Analysis of Interdriver Heterogeneity Based on Trajectory Data with K-means Clustering Method

Tailang Zhu [1, a *], Dongfan Xie[1, b]

[1]School of Traffic and Transportation, Beijing Jiaotong University, Beijing 100044, China

[a]14120767@bjtu.edu.cn, [b]dfxie@bjtu.edu.cn

Keywords: Driver heterogeneity; K-means clustering; Car-following; NGSIM

Abstract. This paper presents a methodology to study the interdriver heterogeneity by using vehicle trajectory data. Different from the existing studies subjectively dividing the drivers into two or three types, this paper explores a K-means clustering methodology to classify the drivers based on real traffic data. So the classification would be more reasonable. In terms of the vehicle trajectory data extracted from the Next Generation Simulation (NGSIM) project, such microscopic variables as velocity, acceleration, spacing (space headway) and headway (time headway) are selected to represent the heterogeneity among drivers. The findings suggest that headway is the best variable to describe drivers' heterogeneity, and spacing is the second best. Additionally, according to the two selected variables, the drivers are divided into three types: stable driver, timid driver and aggressive driver.

Introduction

In recent years, rapid development of the global economy has given further impetus to urban vehicle ownership. It leads to many traffic problems including traffic congestion, traffic safety, environmental pollution, fossil fuel consumption and so on. Recent studies [1] have shown that many traffic phenomena e.g., the stop-and-go waves, are not likely to be reproduced with a single driver type. The heterogeneity of drivers is essential for the realization of a real traffic.

Generically, driver heterogeneity includes interdriver heterogeneity and intradriver heterogeneity two types. Interdriver heterogeneity describes the idea that different drivers may have different reactions to the same stimulus. Intradriver heterogeneity implies that the same driver may react differently to the same stimulus at different times or under different conditions [2]. Stimulus includes spacing, headway and so on. Driver may have a wrong prediction during driving in some cases [3] and there are many factors causing this wrong prediction, e.g. weather conditions, driver age, driver gender, prevailing traffic conditions, vehicle types and so on [4,5,6,7,8,9]. For the above factors, the most basic research is the use of traffic flow model to describe the driving behavior characteristics of different driver based on the model in the corresponding coefficient [10,11,12]. However, it is not enough to calibrate car-following model parameters alone and should take complexity of model and correlation between parameters into account. So some scholars extend the research object from a single driver to multiple drivers and define the complexity of model with the number of model parameters. They also compare the difference between models and calibrate different models parameters [13]. The data used for interdriver heterogeneity and intradriver heterogeneity is different. Interdriver heterogeneity using trajectory data is driver-specific data, but intradriver heterogeneity using trajectory data is time-varying data. Because of the research method used, we focus on interdriver heterogeneity in this paper.

In the paper, a K-means method is proposed to analyze driver heterogeneity and distinguish between different types of drivers. This study may help traffic management to develop a series of laws and regulations to improve driving safety and help drivers adjust their driving habits. Section 2 is a simple introduction of the K-means clustering method, and section 3 analyses the US-101 NGSIM trajectory datasets. Finally, some conclusions are provided in section 4.

K-means Clustering Method

Many studies investigate drivers' behavioral intention, just like a timid driver who has a larger response time and minimum spacing features [14], to classify them into different categorizes. There are few studies which discuss the driver categories deeply and research on different driver types' parameters. That is why we use K-means clustering method to examine driver heterogeneity. In this paper, we classify drivers first and then analyze driver heterogeneity under different clustering number.

Clustering is a process to distinguish and classify things in accordance with certain requirements and rules, without any prior knowledge about the class divide in this process, relying on the guidelines as a category similarity between things belonged to divide. Cluster analysis refers to the classification of research methods and mathematical treatment of the given object. It is the task of grouping a set of objects in such a way that objects in the same group are more similar to each other than to those in other groups. Cluster analysis methods can be divided into four types: (1) hierarchical clustering method, (2) clustering based on equivalence relation method, (3) graph clustering method, (4) clustering based on the objective function. The first three cluster analysis methods are not likely to deal with large capacity data, thus the objective function is introduced here.

The objective function of K-means clustering is distance. Distance is the similarity evaluation criteria, and the closer of the two objects, the greater its similarity. "K" represents number of the final clusters. Steps of the K-means clustering method are:

(1) Choose K objects as the initial cluster centers randomly from the N data objects;

(2) According to the mean (central object) each cluster object, and the object distance is calculated for each selected center of the object;

(3) Re-calculate the mean of each cluster, the cluster centers recalculated position;

(4) Calculation of the standard measurement function, the algorithm terminates when certain conditions are met; if the conditions are not satisfied, return to step (2).

Each object involved in a number of variables, the choice of input variables must follow the following rules:

Rule 1 Selected variables should meet the requirements of clustering;

Rule 2 Variables should not have a strong linear correlation;

Rule 3 The value of each variable shouldn't have on the order of magnitude difference.

Data Analysis

NGSIM Data Filtering. This study employs US-101 vehicle trajectory data from FHWA's Next Generation Simulation (NGSIM) program (FHWA, 2008). The data was extracted from video images of southbound traffic on US-101 in Los Angeles, California. The selected section of US-101 is approximately 640 m, as shown in Fig. 1 [14], and trajectories on this study segment were collected from 7:50 a.m. to 8:05 a.m. on June 15, 2005. Data for these trajectories is available at 0.1s resolution. Lane 1 is farthest left lane; lane 5 is farthest right lane. Lane 6 is the auxiliary lane. Lane 7 is the on-ramp at Ventura Boulevard, and Lane 8 is the off-ramp at Cahuenga Boulevard. This paper only studies the five mainline lanes. Congestion appeared from 7:50 a.m. to 8:05 a.m. Free-flow and congestion are included in this period. So it is the proper data to investigate driver heterogeneity. We can observe the space-time diagrams, as shown in Fig. 2, and found this transitional traffic process.

The data includes vehicle location, velocity, acceleration, length, width, the lane number, spacing, headway and so on. This paper focus on driver heterogeneity in car-following process and the data should be filtered and obey the following constraints:

(1) The vehicle must have a leader vehicle;

(2) Spacing is greater than vehicle length and less than 120m;

(3) The vehicle didn't change lanes.

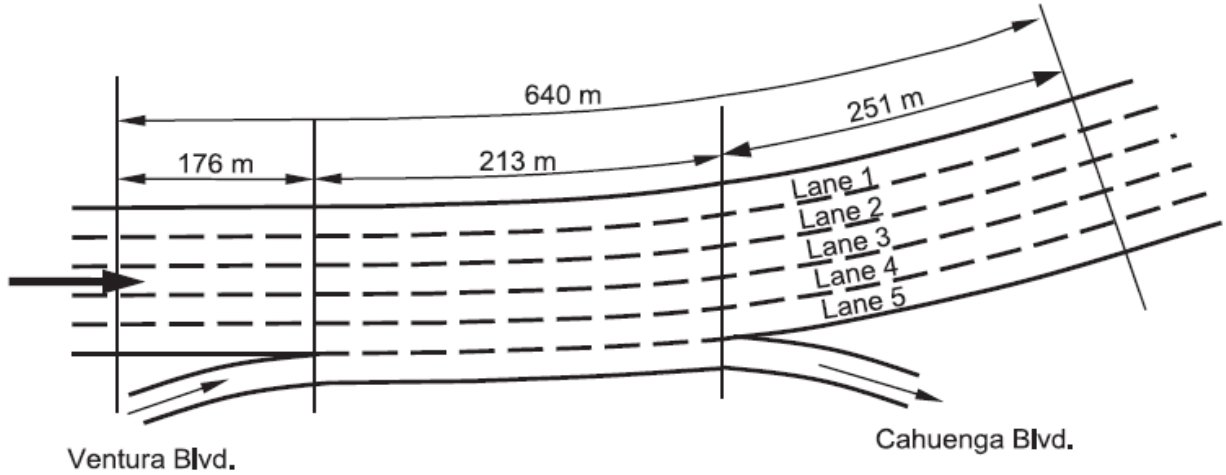

Figure 1. Schematic of the study site: southbound US-101 in Los Angeles, California (NGSIM, 2006).

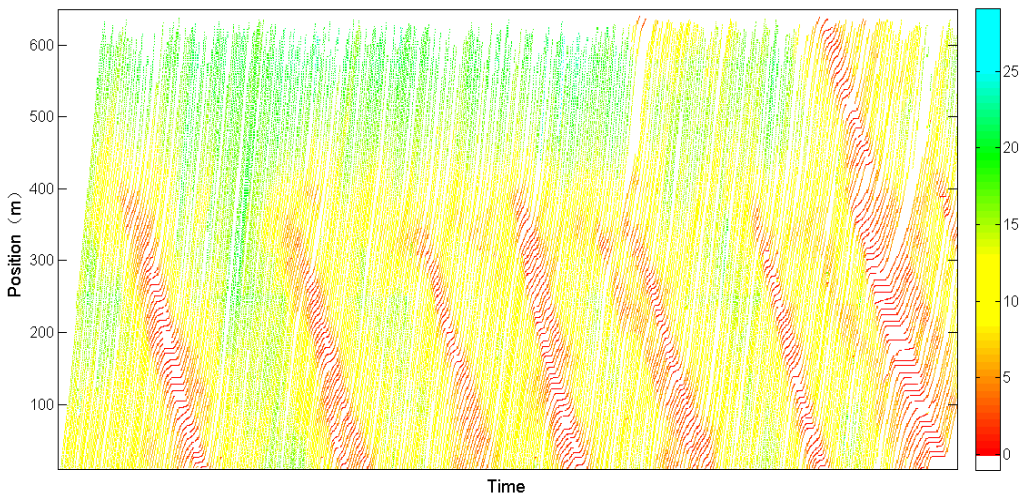

Figure 2. 7:50 a.m.-8:05 a.m. space - time diagrams

We get 1108359 observation data which belongs to 2164 drivers' trajectories through the filtering process. The amount of different vehicles is shown in Table 1. Due to the different driver behaviors with automobile and the insufficient data, the data of both large vehicles and the motorcycles are abandoned. Here, we focus on the heterogeneity of automobile drivers.

Table 1 Vehicle type and number

Vehicle type	Automobile	Truck and Buses	Motorcycle
Number	2081	53	30

According to the rules of variable selection based on clustering analysis, this paper chooses average velocity (AV), average acceleration (AA), average spacing (AS) and average headway(AH) as the variables. Though the four variables meet Rule 1 and 2, their orders of magnitude have a huge difference which violates Rule 3. In order to solve this problem and get rigorous results, we standardize the variables with the following method:

$$x'_{p_{ij}} = \frac{x_{p_{ij}} - \min\{x_{p_i}\}}{\max\{x_{p_i}\} - \min\{x_{p_i}\}} \tag{1}$$

where:

$x'_{p_{ij}}$= the jth data in ith variable of the pth driver after standardization;

$x_{p_{ij}}$= the jth data in ith variable of the pth driver before standardization;

$\min\{x_{p_i}\}$= minimum in the ith variable of the pth driver;

$\max\{x_{p_i}\}$= maximum in the ith variable of the pth driver.

Driver Heterogeneity Analysis. Recent studies [1] have shown that many traffic phenomena, e.g., congestion, the stop-and-go waves, are not likely to be reproduced with homogeneous drivers. To obtain a reliable traffic situation, K is set to be not smaller than 2. Also, the utility of classification could be weaken with a too large number of K. As a result, the clustering number of K varies from 2 to 7. In this study, 100 times clustering iterative will be done for each calculation to avoid local optima, and the clustering results are shown in Fig. 3.

Fig. 3 shows the silhouette value under different clustering K. The abscissa axis and ordinate axis represent silhouette value and the number of cluster, respectively. Silhouette value represents the approximation of each individual to the clustering central points. If silhouette value is close to 1, the individual is close to the centre. Instead, there is no correlation between the individual and centre if silhouette value is equal to 0. Some silhouette value is less than 0 because some individuals don't belong to this category. The biggest silhouette value is close to 0.6 and it shows that clustering results are relatively proper. Because every individual has four variables, high-dimensional has effected on results to a certain degree.

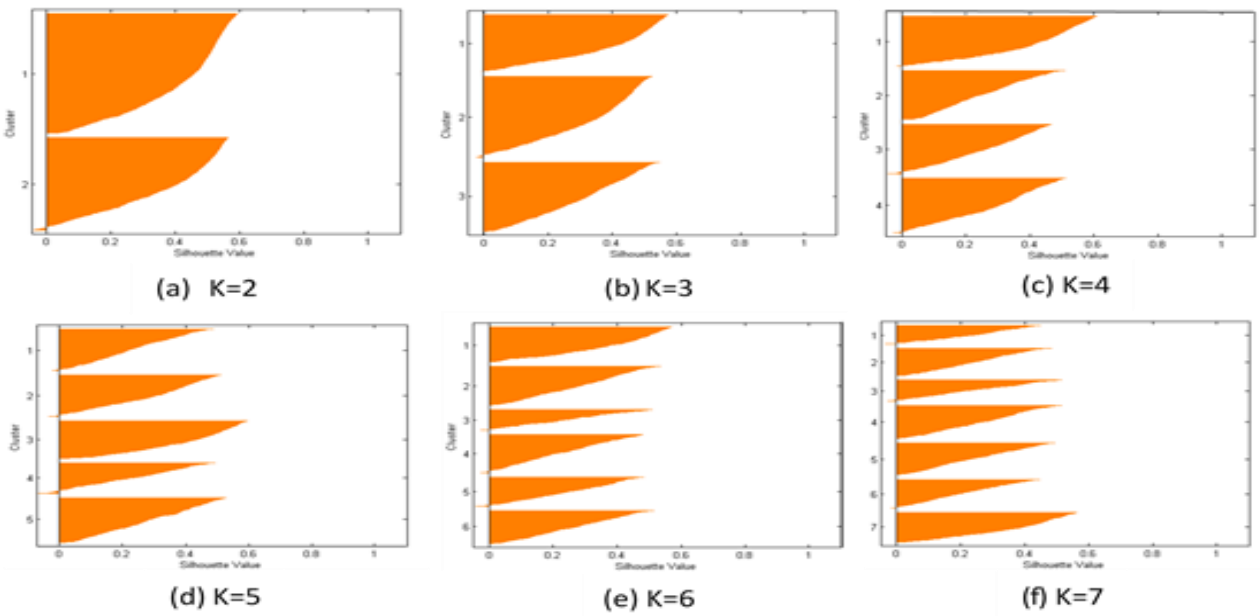

Figure 3. Silhouette value under different K

Table 2 shows clustering central point parameters under various K. The second row of Table 2 is clustering central point parameters. Central point parameters include such four variables as have average velocity (AV), average acceleration (AA), average spacing (AS) and average headway (AH).

Table 2 Clustering central point parameters under different K.

K	Central point parameters				
2	Number	AV	AA	AS	AH
	1	0.4626	0.5184	0.4672	0.4646
	2	0.4882	0.5134	0.3968	0.1672
3	Number	AV	AA	AS	AH
	1	0.5063	0.5109	0.4427	0.1063
	2	0.4744	0.5177	0.5149	0.4944
	3	0.4465	0.5189	0.3394	0.3348
4	Number	AV	AA	AS	AH
	1	0.5034	0.5103	0.4348	0.0913
	2	0.5125	0.5152	0.4997	0.3724
	3	0.4506	0.5198	0.5090	0.5483
	4	0.4332	0.5193	0.3158	0.3312
5	Number	AV	AA	AS	AH
	1	0.5159	0.5140	0.4906	0.3440
	2	0.4734	0.5194	0.5446	0.5472
	3	0.5160	0.5086	0.4668	0.0812
	4	0.4449	0.5178	0.2844	0.2080
	5	0.4201	0.5206	0.3650	0.4282
6	Number	AV	AA	AS	AH
	1	0.5253	0.5085	0.4784	0.0907
	2	0.4457	0.5190	0.3325	0.3313
	3	0.4382	0.5165	0.2828	0.1298
	4	0.5188	0.5142	0.5060	0.3645
	5	0.4788	0.5204	0.5699	0.5626
	6	0.4171	0.5196	0.4099	0.4903
7	Number	AV	AA	AS	AH
	1	0.4567	0.5213	0.5749	0.5993
	2	0.5254	0.5153	0.5370	0.4509
	3	0.4444	0.5162	0.2979	0.1190
	4	0.4381	0.5196	0.3112	0.3341
	5	0.4165	0.5192	0.4120	0.4848
	6	0.5049	0.5160	0.4712	0.3043
	7	0.5299	0.5073	0.4862	0.0820

Table 3 is shown the changes of central point drivers' velocity under different cluster number K. With the increase of K, the maximum velocity (MAV) increases, while the minimum velocity (MIV) decreases. The ratio of them tends to increase as well. So we can induce that driver heterogeneity is amplified with the increasing of K in a certain range.

Table 3 Changes of drivers' velocity

K	MAV	MIV	Ratio (MAV/ MIV)
2	0.4882	0.4626	1.0553
3	0.5063	0.4465	1.1339
4	0.5125	0.4332	1.1831
5	0.5160	0.4201	1.2283
6	0.5253	0.4171	1.2594
7	0.5299	0.4165	1.2723

Form Table 2 and Table 3, one can find that drivers' acceleration keeps stable in spite of differences of K and induce acceleration may not the proper variable to describe driver

heterogeneity. The reason for this situation is that acceleration is the rate of change of the velocity at a moment, so it can't reflect heterogeneity in time period.

Unlike the former two kinds of variables, we observe the line chart of clustering central points in Fig. 4. In Fig. 4, the numbers 1 to 4 denote AV, AA, AS and AH, respectively. We find the data of AS and AH is different prominently. The headway variation range is larger than the range of the spacing under the same K. Particularly, the latter two variables reflect heterogeneity obviously under the different K. The clustering results indicate that headway is the most appropriate representation of the driver heterogeneous variables, followed by headway, and finally the average velocity.

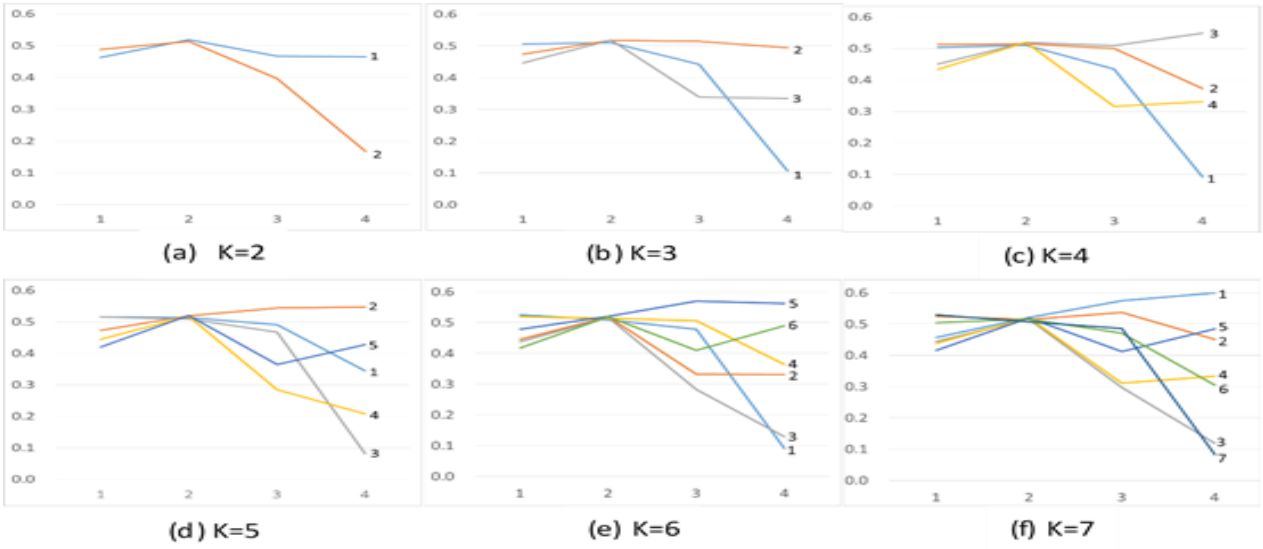

Figure 4. line chart of clustering central points

Further analysis of the obtained clustering results can be found the drivers can be divided into various categories in the different number of clusters, but the three categories may be the most reasonable cluster result. The first type of driver called "stable" that the relationship between headway and spacing is relatively constant. The second type is "aggressive" driver whose headway is less than the spacing, so it proves that aggressive driver tends to chase the leader vehicle. The last category is called "timid" driver whose driving behavior is opposite to aggressive drivers. They aren't eager to chase the leader and like to keep the spacing.

Conclusion

This paper explores a K-means clustering method to divide drivers into various categories based on the NGSIM US-101 data. The data is filtered with car-following behavior and through the standardization. The clustering results show the priority of variables: headway is the most appropriate description of the driver heterogeneous variables, followed headway, and finally the average velocity. With the combination analysis of headway and spacing, drivers are divided into "stable", "aggressive" and "timid" three categories.

The shortcoming of this article is unable to give an accurate variable range of different types of drivers which will be studied in the future work.

Acknowledgements

This work is supported by the Fundamental Research Funds for the Central Universities (Grant Nos. 2016JBM025).

References

[1] J.A. Laval and L. Leclercq: Philosophical Transactions of the Royal Society of London A: Mathematical, Physical and Engineering Sciences, Vol. 368 (2010) No. 1928, p.4519.

[2] J. Taylor, X. Zhou X and N.M. Rouphail: Transportation Research Part B: Methodological, (2015) No. 73, p.59.

[3] S. Ossen and S. Hoogendoorn: Transportation Research Record: Journal of the Transportation Research Board, (2005) No. 1934, p.13.

[4] L. Ma and X. Yan: Accident Analysis & Prevention, (2014) No. 67, p.129.

[5] A. Behnood, A.M. Roshandeh and F.L. Mannering: Analytic Methods in Accident Research, (2014) No. 3, p.56.

[6] P. Deo and H.J. Ruskin: Physica A: Statistical Mechanics and its Applications, (2014) No. 405, p.140.

[7] J.K. Kim, G.F. Ulfarsson and S. Kim: Accident Analysis & Prevention, (2013) No. 50, p.1073.

[8] V. Procher and C. Vance: Transportation Research Record: Journal of the Transportation Research Board, (2012) No. 2320, p.72.

[9] S. Ossen and S.P. Hoogendoorn: Transportation research part C: emerging technologies, Vol. 19 (2011) No. 2, p.182.

[10] S.P. Hoogendoorn, S. Ossen and M. Schreuder: Traffic and Granular Flow'05. Springer Berlin Heidelberg, (2007), p.687.

[11] S. Hoogendoorn, S. Ossen and M. Schreuder: Transportation Research Record: Journal of the Transportation Research Board, (2006) No. 1965, p.112.

[12] S. Ossen, S.P. Hoogendoorn and B. Gorte: Transportation Research Record: Journal of the Transportation Research Board, (2006) No. 1965, p.121.

[13] J. Kim and H. Mahmassani: Transportation Research Record: Journal of the Transportation Research Board, (2011) No. 2249, p.62.

[14] Z. Zheng, S. Ahn and D. Chen: Transportation research part C: emerging technologies, (2013) No. 26, p.367.