NGSIM) vehicle trajectory dataset," *Transportation Research Part B: Methodological*, vol. 105, pp. 362–377, 2017.

[42] 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, pp. 246–253, 2014.

[43] X. Wan, P. J. Jin, F. Yang, and B. Ran, "Merging preparation behavior of drivers: How they choose and approach their merge

positions at a congested weaving area," *Journal of Transportation Engineering*, vol. 142, no. 9, 2016.

[44] X. Wan, P. J. Jin, H. Gu, X. Chen, and B. Ran, "Modeling Freeway Merging in a Weaving Section as a Sequential Decision-Making Process," *Journal of Transportation Engineering, Part A: Systems*, vol. 143, Article ID 05017002, 2017.

[45] G. McLachlan and D. Peel, *Finite mixture models , willey series in probability and statistics*, John Wiley & Sons, New York, USA, 2000.

[46] C. Fraley and A. E. Raftery, "Model-based clustering, discriminant analysis, and density estimation," *Journal of the American Statistical Association*, vol. 97, no. 458, pp. 611–631, 2002.

[47] W. S. DeSarbo and W. L. Cron, "A maximum likelihood methodology for clusterwise linear regression," *Journal of Classification*, vol. 5, no. 2, pp. 249–282, 1988.

[48] M. Wedel and W. S. DeSarbo, "A mixture likelihood approach for generalized linear models," *Journal of Classification*, vol. 12, no. 1, pp. 21–55, 1995.

[49] M. Wedel and W. A. Kamakura, *Market segmentation: Conceptual and methodological foundations*, Springer Science & Business Media, 2012.

[50] W. H. Greene and D. A. Hensher, "A latent class model for discrete choice analysis: contrasts with mixed logit," *Transportation Research Part B: Methodological*, vol. 37, no. 8, pp. 681–698, 2003.

[51] W. H. Greene, *Interpreting estimated parameters and measuring individual heterogeneity in random coefficient models*, 2004, Interpreting estimated parameters and measuring individual heterogeneity in random coefficient models.

[52] A. P. Dempster, N. M. Laird, and D. B. Rubin, "Maximum likelihood from incomplete data via the EM algorithm," *Journal of the Royal Statistical Society Series: B*, vol. 39, no. 1, pp. 1–38, 1977.

[53] G. McLachlan and T. Krishnan, *The EM algorithm and extensions*, John Wiley & Sons, 2007.

[54] J. K. Vermunt and J. Magidson, *Latent GOLD 5.0 upgrade manual*, Statistical Innovations Inc, Belmont, MA, 2013.

[55] H. Akaike, "A new look at the statistical model identification," *IEEE Transactions on Automatic Control*, vol. 19, pp. 716–723, 1974.

[56] G. Schwarz, "Estimating the dimension of a model," *The Annals of Statistics*, vol. 6, no. 2, pp. 461–464, 1978.