A NOVEL CLUSTERING APPROACH TO IDENTIFY VEHICLES EQUIPPED WITH ADAPTIVE CRUISE CONTROL IN A VEHICLE TRAJECTORY DATA

Mohammadreza Khajeh Hosseini
Ph.D. Student
University of Illinois at Urbana-Champaign
205 North Mathews Ave. Urbana, IL 61801
Email: mk49@illinois.edu
Tel: +1 682 208 4187

Alireza Talebpour, Ph.D., Corresponding Author
Assistant Professor
University of Illinois at Urbana-Champaign
205 North Mathews Ave. Urbana, IL 61801
Email: ataleb@illinois.edu
Tel: +1 217 333 8038

Word Count: 5954 words + 4 table(s) $\times$ 250 = 6954 words

Submission Date: December 23, 2020

Submitted for presentation at the 100th Annual Meeting of the Transportation Research Board

ABSTRACT

Vehicle trajectories are one of the cornerstones of modern traffic flow theory with applications in driver behavior studies and automated vehicle (AV) research. The existing vehicle trajectory datasets are limited, mostly due to the high cost of data collection and preparation. Moreover, with the arrival of advanced driver assistance systems (ADAS) such as adaptive cruise control (ACC), there is a potential to see changes in human driving behavior when interacting with these technologies. The existing trajectory datasets fail to provide any information on the utilization of ADAS technologies. This study proposes using a new trajectory dataset that contains multiple instances of vehicles using ACC to identify ACC-type behavior across the entire trajectory dataset. Since the trajectory data is not labeled based on ACC utilization, clustering is an excellent approach to arrange similar trajectories in the dataset into the same group. Using this dataset combined with clustering, this study identifies the vehicle trajectories that have similar traffic dynamics to the vehicles using ACC.

Keywords: Vehicle Trajectory, Aerial Data Collection, Adaptive Cruise Control, Advanced Driver Assistant Systems

INTRODUCTION AND BACKGROUND

Vehicle trajectory is a concise way to store data of an individual or collective group of vehicles for both micro- and macro-level traffic analyses. Vehicle trajectories play a major role in traffic flow studies (including driver behavior, collision analysis, and in automated vehicle (AV) research). With the advancements in sensing and imaging technologies, the trajectories can be generated using cameras, infrared sensors, RADAR, and LiDAR. However, video-based imaging has been the most popular method of extracting vehicle trajectories. The early studies collected vehicle trajectories using pole-mounted cameras at intersections (*1*)(*2*). Aerial imagery for trajectory extraction overcomes issues related to occlusion and cluttering that are associated with using pole-mounted cameras. Satellites, helicopters, and airplanes are the conventional, but expensive to operate means of obtaining aerial videos and images. With the advent of Unmanned Aerial Vehicles (UAVs), the aerial data collection has become much cheaper and accessible.

Some of the existing vehicle trajectory datasets are FHWA Next Generation Simulation Models (NGSIM)(*3*), Strategic Highway Research Program (SHRP2)(*4*) and TrafficNet (*5*). NGSIM is a well-known open-source trajectory dataset collected in 2006 using digital cameras at different locations including US Highway 101 and Interstate 80 freeway. The vehicle trajectories are extracted from the images of multiple cameras combined to create a single image that looks like an aerial shot. The NGSIM trajectory data contains the location of each vehicle at a frequency of 10 Hz over a 1600 to 3200 feet stretch of roadway. However, the NGSIM data suffers from noise and inaccurate detection due to the low-resolution cameras at a considerable distance. Coifman and Li (*6*) analyzed the NGSIM dataset and confirmed inaccuracies in speed and positioning of vehicles. The SHRP2 dataset (*4*), in collaboration with Virginia Tech Transportation Institute (VTTI), had collected naturalistic driving data in 2012. The dataset includes more than 5 million trips that include sensory data such as speed, location, and acceleration, and also vehicle and driver characteristics. This dataset is not freely available to public access and has a limited preview. It is collected using probe vehicles, and the collected data is limited to the field of view of the onboard sensors and does not entirely define the surrounding vehicles and traffic dynamics. The TrafficNet (*5*) provides processed naturalistic data with libraries for researchers to perform data analytics. TrafficNet (*5*) separated driving into six scenarios such as free flow, car-following, cut-in, etc., and classified the entire dataset into these scenarios curated to research. It is a web-based platform, with MYSQL database used to store the information. HighD dataset (*7*) was published in 2018, with naturalistic vehicle trajectories recorded on German highways. It accounts for variability in traffic composition by collecting more data and at six different locations. It has a truck ratio varying from 0 - 50 % and trajectories collected at different times of the day. The trajectories are analyzed and classified into specific maneuver types such as lane changes and critical maneuvers. More recently, pNEUMA dataset (*8*), used swarm of drones to collect arterial traffic data in sequential sessions with blind gaps in between sessions. Their objective was to study Origin-Destination information, travel time, congestion propagation, and lane-changing behavior.

While the aforementioned datasets provide the means to analyze driver behavior, they fail to provide any information on the utilization of the Advanced Driver Assistant Systems (ADAS) by drivers. Utilizing these features by drivers can potentially change the interactions among drivers on the road and can lead to new traffic flow dynamics and possibly new types of high-risk driving instances. Considering that ADAS has a compound annual growth rate of 12% (*9*), it is critical to evaluate the impacts of ADAS technologies on driver behavior and traffic flow dynamics. To address the shortcomings of the aforementioned datasets in considering ADAS technologies,

Khajeh-Hosseini et al. (*10*) introduced a new trajectory dataset with multiple instances of vehicles using adaptive cruise control (ACC) (a common ADAS technology), collected by means of multiple UAVs. This trajectory dataset contains information on five platoons of three ACC operated vehicles mixed with human-driven vehicles on Interstate 35 in Austin, TX. While invaluable information can be extracted from this dataset, considering how widespread ACC systems has become in the past few years, there is a good chance that there exist other vehicles in this trajectory dataset that use ACC. Unfortunately, without such knowledge, capturing the full impacts of ACC system on traffic flow dynamics is impossible. There is a large body of literature on the design of ACC systems with different spacing policies and design criteria as well as comprehensive paper reviews such as Xiao and Gao (*11*) that are recommended for further study to interested readers. Many of the proposed ACC systems are evaluated based on computer simulation data, limited lab platform tests, and few studies such as (*12*) and (*13*) that are based on real experimental data. The majority of the field operated tests are conducted by automotive companies, and one of the challenges with the ACC systems deployed in vehicles is that those systems are protected as intellectual properties by their industry developers with limited publicly available data on their system design (*14*).

Considering the aforementioned limitations, the main contributions of this study are: (1) to develop a robust methodology to identify vehicles with ACC-type behavior in a vehicle trajectory dataset, and (2) to investigate the difference in behavior of conventional vehicles and the vehicles using a full range ACC or the ones with similar dynamics in a real-world setting. The application of the proposed methodology, however, can go far beyond traffic analysis and can be even utilized by AV developers for accurate prediction of vehicle maneuvers in motion planning algorithms.

The remainder of this paper is organized as follows: the next section presents the vehicle trajectory dataset utilized in this study. This section is followed by the details and steps of the proposed methodology towards the clustering of vehicle trajectories. The paper continues by presenting the clustering results of different distance measures and statistical and qualitative discussion on the results. Finally, the paper is concluded with a summary of findings and future research needs.

VEHICLE TRAJECTORY DATA

This study adopts the vehicles trajectory data collected by Khajeh-Hosseini et al. (*10*) using multiple UAVs and aerial videography of the traffic stream. The trajectory of the vehicles are extracted from the video frames recorded in the bird's-eye view from a segment of the roadway. In every video frame, the location of the vehicles are estimated with respect to a fixed coordinate system and reference point on the ground. Every video recording is converted to a sequence of image frames separated at a constant rate over time. Tracking the location of any vehicle over the sequence of images enabled extracting the vehicle's trajectory over time.

Adaptive Cruise Controlled Trajectories

The dataset of this study include trajectory of vehicles utilizing ACC. A platoon of three probe vehicles including two Toyota Prius and one Toyota Avalon, were used under full-range ACC for data collection. The leader of the platoon was following an arbitrary vehicle on the roadway in front of it using ACC. The other two vehicles were also car-following their leaders with ACC.

The data is collected on the southbound of Interstate Highway 35 between Exit 237B and Exit 238A in Austin, Texas (see Figure 1). A single stretch of 150 meters roadway was recorded for 2 hours between 07:30 AM to 09:30 AM during the morning peak on a Friday. The video

recordings were collected using two drones alternately to record the traffic stream continuously. Besides, the drones were operated in uncontrolled airspace at an altitude of 121 m (400 feet) in compliant with Federal Aviation Authority (FAA) under small unmanned aircraft regulations part 107.

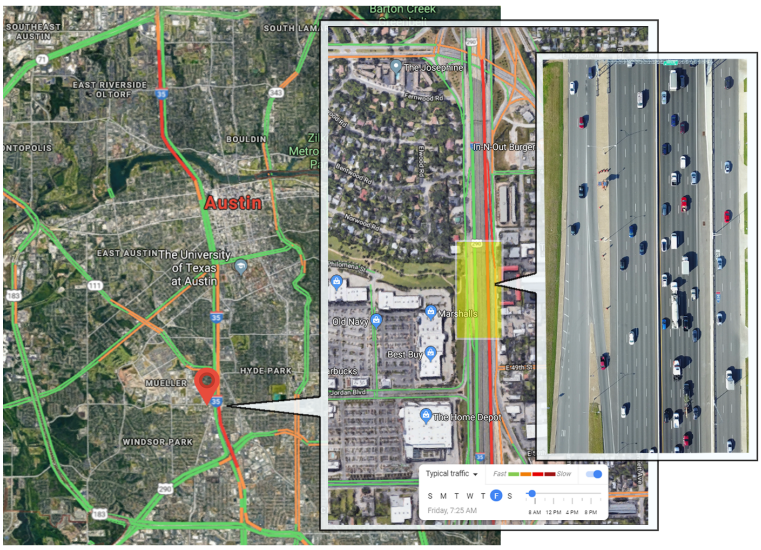


FIGURE 1: **Data collection location.**

During the data collection, five runs of the platoon of probe vehicles using ACC are recorded. The first three runs are conducted in the rightmost lane (lane 4), and the last two runs are performed in the second rightmost lane (lane 3). Figures 2 and 3 illustrate the overview of the platoons and the traffic dynamics for each of the five runs. The platoon overview, figure 2, presents the identification number of the ACC vehicles in the platoon, as well as the leader of the first ACC vehicle and the platoon of three human-driven vehicles behind the last ACC vehicle. The identification numbers are arbitrary and unique numbers assigned to each of the vehicle trajectories in the dataset. The time-space diagrams of figure 3 are generated for the period of 30 seconds before the ACC platoons entering the segment and up to 30 seconds after exiting the study segment. The trajectories of the ACC vehicles are depicted with blue lines in the time-space diagrams.

Reducing Noise in Trajectory Data: Kalman Filter

One of the main challenges in using the real-world collected data is the uncertainties that arise in measurements. The original dataset considers the front bumper as the location of the vehicle on the roadway, and trajectory of the vehicle is the list of its location over space and time (x, y, t). The data from image stabilization and vehicle detection has resulted in noisy estimation of the front bumper and consequently the positions of the vehicles. Moreover, estimating the current state from the previous noisy state results in uncertain estimation. Accordingly, a Kalman filter is applied to reduce the noise in state estimation of the vehicles. The vehicle state at each point, $x_i^{t_i}$, is characterized by its location and kinematic state. The state attributes include the position information, $p_i^{t_i}$, speed, $v_i^{t_i}$, and acceleration, $a_i^{t_i}$. The expected state of the vehicle after t (i.e., rate of data generation) seconds, $\hat{x}_i^{t_i+t}$, can be estimated by multiplying the transition matrix A by the

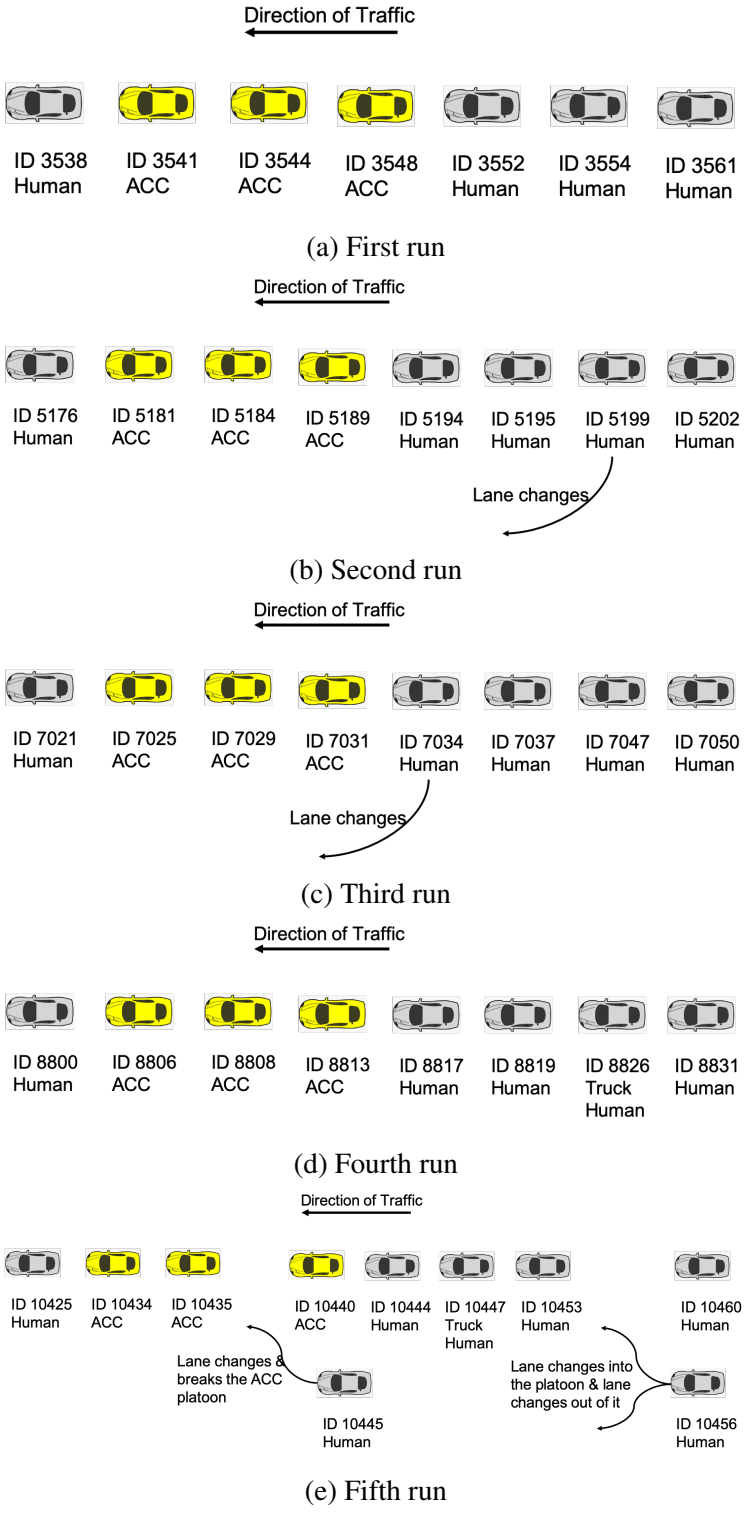
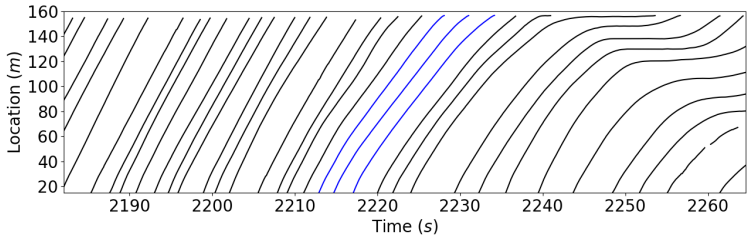
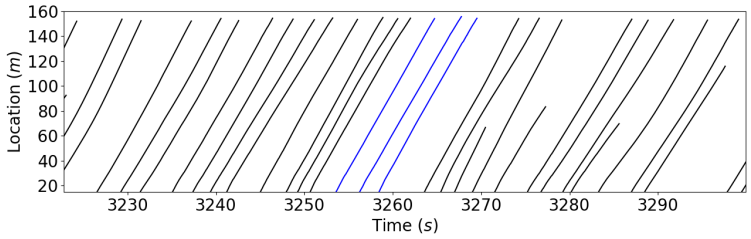


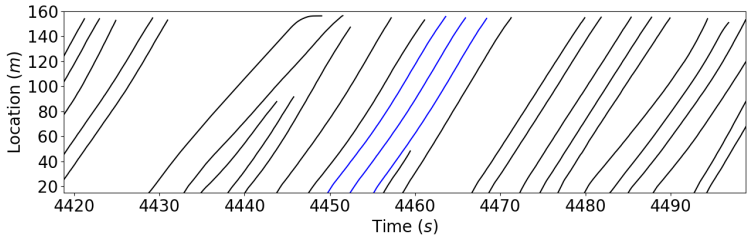
FIGURE 2: Overview of the platoon of the probe vehicles over five runs of data collection.



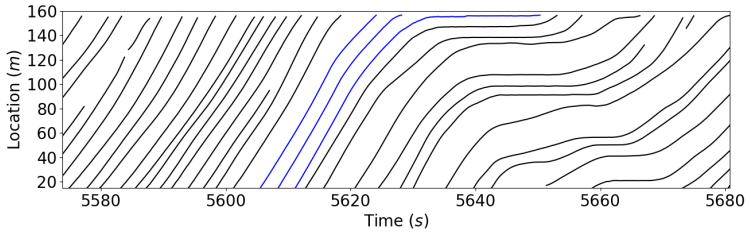
(a) First run



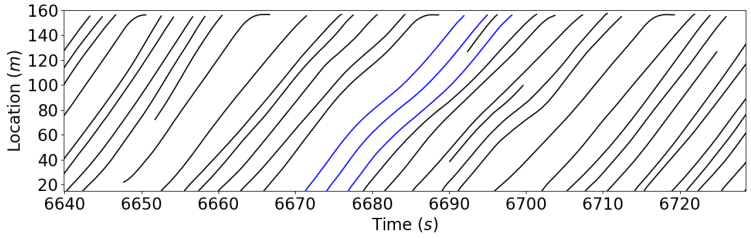
(b) Second run



(c) Third run



(d) Fourth run



(e) Fifth run

FIGURE 3: Time-space diagram of the probe vehicles over five runs of data collection.

initial state vector.

$$\hat{x}_i^{t_i+t} = Ax_i^{t_i} = \begin{bmatrix} 1 & t & \frac{t^2}{2} \\ 0 & 1 & t \\ 0 & 0 & 1 \end{bmatrix} [p_i^{t_i}, v_i^{t_i}, a_i^{t_i}]^T \quad (1)$$

In the state estimation process, the Kalman filter is usually applied to estimate the best guess on the current state of the vehicle considering the previous state and current measurements (i.e., from aerial images). Previous vehicle state, $x_i^{t_i-t}$ is transitioned to the expected current state, x_i^t, based on the process model and applying the transition matrix A:

$$\text{Process model: } x_i^{t_i} = A\hat{x}_i^{t_i-t} + \omega \quad (2)$$

where ω is the process noise. In this study, the process model considers $\omega = [t^2, t, 1]^T \sigma_{ap}^2$, where σ_{ap}^2 is the acceleration variance. ω is assumed to be normally distributed with covariance matrices of Q:

$$Q = \begin{bmatrix} \frac{t^4}{4} & \frac{t^3}{2} & \frac{t^2}{2} \\ \frac{t^3}{2} & t^2 & t \\ \frac{t^2}{2} & t & 1 \end{bmatrix} \sigma_{ap}^4 \quad (3)$$

The expected current state is converted to the expected measurement, $z_i^{t_i}$, through the following measurement model:

$$\text{Measurement model: } z_i^{t_i} = Hx_i^{t_i} + \nu \quad (4)$$

where ν is the measurement noise. Since only the position of the vehicles are directly measured from the aerial images, the resulting state to measurement conversion matrix, H, is [1,0,0]. ν is assumed to be normally distributed with covariance matrices of R:

$$R = \left[\sigma_p^2\right] \quad (5)$$

where σ_p^2 is the position variance. Note that the measurement covariance matrix considers the variance in position alone. A 2D Cartesian coordinate system is considered for the measurements, and the state of the vehicle is evaluated along the two axes, x and y, separately. Taking σ_{ap}^2 and σ_p^2, equal to $0.5(\frac{m}{s^2})^2$ and $0.5m^2$ respectively, performed well in addressing the noise in the state estimates.

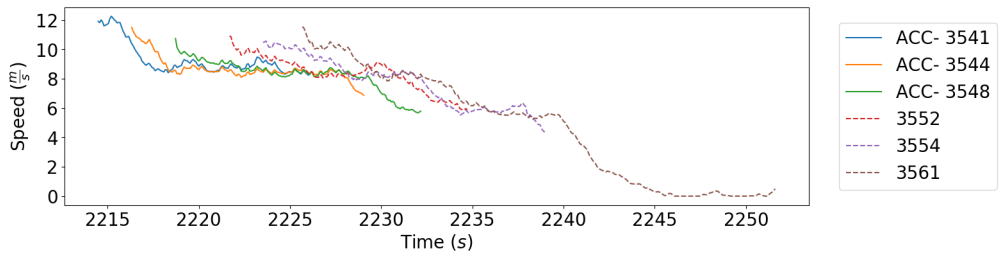
METHODOLOGY

The vehicle trajectory dataset used in this study does not differentiate between the vehicles using ADAS and conventional vehicles except for the three probe vehicles in the traffic stream. Clustering is an unsupervised approach to identify and group similar data points. Since the trajectory dataset is not labeled based on ADAS utilization, clustering is an excellent approach to arrange similar trajectories in the dataset into the same group, and potentially identifying the trajectories

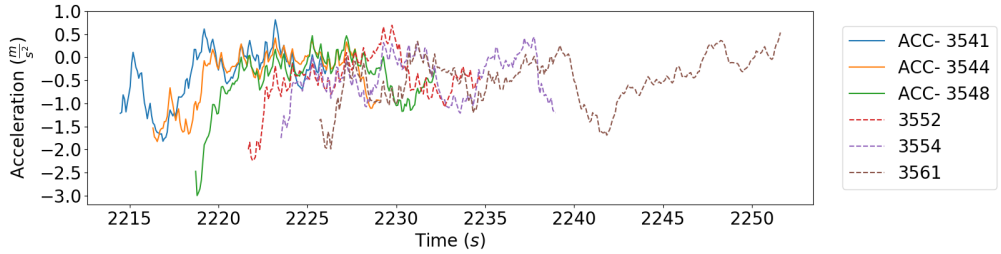
that have comparable dynamics to the probe vehicles that are using ACC.

Vehicle Trajectory as a Time-Series of Data Points

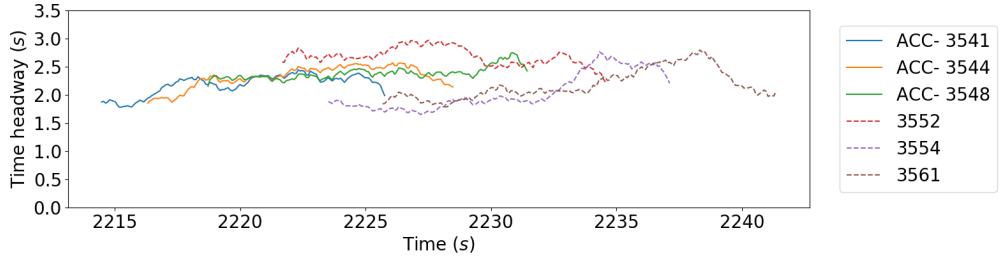
A full range ACC system, which is among the core features of automated vehicles, can adjust the vehicle's speed (i.e., longitudinal driving) in all ranges of traffic state from the stop and go to free flow. The ACC system automatically adjusts the vehicle's speed using the throttle and brake to



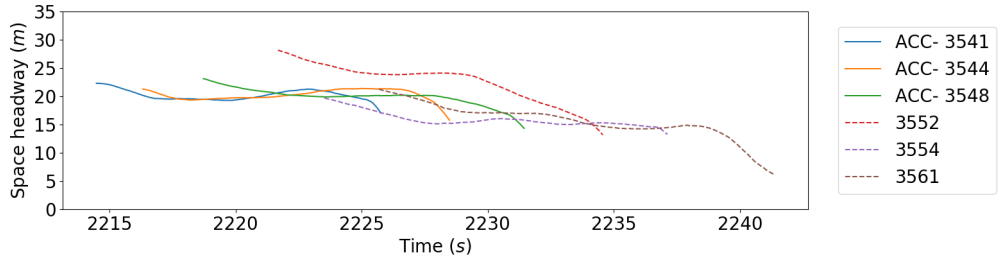
(a) Speed ($\frac{m}{s}$)



(b) Acceleration ($\frac{m}{s^2}$)



(c) Time headway (s)



(d) Space headway (m)

FIGURE 4: **Trajectory data example: first run.**

maintain a desired distance or a desired time headway to the leading vehicle (*14*). Consequently, this study considers the speed, acceleration, time headway, and space headway as well as their changes over time as the potential features in the clustering process. Each vehicle trajectory is a

TABLE 1: **Feature statistics.**

Feature	Mean	Standard Deviation	Unit
Speed	7.36	2.98	$\frac{m}{s}$
Acceleration	-0.01	0.76	$\frac{m}{s^2}$
Time headway	9.26	26.30	s
Space headway	34.78	30.91	m
Speed change	-0.02	0.14	$\frac{m}{s}$
Acceleration change	-0.02	0.18	$\frac{m}{s^2}$
Time headway change	0.10	4.73	s
Space headway change	0.31	5.21	m

time-series of data points with features, including time, location, speed, and acceleration. Based on the location and time, the vehicle's leader and follower are identified, and consequently, the time and space headways are estimated. Figure 4 presents examples of time-series of speed, acceleration, time headway, and space headway of the probe vehicles using ACC and their three immediate followers in the first run. According to this figure, the time-series of the vehicles using ACC (3541, 3544, and 3548) are more similar compared to the other three immediate following vehicles. This similarity is more noticeable for the time headway (figure 4c) and space headway (figure 4d) series.

The trajectory dataset includes five runs of data collection for probe vehicles using ACC collected over two hours. Each vehicle trajectory (i.e., time-series) is unique due to different vehicle dynamics, driver behavior, and the time of data collection (and existence of different traffic states). This study investigates the vehicle trajectories of each run separately to control for the difference in traffic states. The period of each run is considered ten seconds before the first probe vehicle entrance on to the study segment to ten seconds after the last probe vehicle exiting the segment. For each run, all the trajectories of the vehicles observed during that period are considered in the analysis.

Feature Normalization and Feature Selection

Acceleration, speed, time headway, and space headway, as well as their changes from the previous time step, are the eight features that are considered for each data point of the trajectories (i.e., time-series). For the instances that a vehicle did not have a leader, a space headway of 100 meters is considered to avoid missing features for any data point. Each feature has a different scale, and it could contribute differently to the measurement based on the similarity/dissimilarity measure adopted in the clustering process. For example, in the case of using Euclidean distance as the dissimilarity measure, the feature with a larger scale and dispersion could dominate the measurement. Accordingly, all the features are normalized using the mean and variance of the data points in all five runs (as shown in table 1) to maintain a similar scale and dispersion. Besides, normalizing features before principal component analysis helps prevent the domination of the first component with the feature with the highest variance.

The eight normalized features of acceleration, speed, time and space headway and their changes have high correlations. In the cases that the features are highly correlated, the same information contributes higher in the measurements (*15*). Principal component analysis (PCA) is

applied to transform the eight correlated features to construct new uncorrelated features and potentially reducing the number of features. The principal components are estimated, considering all the normalized data points for all trajectories of the five runs. The first seven principal components are kept to maintain a minimum of 95 percent of the variance to be retained. Each principal component is a weighted combination of the eight original features. Also, a whitening transformation is applied by multiplying the components by the square root of the number of samples divided by the singular values to keep the variance of all features as unit components.

Distance Measure Between Trajectories

A vehicle trajectory is a time-series of different features, including location and other features such as speed and acceleration. Comparing the similarity or dissimilarity between the trajectories is an essential step in grouping them into the same or different clusters (*16*). A typical distance measure used in the clustering approaches for static data points is the Euclidean distance. The Euclidean distance is also used to compare the distance between the data points of two trajectories referring to the same time step. The Euclidean distance can be used to measure the similarity of trajectories, T^i, and T^j with a similar length of n time steps and d dimensional data points, p:

$$D_{Euclidean}(T^i, T^j) = \frac{1}{n}\sum_{k=1}^{n}\sqrt{\sum_{m=1}^{d}(p_k^{i,m} - p_k^{j,m})^2} \tag{6}$$

In this equation, $p_k^{i,m}$ refers to the m^{th} feature of k^{th} point of trajectory T^i, and $p_k^{j,m}$ refers to the m^{th} feature of k^{th} point of trajectory T^j. One of the challenges with the Euclidean distance is that each data point in one trajectory is only compared to one data point in the second trajectory with the same time step. However, in most cases, including this study, the length (time steps) of trajectories are not equal, and the euclidean distance is not a suitable distance measure. Besse et al. (*17*) divides the distance measurements between trajectories with different length into two groups of warping based distances and shape-based distances. The warping distances and shape-based distances allow measuring the distance between trajectories with different lengths.

Warping based distances address the challenge of different lengths by finding an optimal (i.e., matching) alignment between the two trajectories regardless of their lengths. The objective of these distances is to find the warping path, w, between two trajectories, T^i (n^i points) and T^j with (n^j points), with the optimal cost when arranging two trajectories in the form of a $n^i \times n^j$ grid. The minimum or maximum warping cost depends on the cost function to measure dissimilarity or similarity between points. The warping path allows us to match data points from one trajectory to the points with different indexing in the second trajectory. Two of the common warping distances include the dynamic time warping (DTW) (*18*) and longest common subsequence (LCSS) (*19*).

DTW distance, equation 7, identifies the minimum cost of the warping path between two

trajectories T^i and T^j recursively:

$$D_{DTW}(T^i,T^j) = \begin{cases} 0 & \text{if } n^i = n^j = 0 \\ \infty & \text{if } n^i = 0 \text{ or } n^j = 0 \\ cost_{DTW}(p_1^i,p_1^j) + min \begin{cases} D_{DTW}(rest(T^i),rest(T^j)), \\ D_{DTW}(rest(T^i),T^j), \\ D_{DTW}(T^i,rest(T^j)), \end{cases} & \text{otherwise} \end{cases} \tag{7}$$

where $rest(T^i)$ refers to the time-series T^i without its first time step data point, and the cost structure in the DTW is the dissimilarity between the data points based on the Euclidean distance:

$$cost_{DTW}(p_1,p_2) = ||p_1p_2||_2 \tag{8}$$

LCSS distance finds the longest common subsequence between two trajectories, T^i and T^j, by counting the number of times the difference between pairs of data points is less than ε recursively:

$$D_{LCSS,\delta,\varepsilon}(T^i,T^j) = \begin{cases} 0 & \text{if } n^i = 0 \text{ or } n^j = 0 \\ 1 + D_{LCSS}(rest(T^i),rest(T^j)) & \text{if } cost_{LCSS}(p_1^i,p_1^j)\text{=1 and } |n-m| <= \delta \\ max \begin{cases} D_{LCSS}(rest(T^i),T^j) \\ D_{LCSS}(T^i,rest(T^j)) \end{cases} & \text{otherwise} \end{cases} \tag{9}$$

where δ controls how far in time the measurement can go to find a matching data point and the cost structure of the LCSS distance is as follows:

$$cost_{LCSS}(p_1,p_2) = \begin{cases} 1 & \text{if } ||p_1p_2||_2 < \varepsilon \\ 0 & \text{otherwise} \end{cases} \tag{10}$$

The final LCSS distance between two trajectories is divided by the minimum of the length of the two trajectories resulting in a value between 0 and 1. Unlike the DTW, the LCSS distance is a similarity measure between two trajectories. δ and ε are the two parameters of the LCSS distance that give some control over noise. If ε is too small, the longest common subsequence could become too small, resulting in a very low similarity, and if the ε id too large, the common subsequence could become too large resulting in a high similarity.

Shape-based distances capture the geometric similarity between two time-series (*17*). Two of the common shape-based distance measures for time series are Fréchet distance (*20*) and Hausdorff distance (*21*). Fréchet distance is a measure of similarity between two curves, A and B:

$$D_{Frechet}(A,B) = \inf_{\alpha,\beta \in X} \max_{t \in [0,1]} \{||A(\alpha(t)),B(\beta(t))||_2\} \tag{11}$$

One intuitive definition of the Fréchet distance is the minimum cord length sufficient to connect two travelers along two different curves, each traveling forward at a different speed. $\alpha(t)$ and $\beta(t)$ are continuous and increasing functions such that $\alpha(0) = 0$, $\alpha(1) = m$, $\beta(0) = 0$ and $\beta(1) = n$, and m and n are the last vertices of curve A and B respectively. Therefore, $A(\alpha(t))$ and $B(\beta(t))$ are the location of the two travelers at time t on curve A and curve B. The Frechet distance between two trajectories can be estimated using the discrete Fréchet distance algorithm proposed by Eiter and Mannila (*22*).

Hausdorff distance is another shape-based distance that can be used to measure the distance between two trajectories T^i and T^j. The maximum distance of the points in one trajectory to the nearest point in the other trajectory is the Hausdorff distance. Accordingly, for each point in trajectory T^i, the infimum of distances from this point to all the points in trajectory T^j are estimated, and the supremum of these infima is found for each trajectory. The Hausdorff distance is the maximum of the two suprema from the two trajectories:

$$D_{Hausdorff}(T^i, T^j) = \max \left\{ \sup_{p^i \in T^i} \inf_{p^j \in T^j} ||p^i p^j||_2, \sup_{p^j \in T^j} \inf_{p^i \in T^i} ||p^i p^j||_2 \right\} \tag{12}$$

Clustering Algorithm for Non-metric Distance

A distance measure is considered a metric if it satisfies non-negativity, symmetry, reflexivity, triangle inequality, and indiscernible identity. Fréchet and Hausdorff distances meet metric requirements, while the warping distances, DTW and LCSS, do not satisfy the triangle inequality (*17*). Hierarchical clustering and affinity propagation (AP) (*23*) are among the clustering algorithms capable of handling non-metric distance measure in the clustering process by directly taking the distance matrix between the trajectories. In contrast, some conventional clustering methods such as K-means clustering are best suited to work with metric distances. To simplify the clustering process, this study adopts the affinity propagation as it does not require specifying the number of clusters, unlike the hierarchical clustering.

Affinity propagation (AP) clustering considers the similarity between the trajectories and evaluates all the potential cluster heads (exemplars). Two types of messages, responsibility, and availability, are exchanged between trajectories in AP. The responsibility message, $r(i, j)$, is sent from the trajectory, T^i to a candidate exemplar trajectory, T^j, quantifies the appropriateness of trajectory T^j to serve as the exemplar for trajectory T^i considering all the other candidate exemplars. The availability message, $a(i, j)$, is sent from exemplar candidate, trajectory T^j, to trajectory T^i, and indicates the fitness of T^j to serve as the exemplar of T^i considering the support from other trajectories that take T^j as their exemplar. AP starts by considering the similarity matrix between the trajectories and setting the availability between all pairs of trajectories to zero. Through an iterative process, AP updates the responsibility and availability messages between pairs of trajectories until they converge. For each trajectory, T^i, the trajectory T^j that maximizes sum of the $r(i, j)$ and $a(i, j)$ is its exemplar (cluster head).

The similarity between a pair of trajectories is defined based on the distance measures considered in the previous section. LCSS is the only distance measure in this study that directly provides the similarity between two trajectories. All the DTW, Hausdorff, and Fréchet distances are dissimilarity measures; thus, the negative of those distances are considered the similarity measure

between two trajectories.

RESULTS AND DISCUSSION

One of the objectives of this study is to compare the traffic dynamics of the conventional vehicles with the vehicles using a full range ACC or the ones with similar dynamics. Since the vehicles using the ADAS technology are not known, the clustering approach is adopted to identify the trajectories that have a comparable traffic dynamics to the three probe vehicles using ACC during the data collection.

Clustering Results Based on Different Distance Measures

Four different distance measures, including DTW, LCSS, Fréchet, and Hausdorff, along with affinity propagation (AP) clustering, are considered to group similar trajectories for each data collection run. LCSS distance requires two hyperparameters of ε and δ that control the similarity margin and how far in time, the measurement can go to find a matching data point. Three values for $\varepsilon \in [0.01, 0.05, 0.1]$ and three values for $\delta \in [10, 50, 100]$ are examined in the clustering process when using LCSS.

The purpose of clustering is to identify trajectories with the traffic dynamics similar to the three probe vehicles using ACC during the data collection. The performance of the different distance measures in the clustering process is compared based on their effectiveness in identifying a similar behavior between the three probe vehicles or assigning them into the same cluster. Table 2 presents the number of clusters discovered for the three probe vehicles using ACC. In this table, a value of one is the most favorable and indicates that the three probe vehicles are grouped into the same vehicle cluster. A value of three is the least desirable value indicating the probe vehicles are grouped into three different clusters. According to table 2, using DTW distance in the clustering process results in a better performance compared to the other distance measures. When using DTW distance combined with the AP, the clustering process grouped the three probe vehicles into the same clusters in all runs except for the fourth run of data collection. In the fourth run, the last vehicle in the probe vehicles platoon is clustered in a single group. Figure 3d presents the time-space diagram of the three probe vehicles (8806, 8808 and 8813) using ACC during the fourth run. According to this figure, the first two probe vehicles leave the study segment just before it becomes congested, and the last probe vehicle is left on the study segment. Since the leader of the last probe vehicle is not on the study segment, the time-series of this trajectory contains a large number of data points with high space headway and very low speed (as indicated previously, when a leader is not available, a large value is utilized for the space headway), which make this trajectory an outlier. As a result, the last probe vehicle using ACC is clustered separately from the other two. In the rest of the analysis, the clusters with less than three members are considered as outliers.

TABLE 2: **Distance measures comparison based on the number of clusters identified for the three porbe vehicles with active ACC.**

Distance Measure	Run 1	Run 2	Run 3	Run 4	Run 5
LCSS ($\varepsilon = 0.01, \delta = 10$)	3	3	3	3	3
LCSS ($\varepsilon = 0.01, \delta = 50$)	3	3	3	2	3
LCSS ($\varepsilon = 0.01, \delta = 100$)	3	3	3	2	3
LCSS ($\varepsilon = 0.05, \delta = 10$)	2	2	3	2	3
LCSS ($\varepsilon = 0.05, \delta = 50$)	1	1	2	3	1
LCSS ($\varepsilon = 0.05, \delta = 100$)	1	1	2	3	2
LCSS ($\varepsilon = 0.10, \delta = 10$)	1	2	3	1	2
LCSS ($\varepsilon = 0.10, \delta = 50$)	1	2	1	2	2
LCSS ($\varepsilon = 0.10, \delta = 100$)	2	2	1	2	2
DTW	1	1	1	2	1
Fréchet	2	3	1	3	2
Hausdorff	1	3	2	3	2

Statistical Analysis of Clustered Trajectories

In the clustering process, DTW distance performed much better than the other distance measures in terms of grouping the three probe vehicles into the same cluster. The remainder of this study evaluates the clustering results when using DTW distance and only for the clusters that have more than three vehicles. Moreover, the statistical comparison between the clusters is performed separately for each run and also separately for time headway and space headway.

Table 3 presents the average and standard deviation of the time headway and space headway for each cluster in every run. In this table, the id of clusters that contain trajectory of the probe vehicles using ACC is complemented by "-ACC". Table 3 also presents the number of vehicles in each cluster, and according to this table, the probe vehicles are grouped with multiple other vehicles into the same cluster, indicating that those vehicles have similar trajectories compared to the probe vehicles using ACC.

A normality test based on D'Agostino and Pearson's (*24*) is applied on the time headway and space headway of the clusters separately. As expected, the normality tests concluded that none of the clusters follow a normal distribution at a significance level of 0.05. The headway distribution is skewed to the right for both time headway and space headway due to headway values being positive and the existence of some significant headways when the vehicle's speed is low or when there is a large distance to the leading vehicle. Following the normality test, Bartlett's test (*25*) is conducted to evaluate the homogeneity of variances of headways between the clusters of each run. Bartlett's test does not require close to normality distribution, unlike the Levene's test (*26*). The result of Bartlett's tests at a significance level of 0.05 suggested that for each run, at least two of the clusters have different variances for both time headway and space headway. Besides, the Kruskal-Wallis test (*27*) is conducted to compare the similarity between the headway distribution of clusters in each run. Kruskal-Wallis is a non-parametric method and does not require the normal distribution of the samples. For each run, and for both time headway and space headway, at a significance level of 0.05, it is concluded that at least two clusters have a different distribution.

For both time headway and space headway, the results of Bartlett's test and the Kruskal-Wallis test suggest that in each run, at least two of the clusters have different variances and distri-

TABLE 3: **Statistics of time headway (s) and space headway (m) for different clusters of the five runs.**

	Cluster Id	# Vehicles	Avg. T-Headway	Std. Dev. T-Headway	Avg. S-Headway	Std. Dev. S-Headway
Run 1	0	7	2.43	0.58	19.70	4.94
	1	16	3.24	1.33	21.12	11.10
	2	8	2.29	0.97	26.71	9.98
	3	4	2.75	0.79	15.64	3.42
	4-ACC	7	2.22	0.65	20.60	4.20
	5	7	3.38	1.28	19.66	7.93
	6	4	2.81	1.03	15.26	2.70
	7	5	2.32	0.68	16.11	4.56
Run 2	1	8	2.26	0.83	24.72	7.66
	2	4	4.66	1.12	32.15	9.28
	3-ACC	16	2.74	1.11	32.82	10.31
	4	18	2.33	0.80	16.47	4.73
	5	7	4.15	1.07	29.05	6.07
Run 3	6	7	3.41	1.09	27.67	9.27
	7	7	2.44	1.27	20.64	8.55
	8-ACC	15	2.58	1.05	24.60	7.91
	11	9	2.71	1.25	23.38	6.26
	12	6	2.96	1.13	34.67	14.28
Run 4	0	7	3.50	1.27	18.6	6.26
	1	8	1.98	0.59	14.99	2.80
	2	15	3.29	1.07	26.81	7.59
	3	6	3.81	2.02	40.49	20.16
	4	4	4.57	0.76	27.57	3.35
	8	3	3.44	1.09	22.93	4.36
	10-ACC	4	3.45	0.91	28.8	5.28
Run 5	0	11	4.18	1.51	33.82	11.61
	2	11	2.89	0.91	16.54	3.84
	3	11	2.86	1.07	31.52	10.83
	4	5	3.39	1.46	18.85	5.20
	5-ACC	16	2.82	1.40	18.87	8.00
	7	3	4.44	1.38	24.81	8.72

butions. Following these two tests, each run's clusters are compared pairwise for the similarity in their variances and means at a significance level of 0.05. Bartlett's test is applied for the pairwise comparison of the variances, and Welch's t-test is adopted to evaluate the similarity of means. The advantage of Welch's t-test over the student t-test is that it does not require equal variances and number of samples. Most of the pairwise comparisons concluded that there is a statistical difference between the means and variances of the clusters in each run for both time headway and space headway. The few pairwise tests that failed to reject the null hypotheses are presented in table 4. According to this table, the pairwise tests was unable to reject equal means and variances between the time headways of clusters 0 and 10 in run 4; however, the means and variances of space headways of these two clusters were statistically different at a significance level of 0.05. From the pairwise comparison between the clusters, it can be concluded that detected clusters have a different distribution of time headway and space headway, and the clustering approach is capable of grouping the trajectories into different traffic flow dynamics.

TABLE 4: **Hypotheses that are failed in the pairwise comparisons of the clusters for each run.**

	Between	Null hypothesis	P-value
Run 1	clusters 0 and 5	equal space headway means	0.37
	clusters 1 and 5	equal time headway variances	0.33
	clusters 2 and 3	equal time headway variances	0.53
	clusters 4-ACC and 7	equal time headway variances	0.12
Run 2	clusters 1 and 4	equal time headway variances	0.79
	clusters 2 and 5	equal time headway variances	0.35
Run 3	clusters 7 and 8-ACC	equal time headway means	0.54
Run 4	clusters 0 and 2	equal time headway variances	0.56
	clusters 0 and 10-ACC	equal time headway variances	0.48
	clusters 0 and 10-ACC	equal time headway means	0.37
	clusters 2 and 10-ACC	equal time headway variances	0.73
	clusters 4 and 8	equal space headway variances	0.44
	clusters 8 and 10-ACC	equal time headway means	0.10
Run 5	clusters 0 and 3	equal space headway variances	0.05
	clusters 0 and 4	equal time headway variances	0.92
	clusters 2 and 3	equal time headway means	0.58
	clusters 2 and 5-ACC	equal time headway means	0.63
	clusters 3 and 5-ACC	equal time headway means	0.40
	clusters 4 and 5-ACC	equal space headway means	0.87
	clusters 5-ACC and 7	equal space headway variances	0.55

Traffic Flow Dynamics within Each Cluster

This section discusses the macroscopic traffic flow dynamics within each cluster and compares them with the dynamics in the entire segment. In particular, speed-flow diagrams are created for each cluster and the entire segment. In order to create these diagrams for each cluster, the average value of time headway and speed for all the data points of trajectories that fall within a cluster are estimated for each time step during each run. Besides, the trajectory data points that did not have a leader or have a speed value of less than 0.1 $\frac{m}{s}$ are ignored in the calculation to ensure meaningful time headways. Each cluster's flow rate at each time step is approximated by the inverse of the average time headway of that cluster at that time.

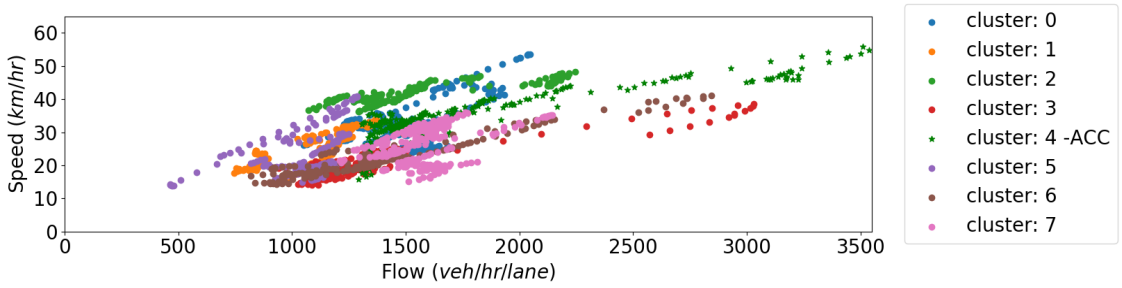
Figure 5 shows the speed-flow graphs for each cluster in each run. This figure indicate a clear distinction between the macroscopic behavior of clusters that contain ACC vehicles and other clusters in runs 1and 2. In the remaining runs, the macroscopic behavior of clusters that contain ACC vehicles is fundamentally different from the majority of the clusters. In fact, in runs 3, 4 and 5, only two, one, and two other clusters show similar behavior, respectively. From the perspective of scatter in the speed-flow diagram, interestingly enough, clusters that contain ACC vehicles have the least amount of scatter in all five runs. This shows that the behavior of these vehicles are more predictable and they are the least likely group of vehicles to contribute to traffic flow breakdown.

Figure 6 shows the average speed-flow graphs for the entire segment. Comparing this figure with Figure 5 reveals interesting observations. First, while some clusters have very large flow rates (e.g., the flow rate in cluster 4 in run 1 reaches 3500 veh/hr/lane), the overall segment has a fairly average flow rate (about 1500 veh/hr/lane). This difference suggests that while some runs show potential to significantly increase flow rate through platooning, the impact of platooning in a mix driving environment might not be as significant until high penetration rates of ACC vehicles. Second, the amount of scatter in Figure 6 is significantly less than Figure 5. This observation suggests that while different clusters behavior differently at different time steps, their average behavior stays the same. In other words, the impact of ACC-type behavior on the entire system (if any) remains constant through out the data.

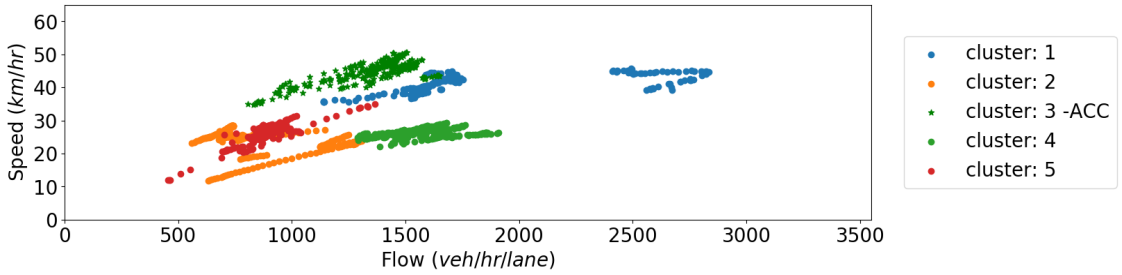
CONCLUSION

The new vehicles equipped with Advanced Driver Assistant Systems (ADAS), such as adaptive cruise control (ACC), can potentially change the interaction among drivers on the road. The existing trajectory datasets fail to provide any information on the utilization of ADAS technologies. This study proposes the use of a new trajectory dataset that contains multiple instances of vehicles using ACC to idnetify ACC-type behavior across the entire trajectory dataset. The trajectory data were collected for over two hours from a 150 meters long segment of I-35 near Austin, TX. The collected trajectories contain five runs of data collection from one platoon consists of three vehicles operated based on an Adaptive Cruise Control (ACC) system. The vehicle trajectory is a time-series of different features, including location and other features such as speed and acceleration. Since the trajectory data is not labeled based on ACC utilization, clustering is an excellent approach to arrange similar trajectories in the dataset into the same group. Comparing the similarity between the trajectories is an essential step in grouping them into the same or different clusters. One of the challenges with vehicle trajectory data is that the trajectories do not have equal lengths (i.e., number of time steps), and the typical Euclidean distance is not a suitable distance measure between trajectories. The distance measures used to compare time-series with different lengths include the warping based distances such as dynamic time warping (DTW) and longest common subsequence (LCSS) and shape-based distance such as Fréchet and Hausdorff. Besides, some of the distance measures between the trajectories do not satisfy the triangle inequality, which limits the clustering method to algorithms such as affinity propagation (AP), which is capable of working directly with the distance matrix. The clustering results with different distance measures indicated that the DTW distance between the trajectories has a better performance in keeping the probe vehicles using ACC in the same group. The statistical analysis of the time headway and space headway indicated a statistical difference between the traffic dynamics of different clusters. The unique trajectory dataset of this study combined with the clustering provides the opportunity to identify vehicle trajectories with comparable traffic dynamics to the vehicles using ACC.

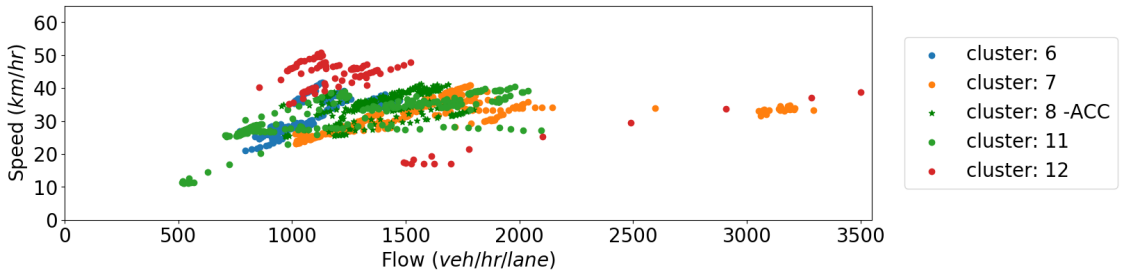
This study proposes a methodology to identify clusters of trajectories with similar traffic dynamics to the vehicles using ADAS systems. The clustering results could be used to calibrate different car following models to gain further information on the behavior of different clusters. This step is left for future studies.



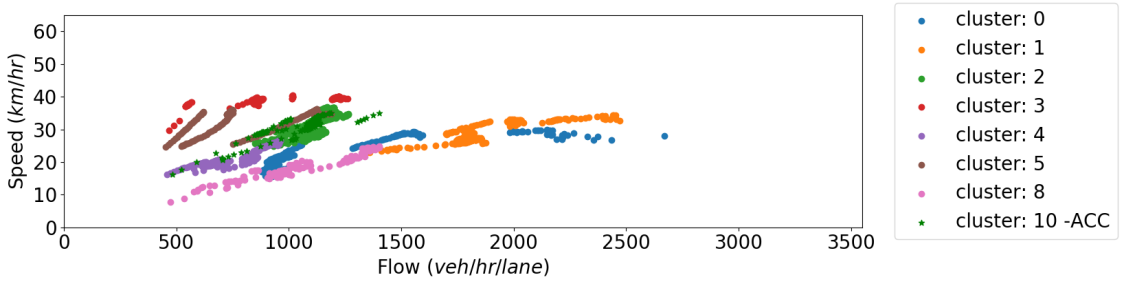
(a) Run 1 - Per cluster



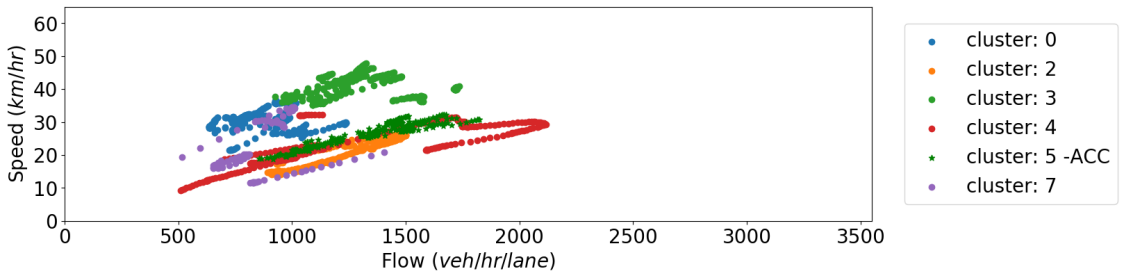
(b) Run 2 - Per cluster



(c) Run 3 - Per cluster

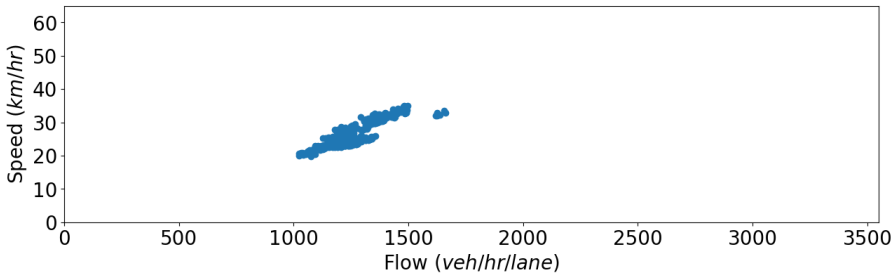


(d) Run 4 - Per cluster

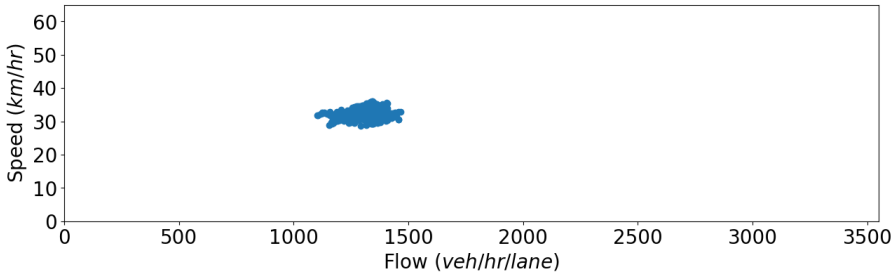


(e) Run 5 - Per cluster

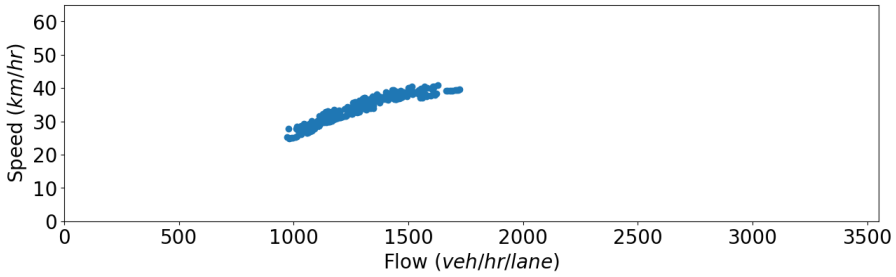
FIGURE 5: Speed and flow for each cluster for each run.



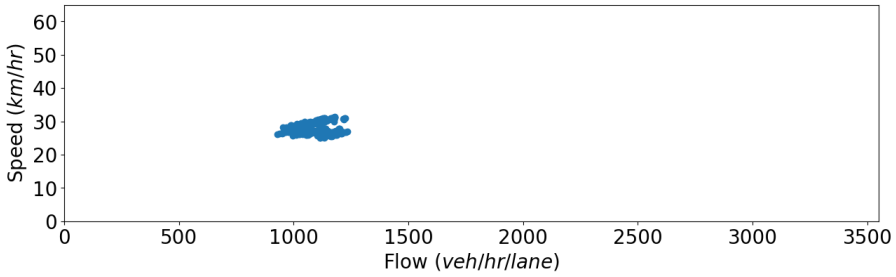
(a) Run 1 - Average of all trajectories



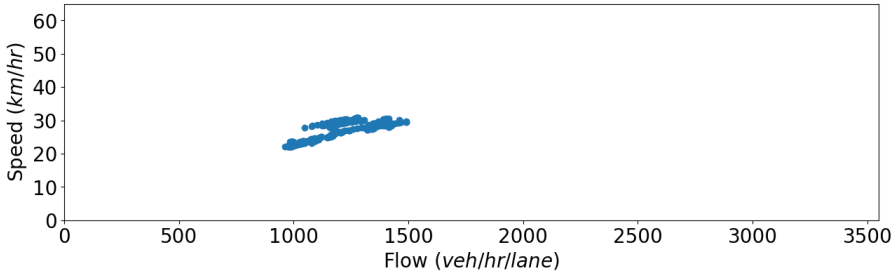
(b) Run 2 - Average of all trajectories



(c) Run 3 - Average of all trajectories



(d) Run 4 - Average of all trajectories



(e) Run 5 - Average of all trajectories

FIGURE 6: Speed and flow for all trajectories for each run.

AUTHOR CONTRIBUTION

All authors contributed to all aspects of the study from conception and design, to data collection, to analysis and interpretation of results, and manuscript preparation. All authors reviewed the results and approved the final version of the manuscript.

REFERENCES

1. Michalopoulos, P. G., Vehicle detection video through image processing: the autoscope system. *IEEE Transactions on vehicular technology*, Vol. 40, No. 1, 1991, pp. 21–29.
2. Zhou, J., D. Gao, and D. Zhang, Moving vehicle detection for automatic traffic monitoring. *IEEE transactions on vehicular technology*, Vol. 56, No. 1, 2007, pp. 51–59.
3. U.S. Federal Highway Administration. Next Generation Simulation (NGSIM). *https://ops.fhwa.dot.gov/trafficanalysistools/ngsim.htm*, 2006.
4. Hankey, J. M., M. A. Perez, and J. A. McClafferty, *Description of the SHRP 2 naturalistic database and the crash, near-crash, and baseline data sets*. Virginia Tech Transportation Institute, 2016.
5. Zhao, D., Y. Guo, and Y. J. Jia, Trafficnet: An open naturalistic driving scenario library. In *2017 IEEE 20th International Conference on Intelligent Transportation Systems (ITSC)*, IEEE, 2017, pp. 1–8.
6. 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.
7. 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 IEEE 21st International Conference on Intelligent Transportation Systems (ITSC)*, 2018.
8. Barmpounakis, E. and N. Geroliminis, On the new era of urban traffic monitoring with massive drone data: The pNEUMA large-scale field experiment. *Transportation Research Part C: Emerging Technologies*, Vol. 111, 2020, pp. 50–71.
9. Bhutani, A. and P. Bhardwaj, Automotive Camera Market Share 2019-2025: Global Industry Report. *https://www.gminsights.com/industry-analysis/automotive-camera-market*, 2019.
10. Khajeh-Hosseini, M., S. Devunuri, A. Talebpour, and S. H. Hamdar, Vehicle Trajectory Data Collection Using Aerial Videography. *Transportation Research Board 99th Annual Meeting*, 2020.
11. Xiao, L. and F. Gao, A comprehensive review of the development of adaptive cruise control systems. *Vehicle system dynamics*, Vol. 48, No. 10, 2010, pp. 1167–1192.
12. Milanés, V. and S. E. Shladover, Modeling cooperative and autonomous adaptive cruise control dynamic responses using experimental data. *Transportation Research Part C: Emerging Technologies*, Vol. 48, 2014, pp. 285–300.
13. Fancher, P., *Intelligent cruise control field operational test. Final report. Volume II: appendices A-F*, 1998.
14. He, Y., B. Ciuffo, Q. Zhou, M. Makridis, K. Mattas, J. Li, Z. Li, F. Yan, and H. Xu, Adaptive cruise control strategies implemented on experimental vehicles: A review. *IFAC-PapersOnLine*, Vol. 52, No. 5, 2019, pp. 21–27.

15. Sambandam, R., Cluster analysis gets complicated. *Marketing Research*, Vol. 15, No. 1, 2003, pp. 16–21.
16. Yuan, G., P. Sun, J. Zhao, D. Li, and C. Wang, A review of moving object trajectory clustering algorithms. *Artificial Intelligence Review*, Vol. 47, No. 1, 2017, pp. 123–144.
17. Besse, P. C., B. Guillouet, J.-M. Loubes, and F. Royer, Review and perspective for distance-based clustering of vehicle trajectories. *IEEE Transactions on Intelligent Transportation Systems*, Vol. 17, No. 11, 2016, pp. 3306–3317.
18. Berndt, D. J. and J. Clifford, Using dynamic time warping to find patterns in time series. In *KDD workshop*, Seattle, WA, USA:, 1994, Vol. 10, pp. 359–370.
19. Vlachos, M., G. Kollios, and D. Gunopulos, Discovering similar multidimensional trajectories. In *Proceedings 18th international conference on data engineering*, IEEE, 2002, pp. 673–684.
20. Fréchet, M. M., Sur quelques points du calcul fonctionnel. *Rendiconti del Circolo Matematico di Palermo (1884-1940)*, Vol. 22, No. 1, 1906, pp. 1–72.
21. Hausdorff, F., *Grundzüge der mengenlehre*, Vol. 7. von Veit, 1914.
22. Eiter, T. and H. Mannila, *Computing discrete Fréchet distance*. Citeseer, 1994.
23. Frey, B. J. and D. Dueck, Clustering by passing messages between data points. *science*, Vol. 315, No. 5814, 2007, pp. 972–976.
24. D'AGOSTINO, R. and E. S. Pearson, Tests for departure from normality. Empirical results for the distributions of b 2 and b. *Biometrika*, Vol. 60, No. 3, 1973, pp. 613–622.
25. Snedecor, G. W. and W. G. Cochran, Statistical Methods, eight edition. *Iowa state University press, Ames, Iowa*, 1989.
26. Levene, H., Robust tests for equality of variances. *Contributions to probability and statistics. Essays in honor of Harold Hotelling*, 1961, pp. 279–292.
27. Daniel, W. W., Kruskal–Wallis one-way analysis of variance by ranks. *Applied nonparametric statistics*, 1990, pp. 226–234.