TRB Annual Meeting

An Investigation of Discretionary Lane-changing Decisions: Insights From the Third Generation SIMulation (TGSIM) Dataset

--Manuscript Draft--

Manuscript Number:	TRBAM-24-06377
Article Type:	Presentation and Publication

Abstract:	The data-driven characterization of discretionary lane-changing behaviors has traditionally been hindered by the scarcity of high-resolution data that can precisely record lateral movements. In this study, we conducted an exploratory investigation leveraging the Third Generation SIMulation (TGSIM) dataset to advance our understanding of discretionary lane-changing behaviors. In this paper, we developed a discretionary lane-changing extraction pipeline and scrutinized crucial factors such as gaps and relative speeds in leading and following directions. A Dynamic Time Warping (DTW) analysis was performed to quantify the difference between any pair of lane-changing behaviors, and an Affinity Propagation (AP) clustering, evaluated on normalized dynamic time warping distance, was conducted. Our results yielded five clusters based on lead and lag gaps, enabling us to categorize lane-changing behaviors into aggressive, neutral, and cautious for both leading and following directions. Clustering based on relative speeds revealed two distinct groups of lane-changing behaviors, one representing overtaking and the other indicative of transitioning into a lane with stable and homogenous speed. The proposed DTW analysis, in conjunction with AP clustering, demonstrated promising potential in categorizing and characterizing lane-changing behaviors. Additionally, this approach can be readily adapted to analyze any driving behavior.

An Investigation of Discretionary Lane-changing Decisions: Insights From the Third Generation SIMulation (TGSIM) Dataset

Yanlin Zhang
Research Assistant
Department of Civil and Environmental Engineering, University of Illinois at Urbana-Champaign
205 N Matthews, Urbana, IL 61801, USA
yanlinz4@illinois.edu

Alireza Talebpour, Ph.D., Corresponding Author
Assistant professor
Department of Civil and Environmental Engineering, University of Illinois at Urbana-Champaign
205 N Matthews, Urbana, IL 61801, USA
ataleb@illinois.edu

Hani S. Mahmassani, Ph.D.
William A. Patterson Distinguished Chair in Transportation, Professor
Northwestern University Transportation Center
600 Foster St., Evanston, IL 60208
masmah@northwestern.edu

Samer H. Hamdar, Ph.D.
Associate Professor
Department of Civil and Environmental Engineering, George Washington University
800 22nd Street NW Washington, DC 20052
hamdar@gwu.edu

Word Count: 4740 words + 1 table(s) $\times$ 250 = 4990 words

Submission Date: August 2, 2023

ABSTRACT

The data-driven characterization of discretionary lane-changing behaviors has traditionally been hindered by the scarcity of high-resolution data that can precisely record lateral movements. In this study, we conducted an exploratory investigation leveraging the Third Generation SIMulation (TGSIM) dataset to advance our understanding of discretionary lane-changing behaviors. In this paper, we developed a discretionary lane-changing extraction pipeline and scrutinized crucial factors such as gaps and relative speeds in leading and following directions. A Dynamic Time Warping (DTW) analysis was performed to quantify the difference between any pair of lane-changing behaviors, and an Affinity Propagation (AP) clustering, evaluated on normalized dynamic time warping distance, was conducted. Our results yielded five clusters based on lead and lag gaps, enabling us to categorize lane-changing behaviors into aggressive, neutral, and cautious for both leading and following directions. Clustering based on relative speeds revealed two distinct groups of lane-changing behaviors, one representing overtaking and the other indicative of transitioning into a lane with stable and homogenous speed. The proposed DTW analysis, in conjunction with AP clustering, demonstrated promising potential in categorizing and characterizing lane-changing behaviors. Additionally, this approach can be readily adapted to analyze any driving behavior.

Keywords: Discretionary Lane Changing, Driving behavior, Time Series Data Analysis

INTRODUCTION

Lane-changing and car-following behaviors are the two crucial elements of microscopic traffic flow theories; while car-following only addresses longitudinal movements, lane-changing behaviors involve both lateral and longitudinal maneuvers. Despite the extensive research on car-following behaviors over the years, lane-changing behaviors have garnered interest in the past two decades (*1*). This recent attention is attributed to the growing evidence of LC's adverse effects on traffic safety (*2, 3*), traffic flow oscillation(*4–7*), and capacity drop (*8*).

In light of the influence of lane-changing behavior on traffic safety, efficiency, and stability, there has been a swift increase in attempts to model lane-changing. Yang and Koutsopoulos (*9*) was one of the pioneers in modeling lane-changing behaviors by extending Gipps' car-following model (*10*) and implemented it into the microscopic traffic simulator, MITSIM. Then Kesting et al. (*11*) proposed the MOBIL model ("minimizing overall braking induced by lane changes") by simplifying the anticipated advantages and disadvantages of lane-changing as single-lane accelerations, which can be integrated with the Intelligent Driver's Model (IDM) (*12*). Another line of modeling lane-changing is based on discrete choice models. Ahmed et al. (*13*) is one of the earliest works defining a utility function for lane-changing behavior. Generally, lane-changing behaviors are categorized as either 'discretionary' or 'mandatory' (*1*). The primary intent behind a discretionary lane change is to choose a better driving condition, whereas the key incentive for a mandatory lane change is to arrive at the destination. Toledo et al. (*14*) developed a discrete choice framework capturing the trade-off between mandatory and discretionary lane changes with a single utility function. As lane-changing behaviors can be regarded as a competition and cooperation between the lane-changing and surrounding vehicles, Kita (*15*) first formally formulated merging with game theory to explain the real-world merging and giveaway behaviorally. Following similar strategies, Ban (*16*) then proposed robust payoff matrices for defining the strategies. As communication technology enhanced the connectivity among vehicles, Talebpour et al. (*17*) pioneered the effort in game-theoretic modeling lane-changing behaviors in a connected environment. More recent work by Ali et al. (*18*) extended Talebpour et al. (*17*)'s previous work and utilized an advanced driving simulator to collect high-quality vehicle trajectory data for the connected environment.

Furthermore, data-driven analysis of lane-changing behaviors has built a presence in the literature as sensing technology advances and data becomes more accessible, such as naturalistic driving data and trajectory data. Yang et al. (*19*) investigated the gap and relative speed based on naturalistic driving data collected from Shanghai, China, and results showed that gaps are significantly impacted by surrounding environments. However, this dataset only contains the driving data from 60 drivers. Then Das et al. (*20*) investigated gaps in lane-changing with a much more comprehensive dataset, the Strategic Highway Research Program 2 (SHRP 2), and demonstrated that different factors, including relative speed, traffic conditions, and acceleration, will influence gap acceptance decisions. Yet, such efforts in analyzing lane-changing behavior are hampered by the complication and difficulty of accurately capturing the longitudinal and lateral position of the lane-changing and surrounding vehicles and identifying the lane boundaries from a naturalistic driving dataset. As for the efforts from vehicle trajectories data, Wang et al. (*21*) utilized NGSIM (*22*) dataset and designed heuristic rules to extract discretionary lane changes. Later Li et al. (*23*) coupled a control model with trajectories to infer common discretionary lane change steering characteristics from NGSIM data. However, researchers are advised not to place excessive trust in the lateral position data provided in the raw NGSIM database, and Coifman and Li (*24*) indicated that refinement on the lateral position from the NGSIM dataset is needed. A recent study by Li

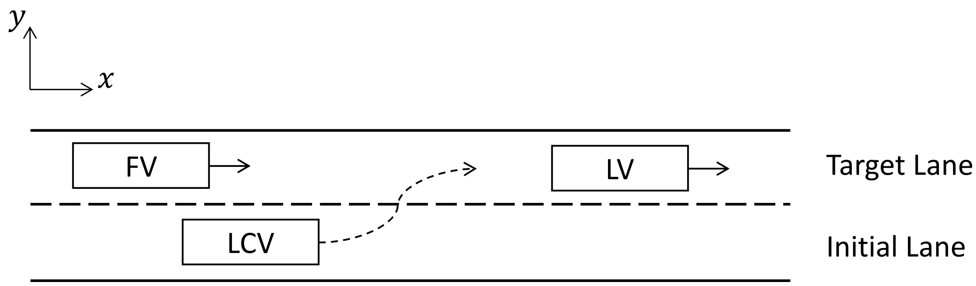

FIGURE 1: A demonstration of lane-changing decision

et al. (*25*) used the HighD dataset (*26*) to analyze how drivers' heterogeneity impact lane-changing duration. The HighD dataset is considered more accurate in lateral movements (*26*), but each drone only covers 420 meters of fast-flowing highway traffic in Germany. Therefore, each vehicle will only appear in the frame for a very short period. However, some multi-regime lane-changing models need longer trajectories across traffic states for calibration and validation.

One missing aspect from the previous studies is the categorization and characterization of discretionary lane-changing behaviors based on real-world data, and the critical bottleneck that impedes the development in analyzing lane-changing behaviors is the lack of abundant and accurate datasets that can capture lateral movements accurately in traffic flow. Therefore, the objective of this paper is to conduct an exploratory investigation on discretionary lane-changing characteristics with the Third Generation SIMulation (TGSIM) Dataset (*27*) and gain insights on how different types of lane-changing maneuvers should be categorized. These findings can provide insights into modeling lane-changing behavior under different traffic states and improve microscopic traffic simulation's soundness.

The paper is organized as follows. The next section elaborates on the data preparation steps, then followed by the methodologies in this study: a dynamic time Warping (DTW) analysis on lane-changing behaviors and an affinity propagation clustering method based on the previous DTW analysis. This paper proceeds with a section presenting the results of characterizing lane-changing behaviors with gaps and relative speeds. Finally, this paper is summarized with some conclusions and discussions.

DATA PREPARATION

This section will first briefly introduce the TGSIM dataset and then describe the extraction process of lane-changing cases, followed by presenting some descriptive analysis of the lane-changing measurements and finally, perform hypothesis tests to justify the benefit of splitting the behavior measurements into leading and following when categorizing and characterizing discretionary lane-changing behaviors.

The data used in this paper is collected from a 3-mile segment on I294 near Hinsdale, IL. A helicopter followed a fleet of three vehicles with activated adaptive cruise control. The helicopter was equipped with a RED camera at 30 frames per second at 8K resolution and moved at 1000 feet. The dataset contains 10 experiment runs, each covering four lanes in the same direction. By introducing the idea of "auxiliary lanes" and excluding all the movements from or towards the outside of the four lanes on the main traffic, only the lane changes among the four lanes in the mainstream are considered in this study. Since the merging from the on-ramp and diverging toward off-ramp cases are removed, all the lane changes are regarded as discretionary lane changes in this study. Ammourah et al. (*27*) has more information on data collection trajectory extraction in detail.

One advantage of TGSIM is that all the boundaries of lanes were carefully marked, and each trajectory point was assigned a lane index after trajectory extraction. As TGSIM data used the center of a vehicle as the proxy of its location, when the lane index changes for the same vehicle index, the center of this vehicle passes the lane markings. Then according to recent work by Ali et al. (*28*), a time window of 6 seconds is preferable for evaluating lane-changing models. Thus, this paper took 3 seconds before and 3 seconds after the lane index changes for a vehicle, then each lane-changing case has 6 seconds of data, accordingly.

Identifying the leader and follower is essential to calculate gaps and relative speeds. Regarding the lane assignments to every trajectory point, taking the leader and follower is straightforward when the lane-changing vehicle proceeds to the target lane. This paper only considers the cases when both the leader and follower interact with the lane-changing vehicle. Therefore, if the longitudinal distance between the leader/follower and the lane-changing vehicle is larger than 500m when the center of the lane-changing vehicle crosses the lane boundary, the lane-changing case will not be extracted into the final lane-changing dataset. There are 477 lane-changing cases after the extraction mentioned above, and each case contains 6 seconds of data. As the time resolution of TGSIM data is every 0.1 seconds, each lane-changing behavior measurement for a lane-changing case is a time series of length 60.

Figure 1 is an illustration of a typical lane-changing maneuver, and the definitions from Zheng (*1*) are utilized as a basis for constructing our lane-changing measurements. The lane-changing vehicle (LCV) changes from the initial lane to the target lane, and the leader (LC) and the follower (FV) are the immediately preceding and following vehicles, respectively. The location of vehicle i at time t is defined as the center of a vehicle, $(x_i(t), y_i(t))$, and this paper denotes the speed of vehicle i at time t as $v_i(t)$ and the length as l_i. Zheng (*1*) used the front-rear distance to represent the gaps, but the lane-changing cases in TGSIM dataset cover a wide range of speeds, from 22.22 m/sec to 47.49m/sec, and a higher speed would lead to a larger gap measured with spatial distance. For the purpose of this paper in characterizing lane-changing decisions, using spatial gaps will induce bias in measuring the difference between lane-changing cases. Therefore, this paper follows the definition presented by Yang et al. (*19*) of the lead gap $g_i^{lead}(t)$ and lag gap $g_i^{lag}(t)$ for a lane-changing case j at time t and uses time to represent gaps as follows:

$$g_j^{lead}(t) = \frac{x_j^{LV}(t) - x_j^{LCV}(t) - l^{LV}}{v_j^{LCV}(t)}, \tag{1}$$

$$g_j^{lag}(t) = \frac{x_j^{LCV}(t) - x_j^{FV}(t) - l^{LCV}}{v_j^{FV}}. \tag{2}$$

The relative speeds are also useful in measuring lane-changing decisions (*19*) and are defined as follows:

$$\Delta v_j^{lead}(t) = v_j^{LCV}(t) - v_j^{LV}(t), \tag{3}$$

$$\Delta v_i^{lag}(t) = v_j^{LV}(t) - v_j^{FV}(t). \tag{4}$$

Table 1 shows the key descriptive statistics of the extracted lane-changing cases from the TGSIM I-294 dataset. The speed distribution of the lane-changing vehicles ranges from 1.68m/sec to 47.29m/sec and covers a broad spectrum of traffic states. To justify if splitting gaps and relative

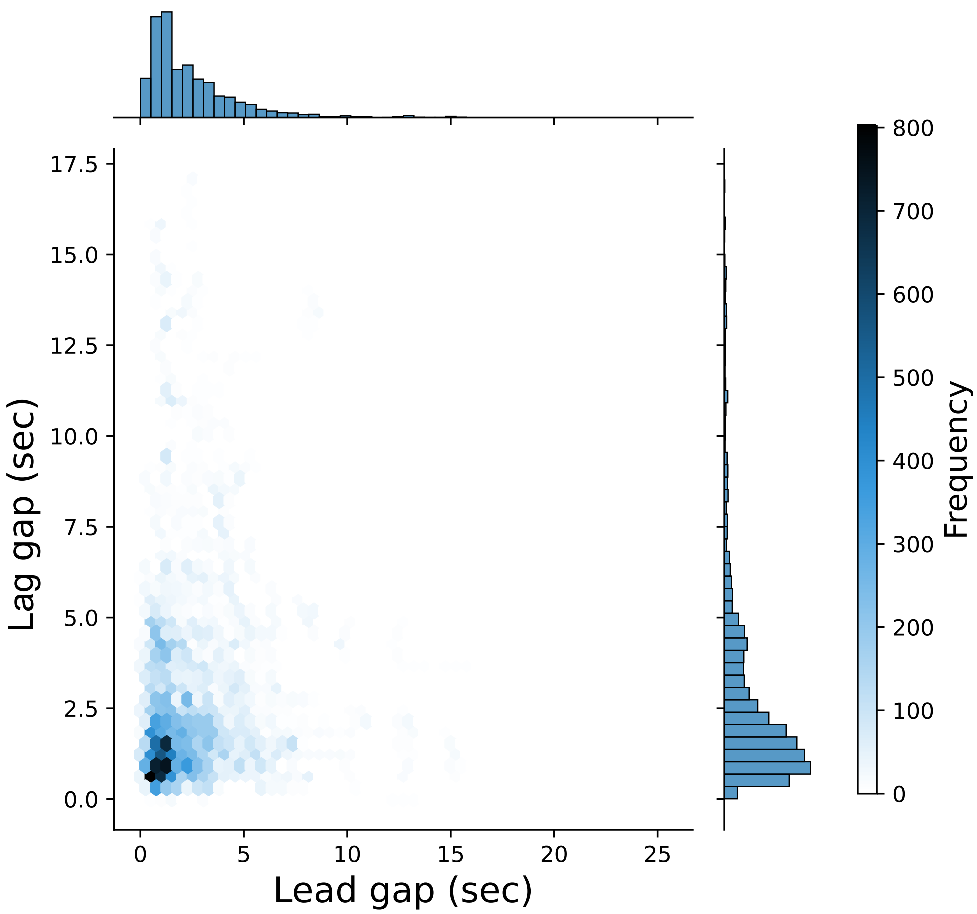

FIGURE 2: The distribution of lead and lag gaps

speed benefits modeling lane-changing decisions, hypothesis tests are helpful to identify whether measurements considering leaders or followers are drawn from the same distribution. Figure 2 shows the joint and marginal distributions of the follow and lead gaps, and they can be easily perceived as not normally distributed and right-skewed. Therefore, the two-sample Kolmogorov-Smirnow (K-S) test is performed with the null hypothesis H_0 as the follow and lead gap have the same distribution. The KS statistic is 0.0485, and the p-value is 0.0000, which rejected the null hypothesis. This result means the lead and lag gaps are not drawn from the same distribution under a 99% confidence level. Figure 3 shows the two relative speeds may share similar standard deviations but differ in the mean value. Following the same logic, t-test and K-S test results on lead and lag relative speed also show that the two measures are statistically significantly different from each other. Therefore, it is worthwhile to treat gaps and relative speed separately for leading and following directions in categorizing and characterizing discretionary lane-changing behaviors.

METHODOLOGY

This section presents two steps to categorize and characterize the discretionary lane-changing behaviors from real-world trajectory data. The first step is to construct a dynamic time warping

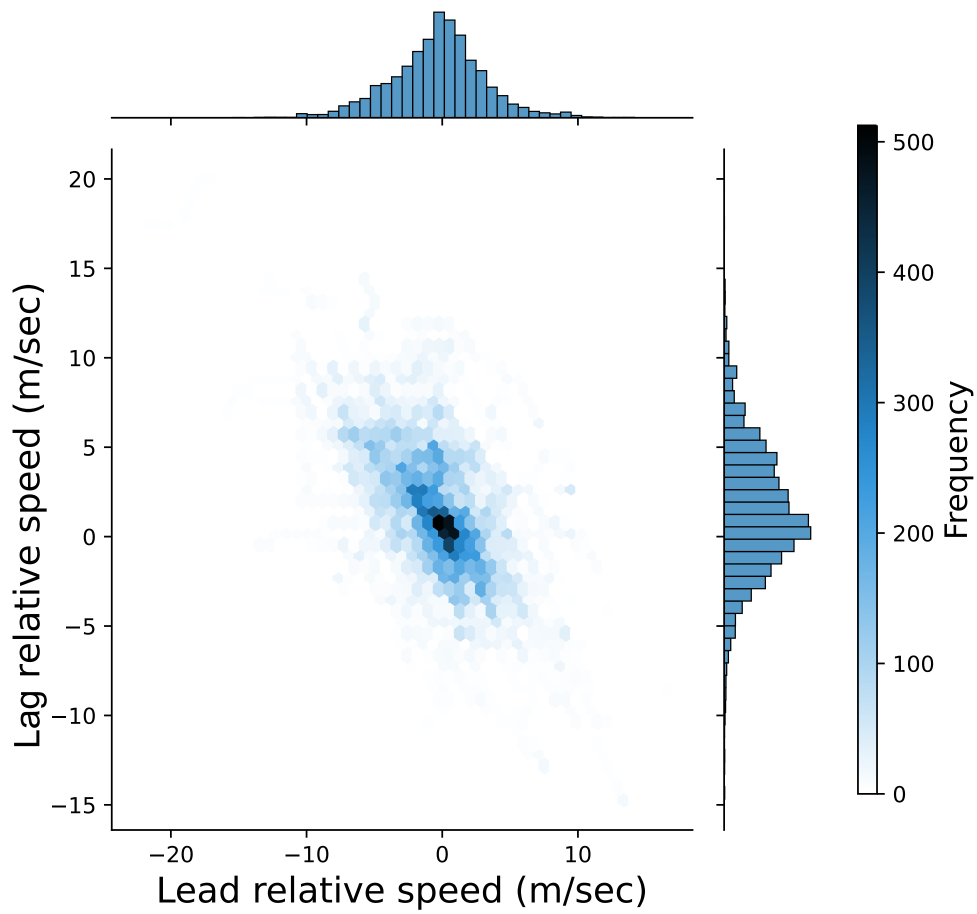

FIGURE 3: The distribution of relative speed compared to leader and follower

TABLE 1: Descriptive Statistics of Lane-changing Decision Parameters

Statistics	Speed (m/sec)	Acceleration (m/sec^2)	Gap (sec)		Relative Speed (m/sec)	
			Lead	Follow	Lead	Follow
Count	28620	28620	28620	28620	28620	28620
Mean	24.2242	0.0842	2.4498	2.7630	-0.3469	1.4374
Std. Dev	7.7736	1.0819	2.2600	2.6953	3.5086	3.7038
Min	1.6767	-9.9500	0.0008	0.0077	-22.4115	-14.6755
25%	19.1380	-0.2656	0.9782	1.0517	-2.3186	-0.8724
50%	25.8543	0.0600	1.6842	1.8450	-0.1772	1.1466
75%	30.0967	0.5000	3.2082	3.5727	1.6160	3.8336
Max	47.3890	12.6500	25.4276	17.0511	16.5015	19.9250

analysis framework to evaluate the similarity between any lane-changing cases, and the second step is to perform affinity clustering based on the dynamic time warping similarities to identify different categories in discretionary lane-changing behaviors.

Dynamic Time Warping (DTW) analysis on Lane-changing behaviors

The idea of using dynamic programming algorithms to find the matching patterns between time series data and evaluating the difference originated from the seminal work by Bellman and Kalaba (*29*), and then formally formulated and extensively explored and applied to the speech recognition tasks Myers et al., Sakoe and Chiba (*30, 31*). Recent studies using DTW analysis on driving behaviors by Hosseini et al., Talebpour and Zhang (*32, 33*) also demonstrated that the dynamic time warping method offers a distinct advantage due to its potential to evaluate time series with different lengths and finds the matching patterns between time series, rendering it an apt choice for driving behavior analysis.

To evaluate the difference between two time series of driving behaviors, $A = [a_1, a_2, \ldots, a_m]$ and $B = [b_1, b_2, \ldots, b_m]$, using dynamic time warping analysis, the local distance matrix $D \in R^{m\times n}$ is defined as

$$D \in R^{m\times n} : d_{ij} = \|a_i - b_j\| = \sqrt{(a_i - b_j)^2}, \qquad i \in [1:m], j \in [1:n] \tag{5}$$

where d_{ij} is an element in the local cost matrix D.

And the warping path $W = [w_1, w_2, w_3, ..., w_k, ..., w_K]$ represents a set of mapping relationships between A and B. An element in the warping path $w_k = (i_k, j_k) \in [1:m] \times [1:n]$ means a_{i_k} and b_{j_k} form a pair in the optimal matching.

Then Senin (*34*) introduced the following formulation:

$$DTW(A,H) = \min_{W} \sqrt{\sum_{k=1}^{K} d_{i_k j_k}^2}. \tag{6}$$

where $DTW(A,H)$ is the **DTW distance**, defined in (*34*) and $d_{i_k j_k}$ is the (i_k, j_k)-th elements in the local cost matrix D.

And the constraints are:

$$w_1 = (1,1), \tag{7}$$

$$w_K = (m,n), \tag{8}$$

$$(i' - i) \leq 1, \forall w_k = (i,j), w_{k+1} = (i', j'), \tag{9}$$

$$(j' - j) \leq 1, \forall w_k = (i,j), w_{k+1} = (i', j'), \tag{10}$$

$$(i' - i) \geq 0, \forall w_k = (i,j), w_{k+1} = (i', j'), \tag{11}$$

$$(j' - j) \geq 0, \forall w_k = (i,j), w_{k+1} = (i', j'), \tag{12}$$

The six constraints described above embody three fundamental assumptions initially proposed by Sakoe and Chiba (*31*). These assumptions are as follows: (1) **Boundary assumption :** Equations (7) and (8) guarantee that the warping path begins at the first points and ends at the last points of the two time series, representing an alignment assumption in DTW. (2) **Continuous assumption:** Equations (9) and (10) ensure a match with neighboring points, indicating that every time step should be included in the optimal warping path.; and (3) **Monotonous assumption:** Equations (11) and (12) preserve the time order and effectively prevent time from moving backward.

It has been proven by Senin (*34*) that the optimization problem defined by Equations (6) through (12) reduced to a shortest path problem in the cumulative distance matrix $C \in R^{m\times n}$ to compute the warping path K, and such shortest path problem can be solved with dynamic programming.

The pseudo-code for the dynamic programming algorithm in computing the cumulative distance matrix $C \in R^{m\times n}$ described in (*34*) is presented in Algorithm 1.

Algorithm 1 CummulativeDistanceMatrix(A,B,D)

$m \leftarrow \|A\|$
$n \leftarrow \|B\|$
New array $C[1...m, 1...n]$
Initialize $C[1,1] = 0$
for $i = 1; i \leq m; i++$ **do**
  $C[i,1] \leftarrow C[i-1,1] + D[i,1]$
end for
for $j = 1; j \leq n; j++$ **do**
  $C[1,j] \leftarrow C[1,j-1] + D[1,j]$
end for
for $i = 1; i \leq m; i++$ **do**
  for $j = 1; j \leq n; j++$ **do**
    $C[i,j] \leftarrow D[i,j] + \min\{C[i-1,j-1], C[j-1,j], C[i,j-1]\}$
  end for
end for
Return C

It takes $O(mn)$ to compute the local distance matrix D and the cumulative distance matrix C (*34*). Then given the cumulative distance matrix, the optimal warping path W can be recovered in $O(n)$ time by tracing back from $C[m,n]$ to get the warping path W.

Since the time series measured in this study are identical in length (all lane-changing cases are in 6 seconds), a length-based normalization conducted in (*33*) is unnecessary.

To better illustrate the process and output of dynamic time warping analysis, the speed of the lane-changing vehicle in two lane-changing cases is taken as a sample. Figure 4, as an exemplification, showcases its accuracy in finding the warping path and, thus, the matching patterns by minimizing the cumulative costs. Figure 5 Shows the corresponding matching patterns based on the warping path shown with dark blue in Figure 4. The black line is the speed for the vehicle ID=7 in experiment Run 7 when changing lanes from 4 to 5; the blue line is the speed for the vehicle ID=35 in Run 7 changing lanes from 4 to 3. The speed axis has an offset of 2 units to illustrate the matching patterns better.

In accordance with the established framework, a DTW distance matrix corresponding to lane-changing behavior measurements can be computed. The obtained difference measurements can subsequently be converted into similarity by inverting the value of each element. These measurements thus serve as an input to the clustering algorithms presented in the following subsection.

Affinity Propagation (AP) clustering based on DTW

This research aims to categorize discretionary lane-changing behaviors using real-world trajectory data. Current classification standards for lane-changing behaviors are inadequately defined for appropriately labeling the extracted data, underscoring the need for a clustering method. Given the exploratory nature of this study and the indeterminate number of clusters, clustering methods such as k-means (*35*) and spectral clustering (*36*), which necessitate prior knowledge of the number of clusters, do not align with the study objectives. DBSCAN, another prevalent unsupervised method,

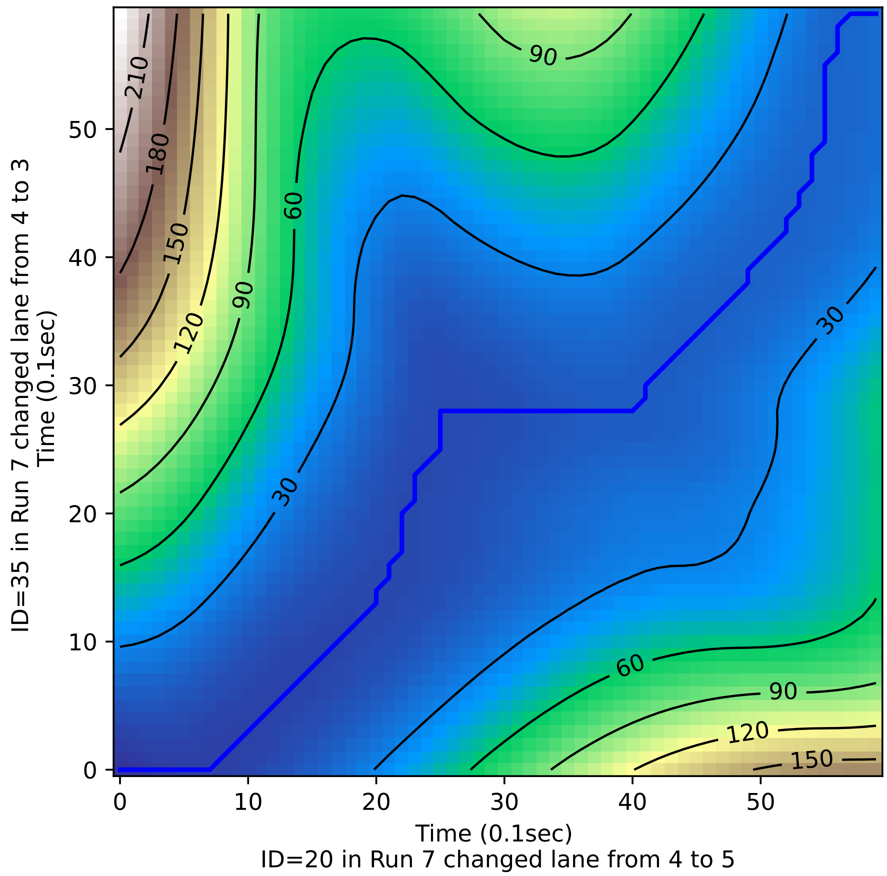

FIGURE 4: A demonstration of warping path computation calculating the distance between two speed time series data

is not ideal due to its inherent limitation of potentially classifying some data points as outliers not belonging to any cluster Ester et al. (*37*). This runs counter to our aim of achieving a comprehensive categorization of discretionary lane changes across various traffic states. Consequently, we have chosen affinity propagation (AP) proposed by Frey and Dueck (*38*) as the clustering method for our analysis due to its suitability for our research needs, providing unbiased categorization of discretionary lane-changing behaviors.

One of the key inputs of AP clustering is the similarity matrix. Inspired by Akl and Valaee (*39*) successful implementation of DTW on multi-dimensional acceleration in gesture recognition using AP clustering, this work designed the similarity matrices as follows:

$$S(i,j) = \begin{cases} -DTW^2(X_i^{lead}, X_j^{lead}) - DTW^2(X_i^{follow}, X_j^{follow}), i \neq j \\ p, i = j \end{cases} \tag{13}$$

where $DTW(X_{lead})$ and $DTW(X_{follow})$ are the dynamic time warping distance for leader-related factors and follower-related factors, respectively. The factors include gaps and relative speed, defined in $(X_j^{lead}, X_j^{follow}) \subset \{(G_j^{lead}, G_j^{follow}), (\Delta V_j^{lead}, \Delta V_j^{follow})\}$,

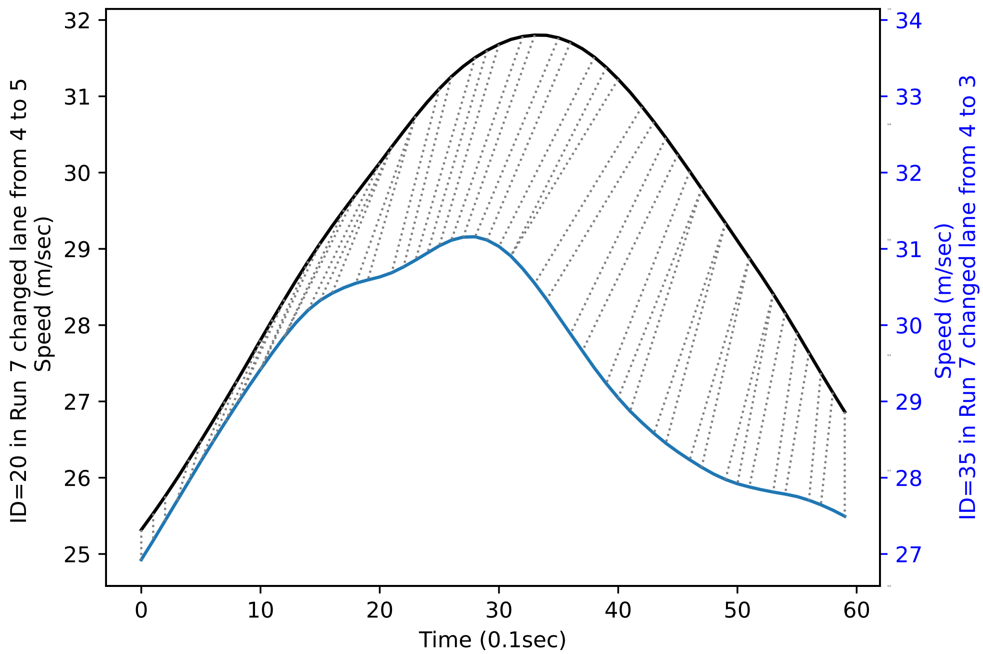

FIGURE 5: A demonstration of matching patterns between two speed time series data based on local cost minimization

where $G_j^{lead} = [g_j^{lead}(1), g_j^{lead}(2), ..., g_j^{lead}(n)]$ and $\Delta V_j^{follow} = [\Delta v_j^{follow}(1), v_j^{follow}(2), ..., v_j^{follow}(n)]$. p is the preference for each data point; points with a larger preference are more likely to be chosen as exemplars. An exemplar means cluster centers (*40*). Therefore, p should be initialized as an array with identical elements to maintain impartiality during clustering.

Affinity propagation can be viewed as exchanging messages between the data points and the message, including the responsibilities $r(i,k)$ and availabilities $a(i,k)$ between point i and k. The update function of responsibilities and availabilities are defined as follows (*38*):

$$r_{new}(i,k) = \lambda r_{old}(i,k) + (1-\lambda)(S(i,k) - \max_{j,j\neq k}\{a(i,j)+S(i,j)\}) \tag{14}$$

$$a_{new}(i,k) = \begin{cases} \lambda a_{old}(i,k) + (1-\lambda)(\min\{0, r(k,k) + \sum_{j,j\neq i,j\neq k} \max\{0, r(j,k)\}\}), i \neq k \\ \lambda a_{old}(i,k) + (1-\lambda)(\sum_{j,j\neq k} \max\{0, r(j,k)\}), i = k \end{cases} \tag{15}$$

where λ is a damping factor between 0 and 1. This is to avoid unstable dynamics in practice (*38*).

The responsibilities and availabilities at any step can be used to identify exemplars. For lane-changing case i, let $k = argmax_j\{a(i,j) + r(i,j)\}$ if $k = i$ then i is an exemplar; and if $k \neq i$, then k is an exemplar for i. Algorithm 2 describes the AP clustering implementation proposed by Frey and Dueck (*38*). The similarity matrix S is computed based on Equation 13, and preference p is a function of the median of the similarity matrix S and how to determine preference will be discussed in detail in the following RESULTS section. Other parameter keeps fixed values in this study, including the damping factor $\lambda = 0.5$, the maximum number of iterations, $max\_iter = 200$, and the number of iterations with no change in exemplars that stops the convergence, $same\_exemplar = 15$.

Algorithm 2 Affinity Propagation Clustering($S, p, \lambda, max\_iter, conv\_iter$)

```
num_iter ← 1
same_exemplar_iter ← 0
while True do
    Compute responsibilities based on (14)
    Computer availabilities based on (15)
    if exemplars changed then
        same_exemplar_iter ← 0
    else
        same_exemplar_iter ← same_exemplar_iter + 1
    end if
    if num_iter > max_iter or same_exemplar_iter > conv_iter then
        break
    end if
    num_iter ← num_iter + 1
end while
Find exemplars and assign each data point to the nearest exemplar to finalize clusters
```

Given that the ground truth for the clustering is unknown, to assess the performance of the clustering, this paper employs two widely-used metrics suitable for situations where the clustering is not known. The first is the average Silhouette Coefficient (*41*), ASC, defined as follows,

$$ASC = \frac{1}{N}\sum_{i}^{N}\frac{b_i - a_i}{max(a_i, b_i)} \tag{16}$$

where N is the number of samples, and $N = 477$ in this study; a_i is the mean dynamic time warping distance between sample i and all other samples in the same cluster; b_i is the mean dynamic warping distance between sample i and all other samples in the next nearest cluster.

A higher average Silhouette Coefficient score means a model with better-defined clusters. The range is between -1 for incorrect clustering and +1 for highly dense clustering, and a score around zero indicates the clusters have too much overlapping.

The second measurement of clustering performance is the Calinski-Harabasz Index (*42*), CHI. For a set of data D of size N is defined as follows,

$$CHI = \frac{tr(B_k)}{tr(W_k)} \times \frac{N-k}{k-1}, \tag{17}$$

$$B_k = \sum_{q=1}^{k} n_q (c_q - c_D)(c_q - c_D)^T, \tag{18}$$

$$W_k = \sum_{q=1}^{k}\sum_{x \in C_q}(x - c_q)(x - c_q)^T. \tag{19}$$

where C_q is the set of points in cluster q; c_q is the center of cluster q; C_D is the set of points in cluster D; and n_q is the number of samples in cluster q.

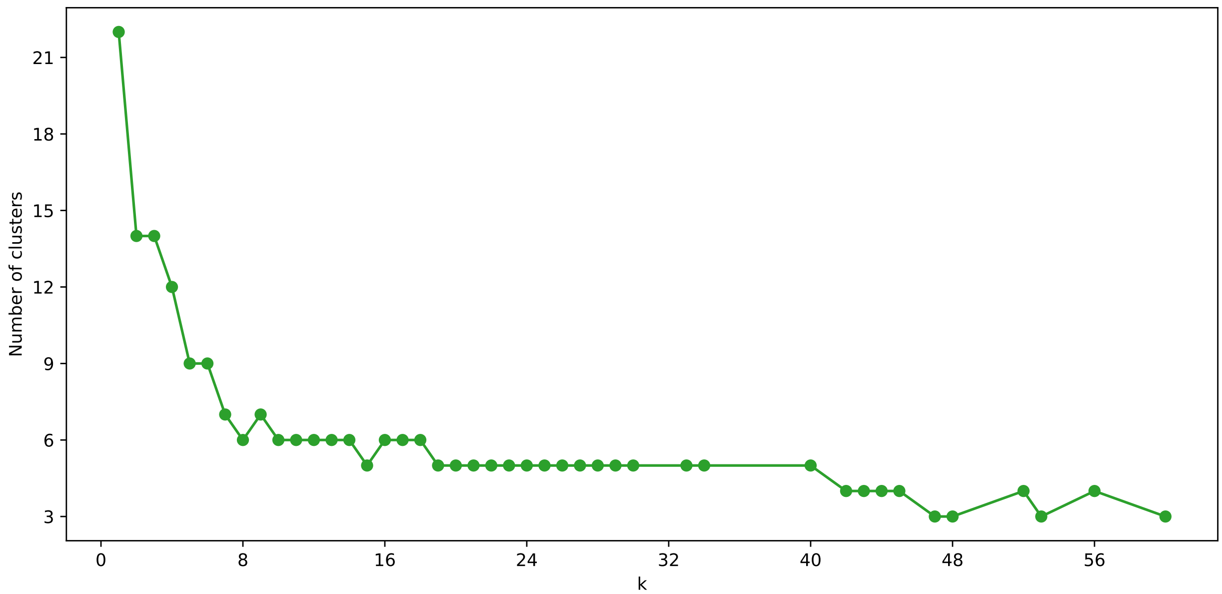

FIGURE 6: Number of clusters based on gaps with different *k* value

A higher *CHI* relates to a more dense and well-separated set of clusters. And both the average Silhouette Coefficient and the Calinski-Harabasz Index can be evaluated using the machine learning python package, scikit-learn (*43*), with a precomputed similarity matrix using dynamic time warping analysis presented in the previous subsection.

RESULTS

The primary goal of this study is to categorize discretionary lane-changing behaviors. The methodologies presented in the previous sections facilitate a nuanced analysis of lane-changing behavior patterns, drawing upon the principles of similarity measurement and cluster analysis. Then this section shows the categorization results and interpretation of how discretionary lane-changing should be modeled.

Characterizing lane-changing decision based on gaps

This subsection outlines the process of selecting the optimal parameter for affinity propagation clustering. This choice gives rise to a discussion on the implications of the smallest number of clusters in relation to modeling lane-changing behaviors considering driver heterogeneity. The subsection then culminates with a detailed classification and characterization of lane-changing behaviors, considering lead and lag gaps.

The AP clustering algorithm takes a preference value as input to denote the likelihood of a sample being chosen to be one of the exemplars, and preference is often defined as an integer multiple of the median of the similarity matrix. Therefore, this paper evaluated the performance score with k-multiple of the median in the similarity matrix defined with lead and lag gaps. k is an integrated value ranging from 1 to 60, and the non-converging cases with some k values are excluded.

Figure 6 shows the number of clusters decreases from 22 as k increases, then becomes stable at five clusters, and finally, as k is larger than 40, the number of clusters will become even

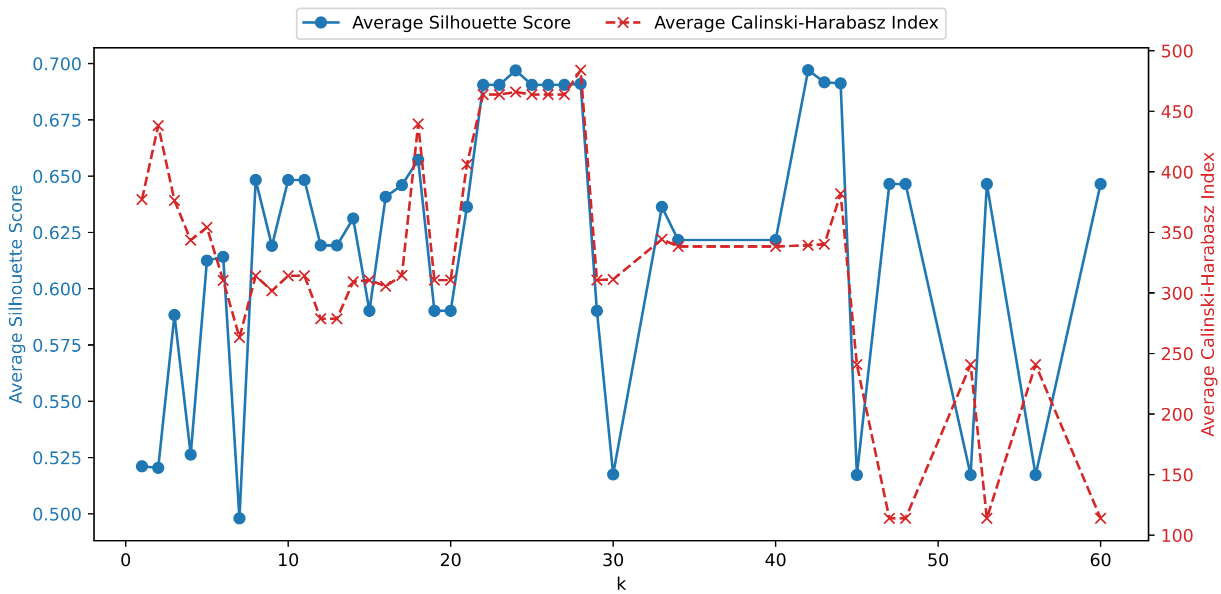

FIGURE 7: Clustering performance based on gaps with different *k* value

less to three and four. Figure 7 shows the variation of the two performance metrics as k changes and clearly demonstrates that both metrics peak when k equals 28; thus, the preference is selected to be 28 multiplied by the median of the similarity matrix with lead and lag gaps, which will lead to five clusters in lane-changing gap acceptance. It is clear from Figure 7 that the average Calinski-Harabasz index becomes significantly less when k is larger than 40, which indicates that less than five clusters in modeling lane-changing with lead and lag gaps may be too simplified to capture important heterogeneity among drivers.

Figure 8 presents the joint and marginal distributions of the five clusters. The discretionary lane-changing lead gaps can be categorized into three distinct regions: (i) aggressive: 0 to 4 seconds; (ii) neutral: 4 to 8 seconds; and (iii) cautious: larger than 8 seconds. Similarly, the lag gaps can also be segmented into three zones: (i) aggressive: 0 to 2.5 seconds; (ii) neutral: 2.5 to 7.5 seconds; and (iii) cautious: larger than 7.5 seconds.

From this, the five clusters within the discretionary lane-changing gaps can be interpreted. Cluster 1 represents aggressive lane-changing behavior in both leading and following directions. While Clusters 2 and 5 exhibit aggressive behavior in the lag gaps, they are neutral and cautious, respectively, when selecting lead gaps. Conversely, Clusters 4 and 3 embody lane-changing behaviors that are aggressive in lead gap selection but neutral and cautious, respectively, on the following end.

Characterizing lane-changing decision based on relative speeds

Adopting a similar methodology for selecting the preference parameter in affinity propagation clustering for lead and lag gaps, the number of clusters, based on relative speeds in the leading and following directions, is depicted in Figure 9. The number of clusters appears to be on a declining trend as k increases and stabilizes at two clusters once k surpasses 39. Figure 7 provides a distinct demonstration of enhanced performance when k is greater than 40. As such, a k-value

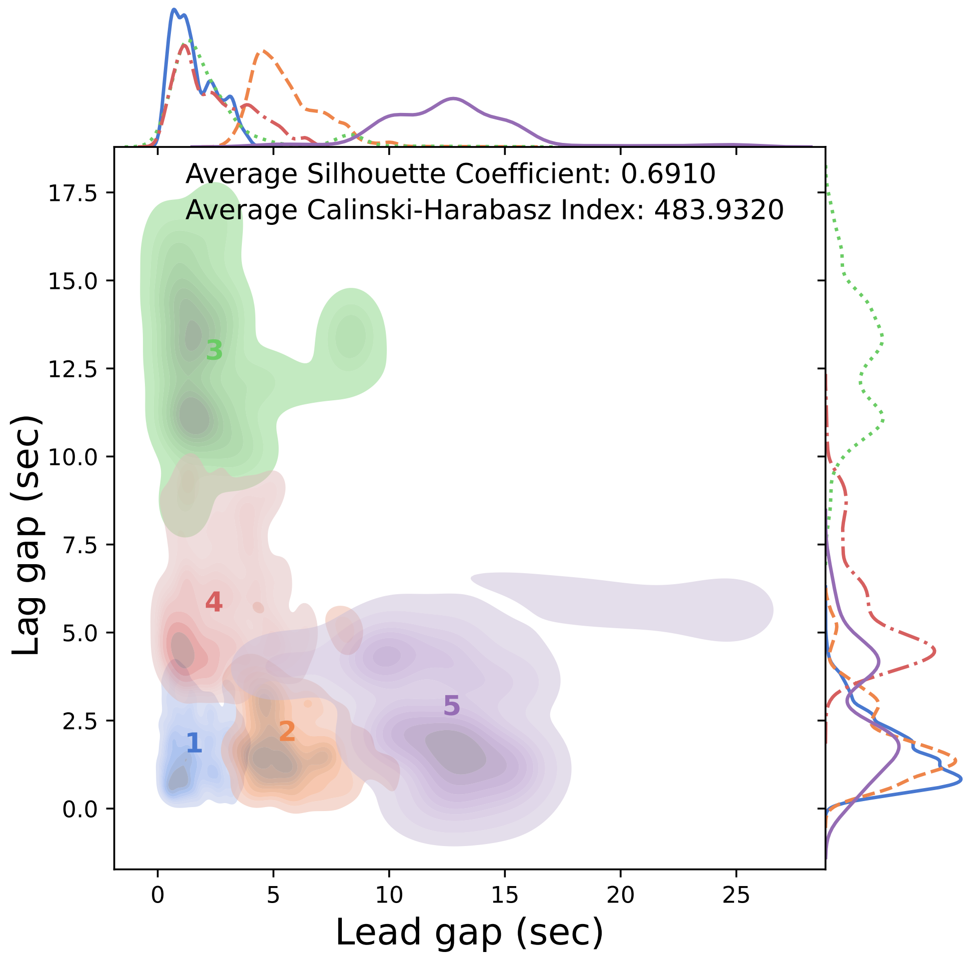

FIGURE 8: Clustering results on lead and lag gap

of 52 is chosen, resulting in two clusters based on the relative speeds in the leading and following directions, forming the basis for discretionary lane-changing behaviors.

Figure 11 depicts the joint and marginal distributions of two clusters based on the lead and lag relative speed. Cluster 1 demonstrates a negative lead relative speed juxtaposed with a positive lag relative speed, suggesting a scenario where a driver aims to surpass the following vehicle in the target lane while avoiding an overly rapid approach toward the leading vehicle. Conversely, Cluster 2 comprises cases with both lead and lag relative speeds approximating zero. This represents instances where the vehicle changing lanes maintain a similar speed to the vehicles in the target lane.

These findings indicate that the similarity measurement derived from dynamic time warping analysis and affinity propagation proposed in the METHODOLOGY section indeed offers an advantage in yielding insights into discretionary lane-changing behaviors.

CONCLUSIONS

In conclusion, this study has shown that a data-driven approach can significantly enhance our understanding of discretionary lane-changing behaviors. The lack of high-resolution data to ac-

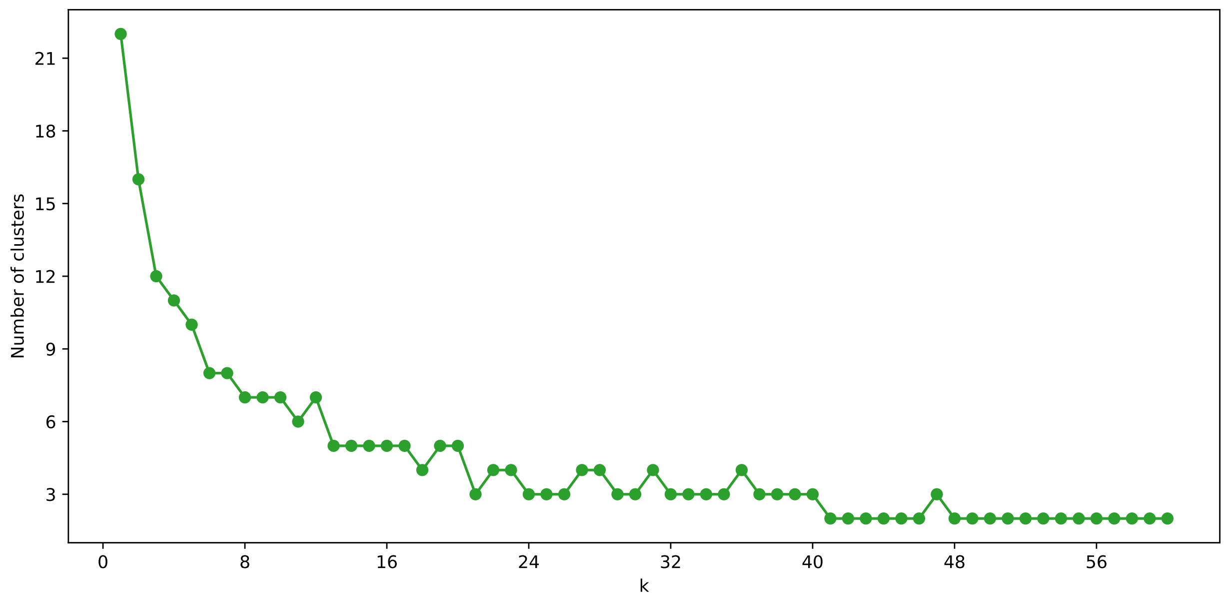

FIGURE 9: Number of clusters based on relative speeds with different *k* value

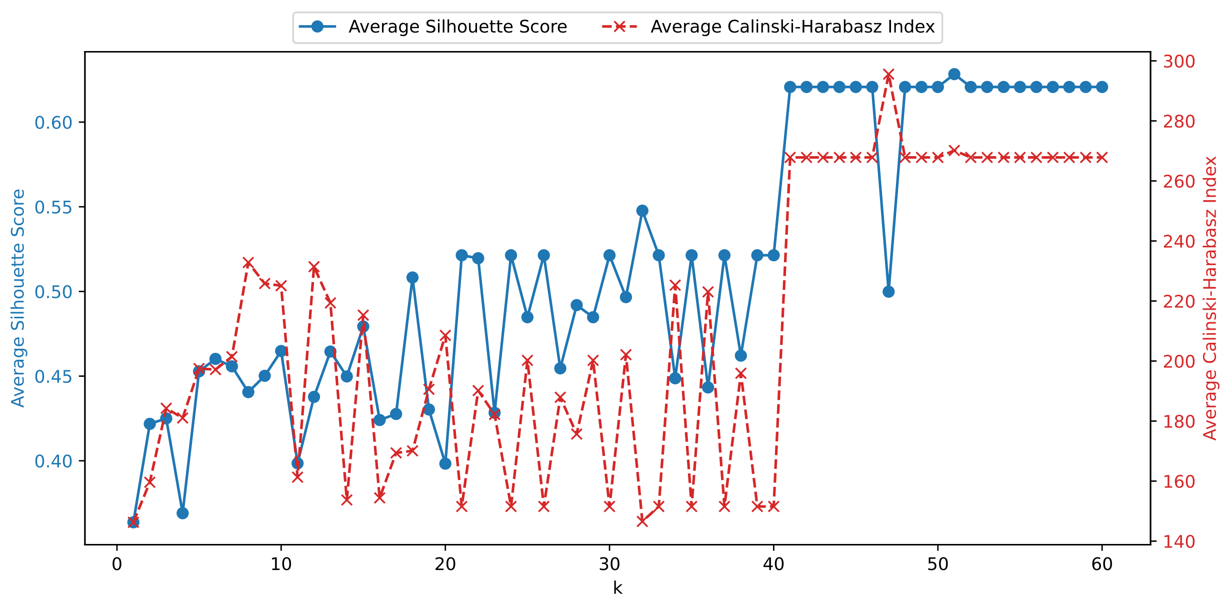

FIGURE 10: Clustering performance based on relative speeds with different *k* value

curately capture lateral movements has long been an impediment to achieving meaningful insight into these behaviors. However, leveraging the Third Generation SIMulation (TGSIM) dataset, our research provides a breakthrough in this area.

This study first developed a discretionary lane-changing extraction pipeline and conducted a descriptive investigation of critical factors such as gaps and relative speeds. Hypothesis tests on if and relative speed should be split into leading and following directions, and results show that analyzing leading and following separately is necessary for modeling lane-changing behaviors.

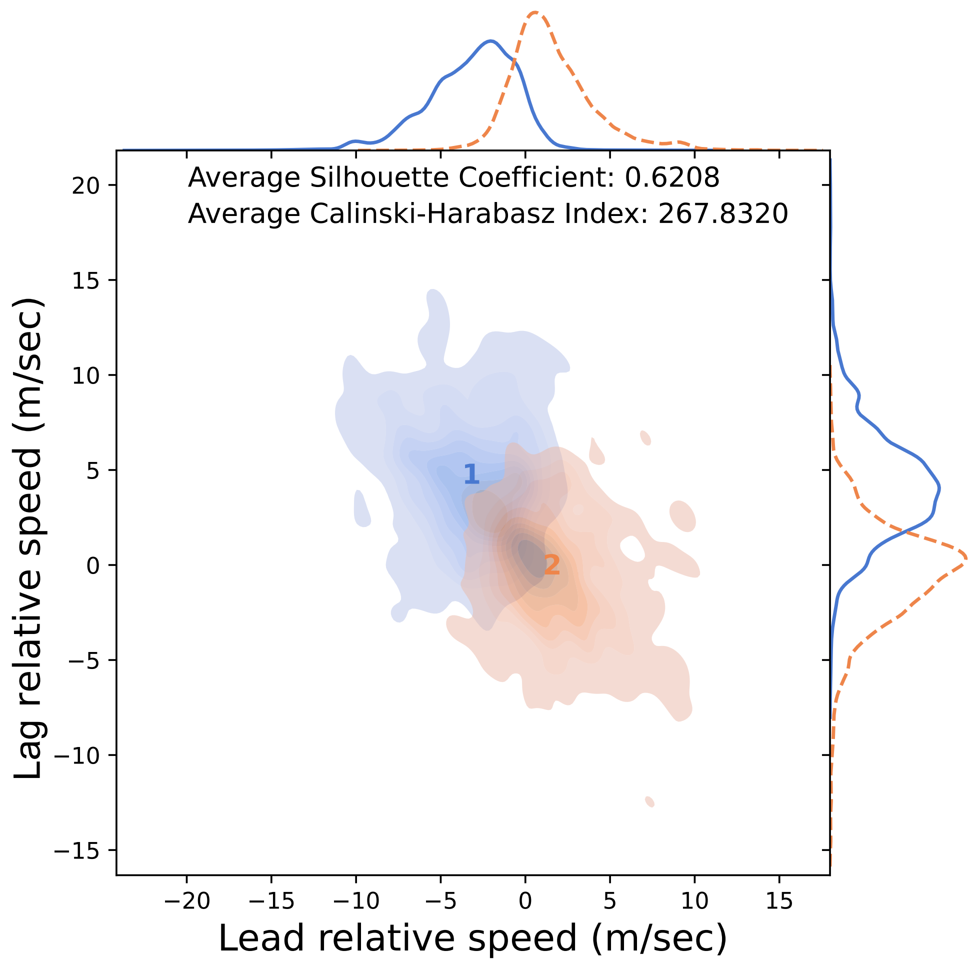

FIGURE 11: Clustering results on relative speed

Then, Our methodological design employing Dynamic Time Warping (DTW) analysis and Affinity Propagation (AP) clustering offers a robust approach to scrutinize and quantify driving dynamics in discretionary lane-changing. The DTW analysis plays a pivotal role in quantifying the differences between an array of lane-changing behaviors, and subsequently, the AP clustering allows us to delve deeper into categorizing and characterizing discretionary lane changes.

For the evaluation and guidance in selecting the hyperparameter, the preference in affinity propagation and two clustering performance metrics are employed, namely the average Silhouette Coefficient and the Calinski-Harabasz Index. The emergence of five distinct clusters from the data underscores the effectiveness of our methodology, particularly in categorizing lane-changing behaviors into aggressive, neutral, and cautious types in both leading and following directions. Our analysis also revealed two distinctive groups based on relative speeds, one representing overtaking and the other depicting a transition into a lane with stable and homogeneous speed.

To summarize, our research has illustrated that the combination of Dynamic Time Warping (DTW) analysis and Affinity Propagation (AP) clustering offers a versatile toolset for exploring driving behavior. The inherent adaptability of this approach holds promise for a multitude of contexts. Coupled with providing a more abundant and accurate data set, it could mark the advent of a new era in data-centric research within the realm of traffic flow studies.

Moving forward, it would be worthwhile to extend this study to incorporate aspects such as automated lane-changing and the spatial trajectories in two dimensions associated with lane changes. This will provide a more comprehensive understanding of driving behaviors.

ACKNOWLEDGEMENTS

The authors would like to express their sincere gratitude to all individuals involved in collecting and processing the TGSIM dataset. The authors also extend their appreciation to the Federal Highway Administration (FHWA) for their invaluable support in this dataset.

REFERENCES

1. Zheng, Z., Recent developments and research needs in modeling lane changing. *Transportation research part B: methodological*, Vol. 60, 2014, pp. 16–32.
2. Pande, A. and M. Abdel-Aty, Assessment of freeway traffic parameters leading to lane-change related collisions. *Accident Analysis & Prevention*, Vol. 38, No. 5, 2006, pp. 936–948.
3. van Winsum, W., D. De Waard, and K. A. Brookhuis, Lane change manoeuvres and safety margins. *Transportation Research Part F: Traffic Psychology and Behaviour*, Vol. 2, No. 3, 1999, pp. 139–149.
4. Kerner, B. S. and H. Rehborn, Experimental features and characteristics of traffic jams. *Physical Review E*, Vol. 53, No. 2, 1996, p. R1297.
5. Mauch, M. and M. J. Cassidy, Freeway traffic oscillations: observations and predictions. In *Transportation and Traffic Theory in the 21st Century: Proceedings of the 15th International Symposium on Transportation and Traffic Theory, Adelaide, Australia, 16-18 July 2002*, Emerald Group Publishing Limited, 2002, pp. 653–673.
6. Ahn, S. and M. J. Cassidy, Freeway traffic oscillations and vehicle lane-change maneuvers. *Transportation and Traffic Theory*, Vol. 1, 2007, pp. 691–710.
7. Zheng, Z., S. Ahn, D. Chen, and J. Laval, Freeway traffic oscillations: microscopic analysis of formations and propagations using wavelet transform. *Procedia-Social and Behavioral Sciences*, Vol. 17, 2011, pp. 702–716.
8. Cassidy, M. J. and J. Rudjanakanoknad, Increasing the capacity of an isolated merge by metering its on-ramp. *Transportation Research Part B: Methodological*, Vol. 39, No. 10, 2005, pp. 896–913.
9. Yang, Q. I. 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, 1996, pp. 113–129.
10. Gipps, P. G., A behavioural car-following model for computer simulation. *Transportation research part B: methodological*, Vol. 15, No. 2, 1981, pp. 105–111.
11. Kesting, A., M. Treiber, and D. Helbing, General lane-changing model MOBIL for car-following models. *Transportation Research Record*, Vol. 1999, No. 1, 2007, pp. 86–94.
12. Treiber, M., A. Hennecke, and D. Helbing, Congested traffic states in empirical observations and microscopic simulations. *Physical review E*, Vol. 62, No. 2, 2000, p. 1805.
13. Ahmed, K., M. Ben-Akiva, H. Koutsopoulos, and R. Mishalani, Models of freeway lane changing and gap acceptance behavior. *Transportation and traffic theory*, Vol. 13, 1996, pp. 501–515.
14. Toledo, T., H. N. Koutsopoulos, and M. Ben-Akiva, Integrated driving behavior modeling. *Transportation Research Part C: Emerging Technologies*, Vol. 15, No. 2, 2007, pp. 96–112.
15. Kita, H., 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, 1999, pp. 305–312.
16. Ban, J. X., A game theoretical approach for modelling merging and yielding behaviour at freeway on-ramp sections. In *Transportation and Traffic Theory: Papers Selected for Presentation at 17th International Symposium on Transportation and Traffic Theory, a Peer Reviewed Series Since 1959*, Elsevier Amsterdam, The Netherlands, 2007, p. 197.

17. Talebpour, A., H. S. Mahmassani, and S. H. Hamdar, Modeling lane-changing behavior in a connected environment: A game theory approach. *Transportation Research Procedia*, Vol. 7, 2015, pp. 420–440.

18. Ali, Y., Z. Zheng, M. M. Haque, and M. Wang, A game theory-based approach for modelling mandatory lane-changing behaviour in a connected environment. *Transportation research part C: emerging technologies*, Vol. 106, 2019, pp. 220–242.

19. Yang, M., X. Wang, and M. Quddus, Examining lane change gap acceptance, duration and impact using naturalistic driving data. *Transportation research part C: emerging technologies*, Vol. 104, 2019, pp. 317–331.

20. Das, A., M. N. Khan, and M. M. Ahmed, Nonparametric multivariate adaptive regression splines models for investigating lane-changing gap acceptance behavior utilizing strategic highway research program 2 naturalistic driving data. *Transportation research record*, Vol. 2674, No. 5, 2020, pp. 223–238.

21. Wang, Q., Z. Li, and L. Li, Investigation of discretionary lane-change characteristics using next-generation simulation data sets. *Journal of Intelligent Transportation Systems*, Vol. 18, No. 3, 2014, pp. 246–253.

22. *Next Generation Simulation: US101 Freeway Dataset*. Federal Highway Administration, Washingtion, D.C., 2007.

23. Li, L., C. Lv, D. Cao, and J. Zhang, Retrieving common discretionary lane changing characteristics from trajectories. *IEEE Transactions on Vehicular Technology*, Vol. 67, No. 3, 2017, pp. 2014–2024.

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

25. Li, G., Z. Yang, Y. Pan, and J. Ma, Analysing and modelling of discretionary lane change duration considering driver heterogeneity. *Transportmetrica B: Transport Dynamics*, Vol. 11, No. 1, 2023, pp. 343–360.

26. Krajewski, R., J. Bock, L. Kloeker, and L. Eckstein, The highD Dataset: A Drone Dataset of Naturalistic Vehicle Trajectories on German Highways for Validation of Highly Automated Driving Systems. In *2018 21st International Conference on Intelligent Transportation Systems (ITSC)*, 2018, pp. 2118–2125.

27. Ammourah, R., P. Beigi, B. Fan, H. S. H., R. James, M. Khajeh-Hosseini, H. S. Mahmassani, D. Monzer, A. Talebpour, Y. Zhang, T. Radvand, and C.-C. Hsiao, Introduction to the Third Generation Simulation (TGSIM) Dataset: Data Collection and Trajectory Extraction, 2023.

28. Ali, Y., Z. Zheng, and M. C. Bliemer, Calibrating lane-changing models: Two data-related issues and a general method to extract appropriate data. *Transportation Research Part C: Emerging Technologies*, Vol. 152, 2023, p. 104182.

29. Bellman, R. and R. Kalaba, On adaptive control processes. *IRE Transactions on Automatic Control*, Vol. 4, No. 2, 1959, pp. 1–9.

30. Myers, C., L. Rabiner, and A. Rosenberg, Performance tradeoffs in dynamic time warping algorithms for isolated word recognition. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, Vol. 28, No. 6, 1980, pp. 623–635.

31. Sakoe, H. and S. Chiba, Dynamic programming algorithm optimization for spoken word

recognition. *IEEE transactions on acoustics, speech, and signal processing*, Vol. 26, No. 1, 1978, pp. 43–49.

32. Hosseini, M. K., A. Talebpour, S. Devunuri, and S. H. Hamdar, An unsupervised learning framework for detecting adaptive cruise control operated vehicles in a vehicle trajectory data. *Expert Systems With Applications*, Vol. 208, 2022, p. 118060.
33. Talebpour, A. and Y. Zhang, Characterizing Human-Automated Vehicle Interactions: an Investigation into Car-following Behavior, 2023.
34. Senin, P., Dynamic time warping algorithm review. *Information and Computer Science Department University of Hawaii at Manoa Honolulu, USA*, Vol. 855, No. 1-23, 2008, p. 40.
35. MacQueen, J. et al., Some methods for classification and analysis of multivariate observations. In *Proceedings of the fifth Berkeley symposium on mathematical statistics and probability*, Oakland, CA, USA, 1967, Vol. 1, pp. 281–297.
36. Shi, Multiclass spectral clustering. In *Proceedings ninth IEEE international conference on computer vision*, IEEE, 2003, pp. 313–319.
37. Ester, M., H.-P. Kriegel, J. Sander, X. Xu, et al., A density-based algorithm for discovering clusters in large spatial databases with noise. In *kdd*, 1996, Vol. 96, pp. 226–231.
38. Frey, B. J. and D. Dueck, Clustering by passing messages between data points. *science*, Vol. 315, No. 5814, 2007, pp. 972–976.
39. Akl, A. and S. Valaee, Accelerometer-based gesture recognition via dynamic-time warping, affinity propagation, & compressive sensing. In *2010 IEEE International Conference on Acoustics, Speech and Signal Processing*, IEEE, 2010, pp. 2270–2273.
40. Mézard, M., Where are the exemplars? *Science*, Vol. 315, No. 5814, 2007, pp. 949–951.
41. Rousseeuw, P. J., Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *Journal of computational and applied mathematics*, Vol. 20, 1987, pp. 53–65.
42. Caliński, T. and J. Harabasz, A dendrite method for cluster analysis. *Communications in Statistics-theory and Methods*, Vol. 3, No. 1, 1974, pp. 1–27.
43. Pedregosa, F., G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay, Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research*, Vol. 12, 2011, pp. 2825–2830.