Dynamic Clustering Analysis for Driving Styles Identification

Maria Valentina Niño de Zepeda[a], Fanlin Meng[a,*], Jinya Su[b], Xiao-Jun Zeng[c], Qian Wang[d]

[a] *Department of Mathematical Sciences, University of Essex, Colchester CO4 3SQ, UK*
[b] *School of Computer Science and Electronic Engineering, University of Essex, Colchester CO4 3SQ, UK*
[c] *Department of Computer Science, University of Manchester, Manchester M13 9PL, UK*
[d] *Department of Computer Science, Durham University, Durham DH1 3LE, UK*

Abstract

For intelligent driving systems, the ability to recognize different driving styles of surrounding vehicles is crucial in determining the safest, yet more efficient driving decisions especially in the context of the mixed driving environment. Knowing for instance if the vehicle in the adjacent lane is aggressive or cautious can greatly assist in the decision making of ego vehicle in terms of whether and when it is appropriate to make particular manoeuvres (e.g. lane change). In addition, vehicles behave differently under different surrounding environments, making the driving styles identification highly challenging. To this end, in this paper we propose a dynamic clustering based driving styles identification and profiling approach where clusters vary in response to the changing surrounding environment. To better capture dynamic driving patterns and understand the driving style switch behaviours and more complicated driving patterns, a position-dependent dynamic clustering structure is developed where a driver is assigned to a cluster sequence rather than a single cluster. To the best of our knowledge, this is the first research paper of its kind on the dynamic clustering of driving styles. The usefulness of the proposed method is demonstrated on a real-world vehicle trajectory dataset where results show that driving style switches and more complex driving behaviours can be better captured. The potential applications in intelligent driving systems are also discussed.

Keywords: dynamic clustering analysis, driving style, mixed driving environment, vehicle trajectory

1. Introduction

Autonomous driving technology is evolving rapidly and with it, a new mixed-traffic environment in which human drivers interact with self-driving cars is arising [1]. This scenario requires intelligent driving systems to be able to react to an uncertain environment [2]. In this context, [3] identifies two types of uncertainties: one originating from noisy sensors and the other coming from human intentions. For noisy sensors, new technologies, such as high precision GPS and vehicle-to-vehicle communication, can help

[*] Corresponding author

Email addresses: wvninodez@gmail.com (Maria Valentina Niño de Zepeda), fanlin.meng@essex.ac.uk (Fanlin Meng), j.su@essex.ac.uk (Jinya Su), x.zeng@manchester.ac.uk (Xiao-Jun Zeng), qian.wang@durham.ac.uk (Qian Wang)

self-driving systems obtain more precise information about their surroundings. For human intentions uncertainty, some studies have been done to predict human behaviours during driving [4][5][6].

As widely recognized, autonomous vehicles should be able to predict the intentions of other road participants taking all the precautions in order to drive safely, yet not overreacting to threats with low probability [7] [8]. When all surrounding vehicles are treated the same, the intelligent system is obliged to make very conservative decisions to reduce all possible collision risks. However, based on real-world observations, drivers are diverse and usually react differently to stimulus, which creates complex driving behaviours and profiles. As a result, understanding the underlying driving behaviours and styles of drivers becomes necessary in enabling a safe and efficient autonomous driving.

In this paper we focus on identifying different driving styles among human drivers, which is particularly important in the mixed driving environment consisting of human drivers and autonomous/self-driving vehicles. By using clustering techniques, we can find groups of drivers that share similar characteristics. In addition, to account for the fact that some driving styles may show only under specific circumstances, a dynamic clustering framework is considered. More specifically, multiple clustering analyses are done by varying the input data using different cutting points of the vehicle trajectory data. In other words, a position-dependent structure of the clusters is proposed. Such an approach has the capability to recognise the evolving behaviours of drivers. For instance, a driver that started as a more impulsive type of driver may change to a more cautious type as he goes on the road when the surrounding environment changes. This dynamic clustering approach helps represent more complex driving profiles than traditional static clustering. The US-101 dataset from the Next Generation Simulation Program (NGSIM) is used in this paper, which contains the trajectories of vehicles in the U.S. 101 Highway of Los Angeles, California - a five-lane road with one on-ramp access and one off-ramp exit [9]. Following the exploratory analysis, we focus on drivers that take the on-ramp to access the free-way since they show a wider range of behaviours than drivers that are passing through.

The main contributions of this study are summarized as follows:

- We propose a dynamic clustering framework for driving style profiling and identification where dynamic and evolving driving styles under changing surrounding environment can be effectively captured. To the best of our knowledge, this is the first paper to present a dynamic clustering approach to tackle the problem.

- We develop a position-dependent dynamic cluster structure where a driver is assigned to a cluster sequence rather than a single cluster. In such a way, it can not only capture the dynamic driving behaviours but also can enable a better understanding of different driving pattern switches and more complicated driving behaviours through improved interpretability.

- A comprehensive experimental study is conducted based on a real-world vehicle trajectory dataset considering different static clustering algorithms to demonstrate the effectiveness of the proposed dynamic clustering approach.

The paper is organized as follows: a literature review is conducted in Section II . In Section III the considered dataset and the data preprocessing are detailed. The proposed dynamic clustering based driving styles analysis framework is given in Section IV while results are demonstrated in Section V. Finally, the paper is concluded in Section VI.

2. Related Work

2.1. Driving Styles Identification

Many studies have been done to identify driving styles (DS) among human drivers. It is a complex task for many reasons. First, DS is a concept with no unique clear definition.[10] states that there is no agreed definition of DS in the literature. Some associate it with subjective factors such as the driver's attitude, mood or way of thinking about driving whereas others attempt to provide a more pragmatic description, restricting it exclusively to the manner a driver operates the vehicle [11]. However, it is commonly accepted that driving styles are influenced by two types of factors: human and environmental factors. This implies that a driver may vary its DS according to the environmental factors he faces.

Most of the research on transportation systems identify two to three types of drivers, usually defining them by their level of aggressiveness (or cautiousness). [4] identifies two types of DS: aggressive and normal. It uses a semi-supervised support vector machine (SVM) approach to label the drivers. Similarly, [5] uses k-means with SVM to classify drivers into aggressive and moderate types. [6] uses three different driving styles in the lane-changing model: cautious, stable and radical.

In the field of psychology, research have also been done to classify drivers into different profiles. For instance, [12] identifies four DS: risky, angry, careful and anxious. The approach links personality traits and socio-demographic variables to the DS recognition, which could open new possibilities for more quantitative approaches to incorporate such variables.

Although the above literature provides valuable insights on driving styles identification, they are mostly classification based approaches, which requires a lot of efforts and costs to collect and label the data. Due to such limitations, the collected data for classification based driving styles identification are usually small. Therefore, they can only identify a few simple behaviour patterns such as aggressive and moderate, which is oversimplified given the fact that there are possibly much more complicated behaviours in real-world scenarios. Instead, our proposed dynamic clustering based approach is an unsupervised learning approach which does not have the above limitations and can also better capture the dynamic and complex driving behaviours.

2.2. Clustering based approaches

Cluster analysis is the formal study of algorithms and methods for grouping objects. One of the most popular algorithms for clustering is k-means [13]. It presents many advantages such as its simple interpretation and its scalability to large samples but also some drawbacks such as it requires the number of clusters to be defined in advance and the risk that it may converge to a local minimum (in other words, it is sensitive to the centroids initialization). Some research have been proposed to improve the initialization of k-means clustering. For instance, [14] propose k-means++, a randomized seeding technique to improve the speed of convergence and accuracy of k-means. Another possible problem of k-means is that it does not perform well in identifying clusters with a non-flat geometry. Spectral clustering [15], a technique has its roots in graph theory can deal with the above problem. Same as k-means, in spectral clustering the number of clusters must be pre-defined. Different from k-means and spectral clustering, density-based spatial clustering of applications with noise (DBSCAN) [16] does not need to pre-define of the number of clusters. This technique separates clusters of high density from low density, and identifies areas in the data that have a high density of observations. However, it is not efficient in separating clusters when they have similar densities.

Some research (e.g. [5]) have used clustering algorithms to assist in identifying driving styles in the intelligent transportation systems, however, they are usually considered in a static clustering framework. In other words, the existing clustering approaches cannot recognise the dynamic and evolving driving styles of drivers when the surrounding environments change.

Dynamic clustering (DC), which embraces clustering scenarios with dynamic features, dynamic data objects and/or dynamic clusters has been proposed recently in different research areas (e.g. [17]). In this type of clustering, each observation is not assigned to a single cluster label, but a sequence of labels that change dynamically according to an evolving variable. Typically, the variable considered in the clustering analysis can change with time such that clusters location and memberships will evolve over time, which is similar to time series clustering. Time series clustering has been used in different fields including biology, energy, finances and psychology to discover complex time evolving patterns [18]. For instance, [19] uses DC with a Bayesian approach to find links between genes and stages of the cell cycle. It proposes a structure in which not only the cluster membership can change over time, but also in which clusters can split and merge over time.

In the area of intelligent transport systems, few research have been done along this direction. In this paper, we propose a dynamic clustering based approach where clusters vary depending on the position of vehicles. By embedding existing static clustering algorithms into the proposed dynamic clustering framework, the proposed approach can naturally account for dynamic driving behaviours when drivers face changing surrounding environments. It enables a better understanding of different driving pattern switches and can also help understand more complicated driving patterns to improve the interpretability.

3. Data

In this section, we first describe the selected dataset and the underlying rationale. Second, we give details of the data processing in smoothing and cleaning the data.

3.1. Description of the dataset

In this paper we use the US-101 dataset from the Next Generation Simulation Program (NGSIM). The vehicle trajectory data collected on the US 101 highway was registered on June 15th 2005. Eight syncronized cameras placed on top of a nearby building recorded vehicles passing through the study area. The study area consists of 640 metres of a five lane freeway illustrated in Figure 1. Note that lanes 6, 7 and 8 correspond to the auxiliary lane, the onramp and the offramp respectively. For lanes 1 to 5, the lane numbering is incremented from the left-most lane.

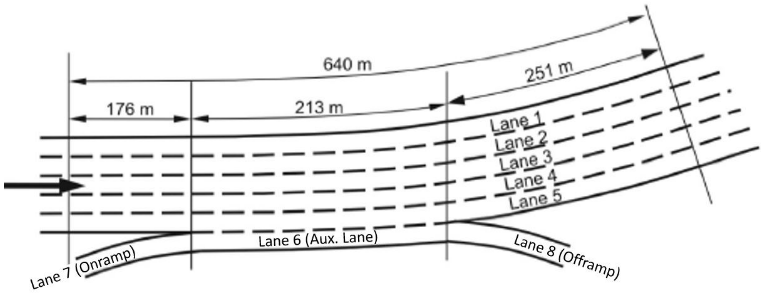


Figure 1: US 101 Highway Diagram

The recording has a duration of 45 minutes, which is segmented into three periods, 15 minutes each. Period 1 represents a transitional traffic in the build up to congestion. Periods 2 and 3 represent primarily congested conditions.

3.2. Defining mandatory and discretionary lane changes

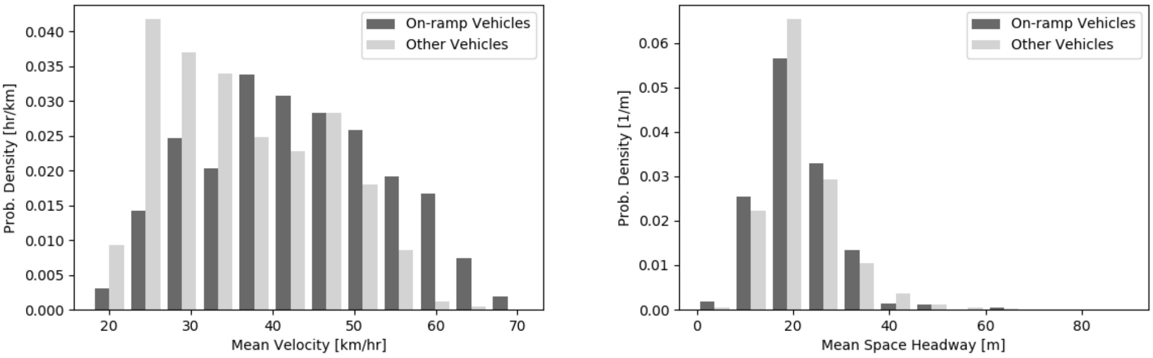
Lane changes are one of the most important factors in analysing driving behaviours and styles. Following [20], we separated lane changes into mandatory (MLC) and discretionary (DLC). In this study, MLC are defined as any lane change action taken to get out of the on-ramp or to access the off-ramp. More specifically, four types of MLC are defined in this paper, which can be found in Table 1. All other lane changes are considered as DLC.

Table 1: MLC Definition

Type of mandatory lane change
Going from on-ramp to auxiliary lane
Going from auxiliary lane to off-ramp
Going from auxiliary lane to rightmost lane when coming from the on-ramp
Any lane change to the right when the vehicle is attempting to access the off-ramp

3.3. Selecting on-ramp drivers

According to [21] which analyses driving behaviour at motorway ramps, in the surroundings of an on-ramp multiple manoeuvres are observed (e.g. changes in speed, changes in headway, lane changes). This phenomenon is referred to as turbulence. When turbulence is present there is a broader range of possible driving behaviours to observe than in less turbulent environments. Figure 2 shows the distribution of mean velocity and mean space headway. Vehicles taking the on-ramp show a wider range of velocities and space headways than other vehicles. Also, they show more discretionary lane changes than the rest: the average number of DLC per driver is 0.8 for on-ramp vehicles vs 0.2 for the rest of the vehicles. For this reason, in this paper we focus on drivers that take the onramp in order to identify different driving styles.



(a) Mean Velocity

(b) Mean Space Headway

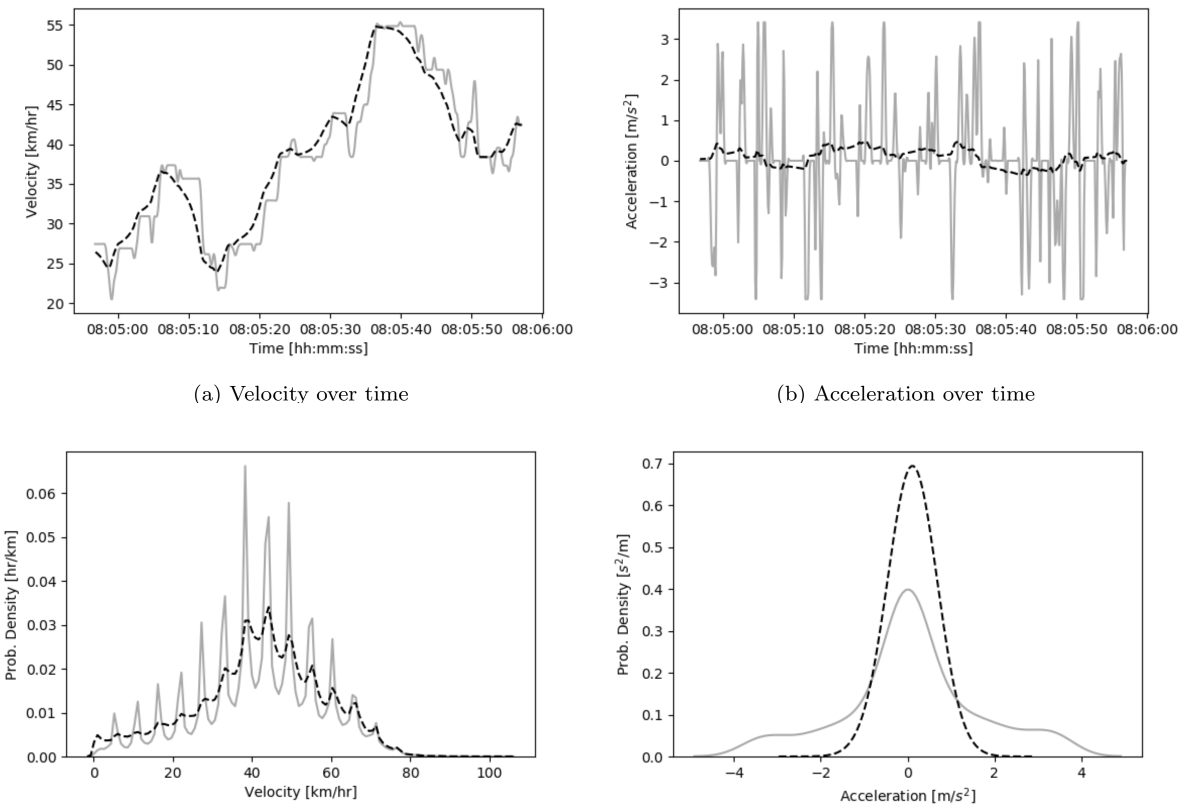
Figure 2: On-ramp Vehicles vs Other Vehicles.

3.4. Data Processing

3.4.1. Smoothing velocity and acceleration

As can be seen in the Figure 3, original velocity and acceleration variables have unrealistic distributions. This has also been reported in other research [22]. For an example vehicle, velocity over time (see

Figure 3(a)) shows abrupt changes between different velocities instead of smooth transitions. In addition, the acceleration (Figure 3(b)) varies between hard acceleration and hard deceleration several times per second. These behaviours are improbable in real-world observations. Similar abnormal observations can be found in the velocity distribution (Figure 3(c)) and acceleration distribution (Figure 3(d)). These unrealistic distributions are suspected to be the result of data collection or data post-processing. For this reason, an exponential moving average filter was applied over these two variables, using a span of 30 for velocity and of 120 for acceleration. The resulting distribution of the variables after smoothing can be found in Figure 3 (dashed lines).



(a) Velocity over time

(b) Acceleration over time

(c) Velocity Distribution

(d) Acceleration Distribution

Figure 3: Velocity and acceleration smoothing.

3.4.2. Cleaning the data from tracking software errors

Vehicle trajectories were transcribed from the video data using software that automatically detects and tracks the vehicles. When analysing the dataset, some inconsistencies were detected on the trajectories of certain vehicles, which were caused by errors of the tracking system. To deal with them two rules were considered. First, we defined that lane changes can only happen between adjacent lanes (e.g. it cannot skip a lane in one single time-step). All vehicles that present this type of unfeasible lane change were removed from the dataset. Second, some vehicles also showed extremely short overtaking manoeuvres. This is when a vehicle changes to the adjacent lane and then comes back to the original lane in an extremely short time. In such situations, the actual lane change did not happen and the vehicle remained in the original lane.

4. Methodology

In this section, we first present the system framework of our proposed dynamic clustering approach for driving styles analysis. Second, technical details of each key component in the system are described.

4.1. System Framework

In this paper, we proposed a new dynamic-clustering based driving style analysis approach to capture the evolving driving patterns in the dynamic driving environment and complex driving behaviours such as driving styles switching behaviours. The methodological and system framework is illustrated in the Figure 4 which mainly consists of the dynamic clustering structure and the dynamic clustering implementation and analysis.

More specifically, by defining the position-dependent dynamic cluster structure that provides fundamentals for the proposed dynamic clustering framework (subsection 4.2), clustering implementation was conducted based on data up to each cutting point, which involves the clustering attributes selection, clustering algorithms and evaluation metrics (subsections 4.3-4.5). The clustering results were analysed and aggregated for the dynamic clustering clusters profiling and driving styles analysis (subsection 4.6).

4.2. Dynamic cluster structure

We use a dynamic clustering approach where the position of the vehicle varies over time. We define four cutting points in this study (p=300m, p=400m, p=500m and p=600m where p represents the position and m stands for meters). The selection of these cutting points follows from a visual and exploratory analysis considering the geometry of the study area of the highway. For instance, the first cutting point p=300m is selected to be slightly after the on-ramp, which is expected to observe different and some extreme driving behaviours (e.g. multiple lane changes in a short time frame). The subsequent cutting points are selected to capture potential dynamic changing driving behaviours over time. For each cutting point we perform a separate cluster analysis using only the information available to that point. This is done to capture the space-dependent structure of clusters and the evolving membership of drivers as they move along the highway. In such a way, cluster memberships will evolve over time. This accounts for the fact that driving styles vary according to environmental factors [10]. For example, a driver may be labelled as impulsive in the first 300 metres, but then changes to a more cautious driver as he moves

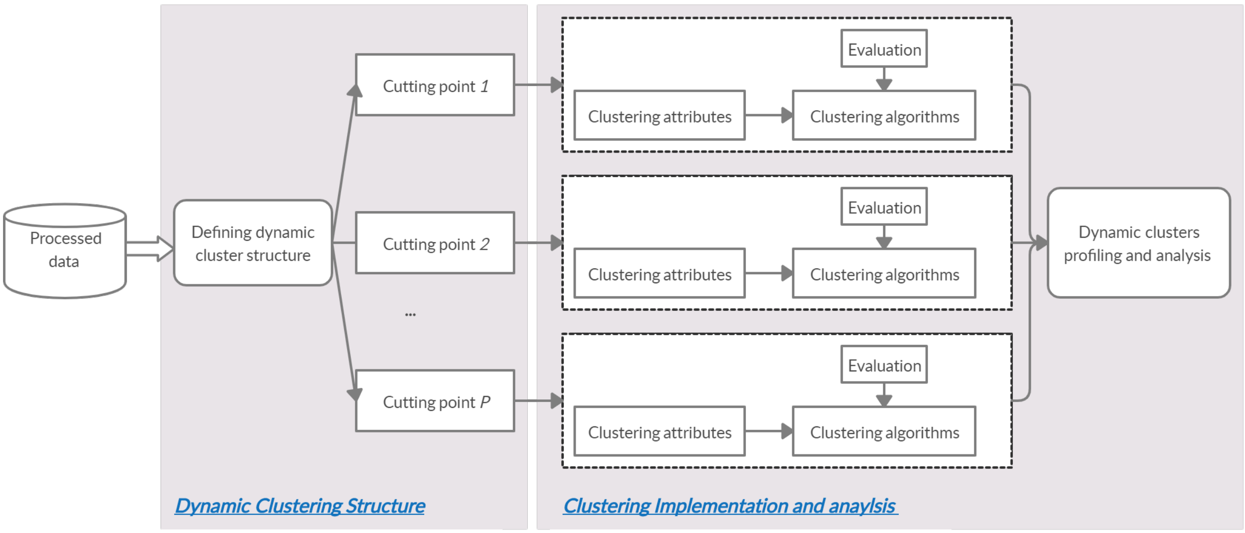


Figure 4: The system framework of proposed dynamic clustering for driving styles analysis.

along the road. The output of this dynamic clustering analysis is a cluster membership sequence for each driver.

The notations for describing the dynamic clustering approach are given in the Table 2 and are used throughout the paper. Note that in our dynamic approach a driver is not assigned to a single cluster but a cluster sequence. In such a way, it accounts for both the cluster membership at a specific point of the highway and also transitions between different driving styles, which helps find more complex driving patterns.

4.3. Clustering attributes

The considered clustering attributes are specified in the Table 3. More specifically, we considered the available information on velocity, acceleration and space headway with the preceding vehicle. We also built four additional variables to identify different behaviours of drivers coming from the on-ramp: first, we included the number of DLCs. Multiple DLC could be a sign of an aggressive driving style; second, we considered the distance the driver uses the auxiliary lane when entering the highway. A driver who uses the auxiliary lane for a safe distance is usually more cautious than a driver who jumps directly into the first lane of the highway; third, we considered two variables to reflect how drivers move horizontally (across lanes)- the difference between the most-right and the most-left lane the vehicle was in and the time elapsed in performing such manoeuvres.

Table 2: Notations for dynamic clustering

N	Total number of vehicles
J	Total number of features
X_{ip}	J-vector of observed features for vehicle i until position p, so that the total data is represented by $\mathcal{X} = \{X_{ip} \mid i = 1, ..., N, p = 300,\ 400, 500, 600\}$
ζ	Set of possible clusters sequences $\zeta = \{S_1, ..., S_M\}$ where M is the number of possible cluster sequences
S_m	Vector of cluster sequence, whose component $C_{k,p}$ is the k-th dynamic cluster of position p and c_{kp} is the centroid of that cluster, $m = 1, 2, ...M$

Table 3: Variables Description

Column Name	Description
VehicleID_Period	Unique vehicle identifier
Period	15-min period the vehicle belongs to
Vel_Mean	Average of the instantaneous velocities of the vehicle in $[km/h]$
Vel_Var	Variance of the instantaneous velocities of the vehicle in $[(km/h)^2]$
Acc_Mean	Average of the instantaneous accelerations of the vehicle in $[m/s^2]$
Acc_Var	Variance of the instantaneous accelerations of the vehicle in $[(m/s^2)^2]$
Space_Hw_Mean	Average of the space with the preceding vehicle in $[m]$
Space Hw_Var	Variance of the space with the preceding vehicle in $[m^2]$
Discretionary_Lc	Total number of discretionary lane changes by the vehicle.
Distance_AuxLane	Distance the vehicle was in the auxiliary lane.
Max_LaneDiff	Number of lanes between the most-right and the most-left lane the vehicle was in.
Time_Max_LaneDiff	Time spent by the vehicle in getting from the most right-lane to the most-left lane.

4.3.1. Percentile variables construction

As aforementioned, the original dataset is segmented into three periods, each with different traffic conditions, going from lower to higher congestion. To account for such facts, the original clustering attributes were transformed to percentile values. That is, for each original clustering variable, a percentile variable was created representing the relative measurements of the vehicle to the rest of the vehicles in the same period. For example, a vehicle in Period 1 that has a percentile variable of 0.20 for velocity is a vehicle whose average velocity (*Vel_ Mean*) is in the 20th percentile of the *Vel_ Mean* distribution of all vehicles in Period 1. These variables were created to have a measurement that is not sensitive to the difference in traffic conditions. Also, using these variables that take values between 0 and 1 serves as a normalisation of the data and could avoid larger scale variables being more influential in the clustering analysis.

4.4. Clustering algorithms

We considered three clustering algorithms in this paper: k-means, spectral clustering and DBSCAN. However, after an initial analysis, DBSCAN is eliminated due to the fact that the density of observations is not heterogeneous enough for this algorithm to work properly. k-means is selected due to its simplicity and good performance. In addition, we adopted k-means++ to improve the initialisation of cluster centroids for k-means. Spectral clustering is considered to identify possible non-flat geometry clusters in the data.

4.4.1. k-means

In order to better describe the k-means algorithm in the context of our problem, we simplify the notations by relaxing the notation p, i.e. assuming a fixed position p. As a result, $C_{k,p}$ becomes C_k, with the centroid denoted as c_k. $|C_k|$ represents the number of observations in the k-th cluster. The feature vector X_{ip} becomes X_i, with x_{ij} being the jth feature value for vehicle i. $\mathcal{X}$ represents all the data points/ observations at the position p. K represents the total number of clusters at the specific position p.

Following [14], the k-means algorithm is described in algorithm 1. In addition, centroids initialization in the step 1 in the algorithm 1 is enhanced using k-means++. Let $D(X_i)$ denote the shortest distance from a data point to the closest cluster centre we have already chosen. The k-means++ is described in the algorithm 2.

4.4.2. Spectral clustering

Spectral clustering has its roots in graph theory where the goal is to identify communities of nodes in a graph based on the edges connecting them. When clustering, the graph nodes are the observations and the edges are given by the affinity matrix, which expresses how similar a pair of data points is to

Algorithm 1 k-means clustering

```
1: Choose the initial cluster centers $c = \{c_1, ..., c_K\}$
2: For each $k$, set the cluster $C_k$ to be the set of points in $X$ that are closer to $c_k$ than to other cluster centers
3: For each $k$, update the cluster center $c_k$ to be the mean attribute vector of all observations in cluster $C_k$: $c_k = \frac{1}{|C_k|} \sum_{i \in C_k} X_i$
4: Repeat steps 2 and 3 until cluster centers $c$ no longer changes
```

Algorithm 2 k-means++

```
1: Select the first center $c_1$ by choosing uniformly at random from $\mathcal{X}$
2: Select a new center $c_k$ by choosing $X_i \in \mathcal{X}$ with weighted probability distribution $\frac{D(X_i)^2}{\sum_{X_i \in \mathcal{X}} D(X_i)^2}$
3: Repeat step 2 until we have chosen all $K$ centers
```

each other. We use the spectral clustering proposed in [15] in this study where algorithmic details can be found in the above cited reference.

4.5. Evaluation

Clustering validation can be categorized into external and internal clustering validation [23]. Due to the fact that the external clustering validation requires the ground truth information (e.g. class labels) which is usually unavailable as in this study, internal clustering validation needs to be used for clustering evaluation. In this paper, we consider two internal evaluation metrics: the Silhouette Coefficient (SC) and the Davies-Bouldin Score (DB). These metrics simultaneously measure the cohesion of the objects within clusters and the separation between clusters. To determine the optimal number of clusters we test different values of K, from 2 to 15, and choose the value that optimizes the evaluation metrics.

4.5.1. Silhouette Coefficient

This coefficient is obtained by contrasting the average distance of objects within the same cluster with the average distance to objects in other clusters. That is, clusters with high similarity within a cluster and low similarity between clusters will have higher SC values.

The SC for the cluster configuration is obtained from the Individual Silhouette Coefficients (ISC). The ISC of a data point X_i is denoted as sil_i and defined as follows:

$$sil_i = \frac{b(X_i) - a(X_i)}{\max\{a(X_i), b(X_i)\}} \tag{1}$$

where $a(X_i)$ is the average distance between point X_i and all other data within the same cluster and $b(X_i)$ is the smallest average distance of X_i to the points in other clusters.

The SC is then given by:

$$SC = \frac{\sum_{k=1}^{K} \frac{\sum_{i=1}^{|C_k|} sil_i}{|C_k|}}{K} \tag{2}$$

where it takes values between -1 and 1. If the SC value is greater, the corresponding cluster configuration is better.

4.5.2. Davies-Bouldin Score

To define the DB score, we first need the definitions of the average distance within cluster (WCD_k) and the ratio of the intra-cluster distances to the inter-cluster distances (R_{kl}).

For a cluster k, we define its average distance within cluster as:

$$WCD_k = \frac{1}{|C_k|} \sum_{i=1}^{|C_k|} d(X_i, c_k) \tag{3}$$

Then, we define the ratio R_{kl} for a cluster k with respect to another cluster l as:

$$R_{kl} = \frac{WCD_k + WCD_l}{d(c_k, c_l)} \tag{4}$$

where $d()$ is the Euclidean distance function and c_k and c_l are the centroids of clusters k and l respectively.

For each cluster k, we define R_k as the maximum ratio ($\max_l R_{kl}$). Finally, DB is given by

$$DB = \frac{1}{K} \sum_{k=1}^{K} R_k \tag{5}$$

where a lower value indicates a better clustering configuration.

4.6. Clusters profiling and analysis

In this paper, clusters are profiled using 4 variables: the mean of the velocity, the variance of the velocity, the mean space headway and the number of discretionary lane changes. We use these variables to characterize the clusters in terms of impulsivity, taking very cautious drivers on one end of the spectrum and impulsive-aggressive drivers on the other. Although in general high velocity is considered as an indicator of aggressive driving styles, we do not consider drivers with higher mean velocities as more aggressive. This is because all the data are taken from periods of traffic congestion and exhibit relatively low velocities. In particular,

- We consider drivers with a larger space headway to be more cautious than drivers that leave very little space with their preceding vehicles.
- We consider drivers that conduct more discretionary lane changes to be more impulsive.
- We consider drivers show a high variance in the velocity to be more impulsive and unpredictable than others.

Based on the clusters profiling at each cutting point and final clustering results, dynamic driving patterns can be captured between different cutting points. In addition, more complex driving behaviours such as driving styles switches along different cutting points can be identified.

5. Results

In this section, we first choose the optimal number of clusters for each cutting point and select the clustering algorithm. Secondly, the dynamic clustering results are presented. Finally, driving styles switching behaviours are discussed.

5.1. Selecting clustering algorithm

We choose the optimal number of clusters for each cutting point based on SC and DB, testing different values of K ranging from 2 to 15.

5.1.1. Cutting point p=300m

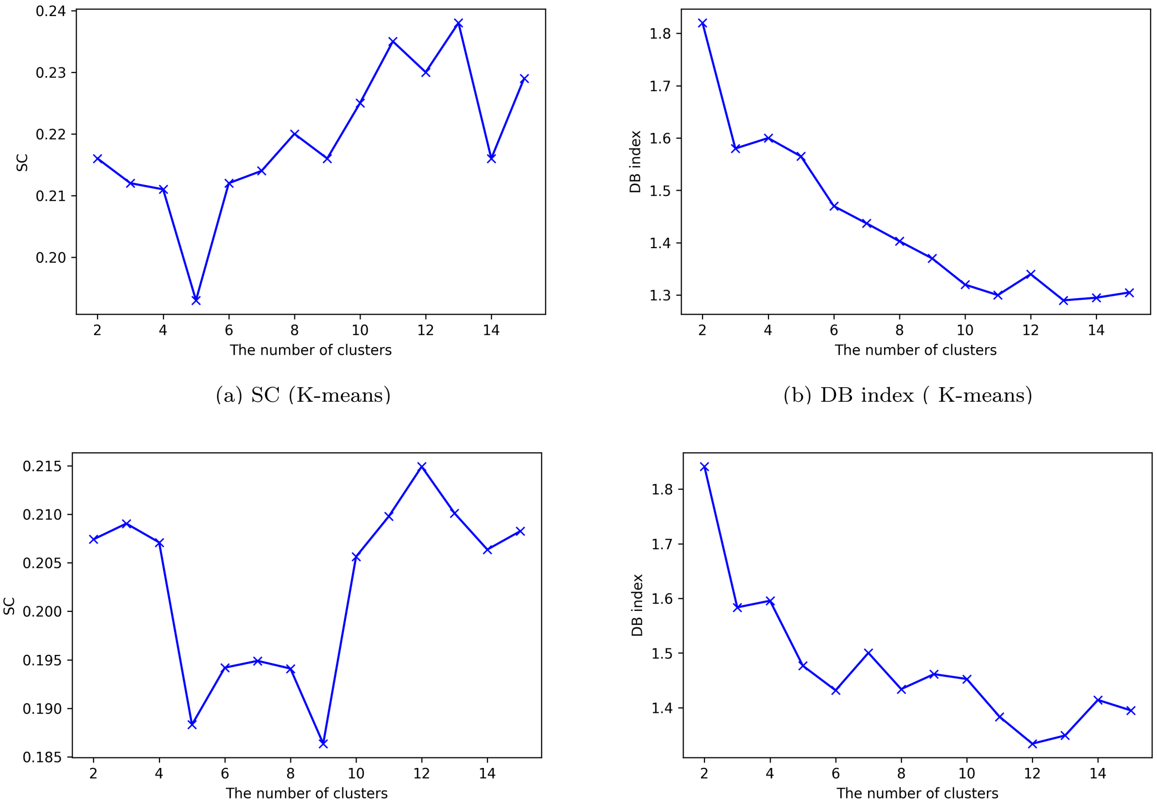
For the cutting point $p = 300m$, the relationship between the number of clusters and the evaluation metrics is presented in the Figure 5. When analysing SC and DB for k-means, we conclude that in both cases the optimal K is 13 where SC reaches its maximum and DB at its minimum. Similarly, for spectral clustering, an optimal K of 12 is selected. However, for the final clustering we choose to use k-means with K=13 as it shows better performance in terms of validity indexes with a SC of 0.238 and a DB of 1.290.

5.1.2. Cutting point p=400m

For this cutting point, the influence of the cluster number on the evaluation metrics is illustrated in the Figure 6. When analysing SC and DB for k-means, we conclude that the optimal K is 7. For spectral clustering, it suggests an optimal K of 6. For the final clustering we choose k-means with K=7 as it shows the best validity indexes, with a SC of 0.211 and a DB of 1.418.

5.1.3. Cutting point p=500m

The influence of the cluster number on the evaluation metrics is illustrated in the Figure 7. For this cutting point, we conclude that the optimal K is 3 for k-means clustering where SC reaches its maximum and DB at its minimum. For spectral clustering, the optimal K is also determined to be 3. For the final clustering, we select k-means with K=3, obtaining a SC of 0.250 and a DB of 1.456.



(a) SC (K-means)

(b) DB index (K-means)

(c) SC (Spectral clustering)

(d) DB index (Spectral clustering)

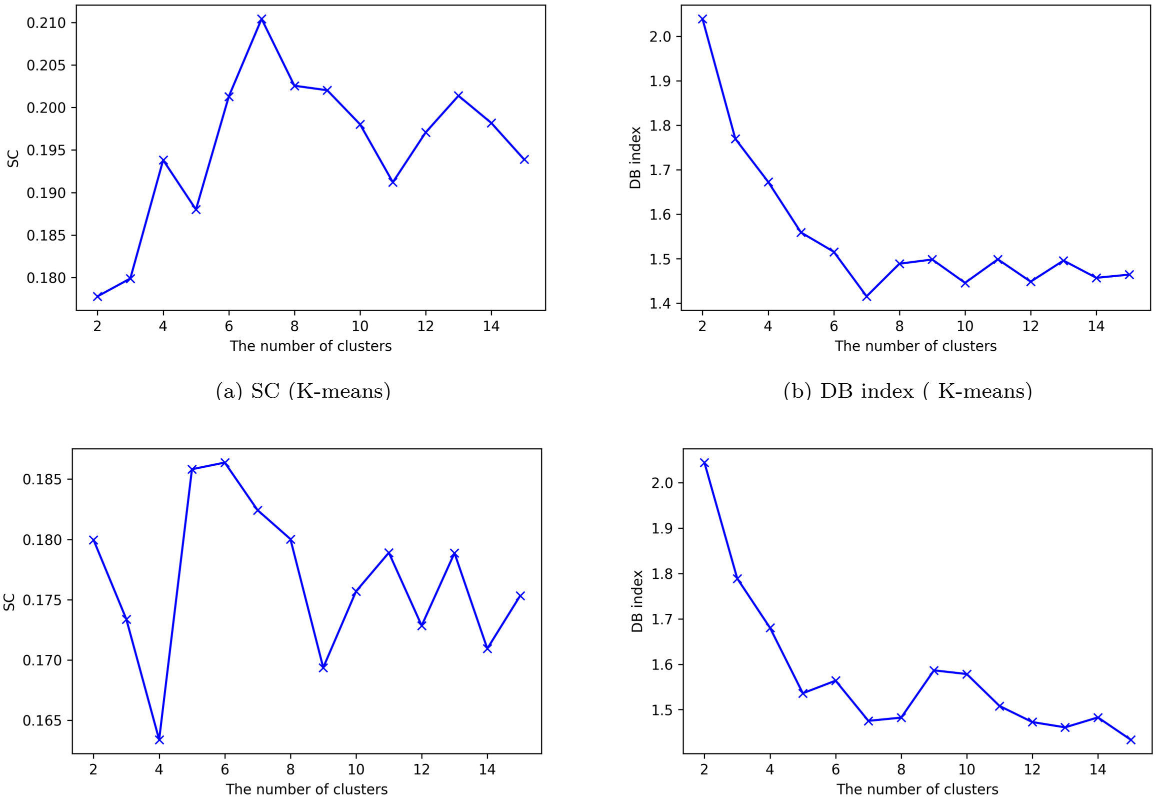
Figure 5: The relationship between the number of clusters and evaluation metrics at cutting point $p = 300m$.

5.1.4. Cutting point p=600m

For the cutting point $p = 600m$, Figure 8 illustrates the influence of the number of clusters on the evaluation metrics. When analysing SC for k-means, we conclude that the optimal K is 2 where SC reaches its maximum. However, DB shows an erratic behaviour with no clear trend. For spectral clustering, the optimal K is also determined to be 2. Since the SC of k-means is higher than spectral clustering and the DB of k-means is lower than spectral clustering at $K = 2$, we select k-means with K=2, obtaining a SC of 0.229 and a DB of 1.702.

5.2. Dynamic clustering results

Table 4 shows the summary statistics for clusters at different cutting points. For each cluster, the mean velocity, the variance of velocity, the mean space headway and the average number of DLC are shown. In brackets there are the standard errors of these statistics. In general, for the first a few clusters one can observe more cautious driving styles with larger space headway, few lane changes and a low



(a) SC (K-means)

(b) DB index (K-means)

(c) SC (Spectral clustering)

(d) DB index (Spectral clustering)

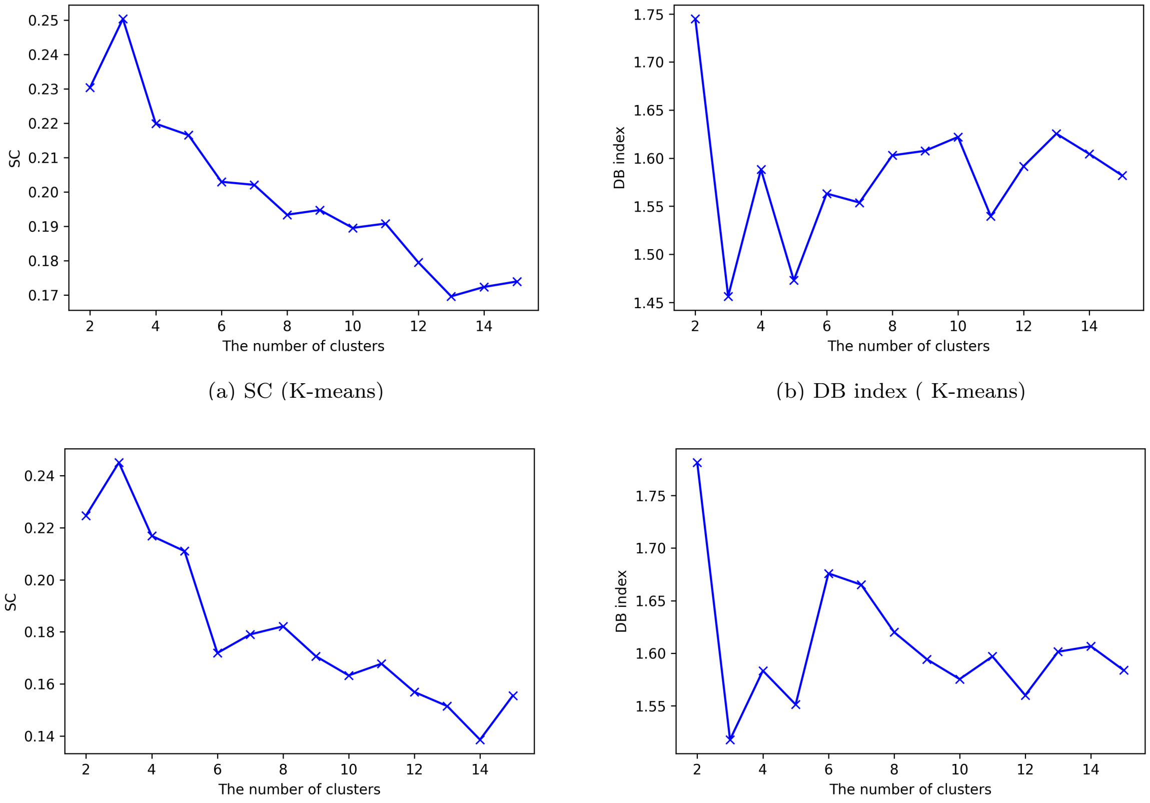
Figure 6: The relationship between the number of clusters and evaluation metrics at cutting point $p = 400m$.

variance in velocity. For the last a few clusters, we can identify more impulsive driving styles with higher velocity variances and more lane changes.

To better understand the lane change dynamics, Figure 9 presents trajectories of vehicles in each cluster at different cutting points where the axes represent space coordinates of the highway. The shaded area in each subfigure represents the section considered in the clustering according to different cutting points. Lanes are termed from 1, the leftmost lane (fastest lane), to 5, the rightmost lane (slowest lane). Lane 6 is the auxiliary lane that comes immediately after the on-ramp.

Specifically, for the cutting point $p = 300m$, clusters $C_{1,300}$ to $C_{4,300}$ do not have many discretionary lane changes (DLC) compared with other clusters. Instead, Clusters $C_{10,300}$ to $C_{13,300}$ show much more discretionary lane change behaviour (see Figures 4 (j)-(m)). Other clusters such as $C_{5,300}$, show a high amount of lane changes but are considered less impulsive. This is because they have larger headway spaces and drive at a more constant velocity.

For $p = 400m$, clusters $C_{1,400}$ and $C_{4,400}$ both show fewer lane changes. However, $C_{4,400}$ is less cautious



(a) SC (K-means)

(b) DB index (K-means)

(c) SC (Spectral clustering)

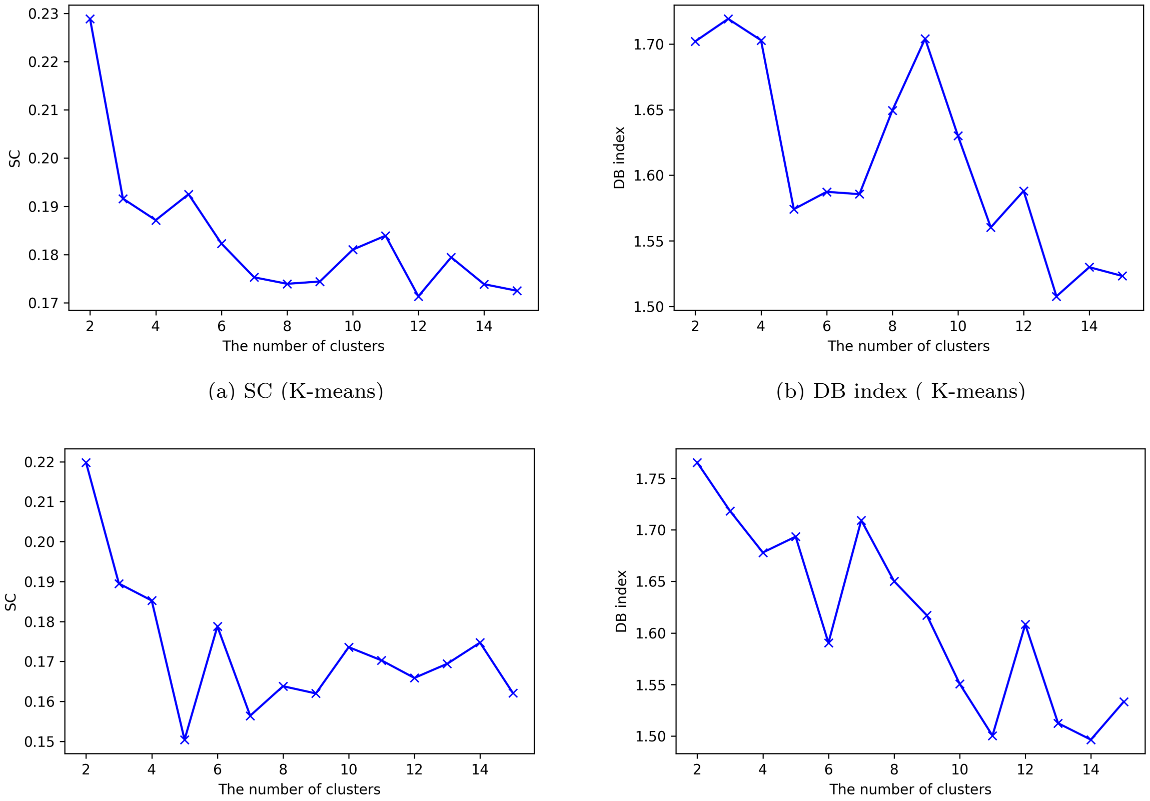
(d) DB index (Spectral clustering)

Figure 7: The relationship between the number of clusters and evaluation metrics at cutting point $p = 500m$.

because it has the smallest space headway. Cluster $C_{7,400}$ is considered to be the most aggressive with the most lane changes, a high variance in the velocity and a small space headway.

For $p = 500m$, we identify three driving styles. Cluster $C_{1,500}$ is the most cautious profile, which shows a large space headway and fewer discretionary lane changes, and low velocity variance. Cluster $C_{2,500}$ is considered to be the moderate driving style. Compared with cluster $C_{1,500}$, it has more lane changes and higher velocity variance. Cluster $C_{3,500}$ is considered to be impulsive with very high velocity variance, small headway space and more lane changes.

Finally, for $p = 600m$, drivers are classified into two groups: a larger group of cautious-to-moderate drivers and a smaller group of impulsive-aggressive drivers. Having a look at the vehicle trajectories, we can find all extreme lane changes (some drivers move from the auxiliary lane to lane 1 in less than 500) occur in the cluster 2.



(a) SC (K-means)

(b) DB index (K-means)

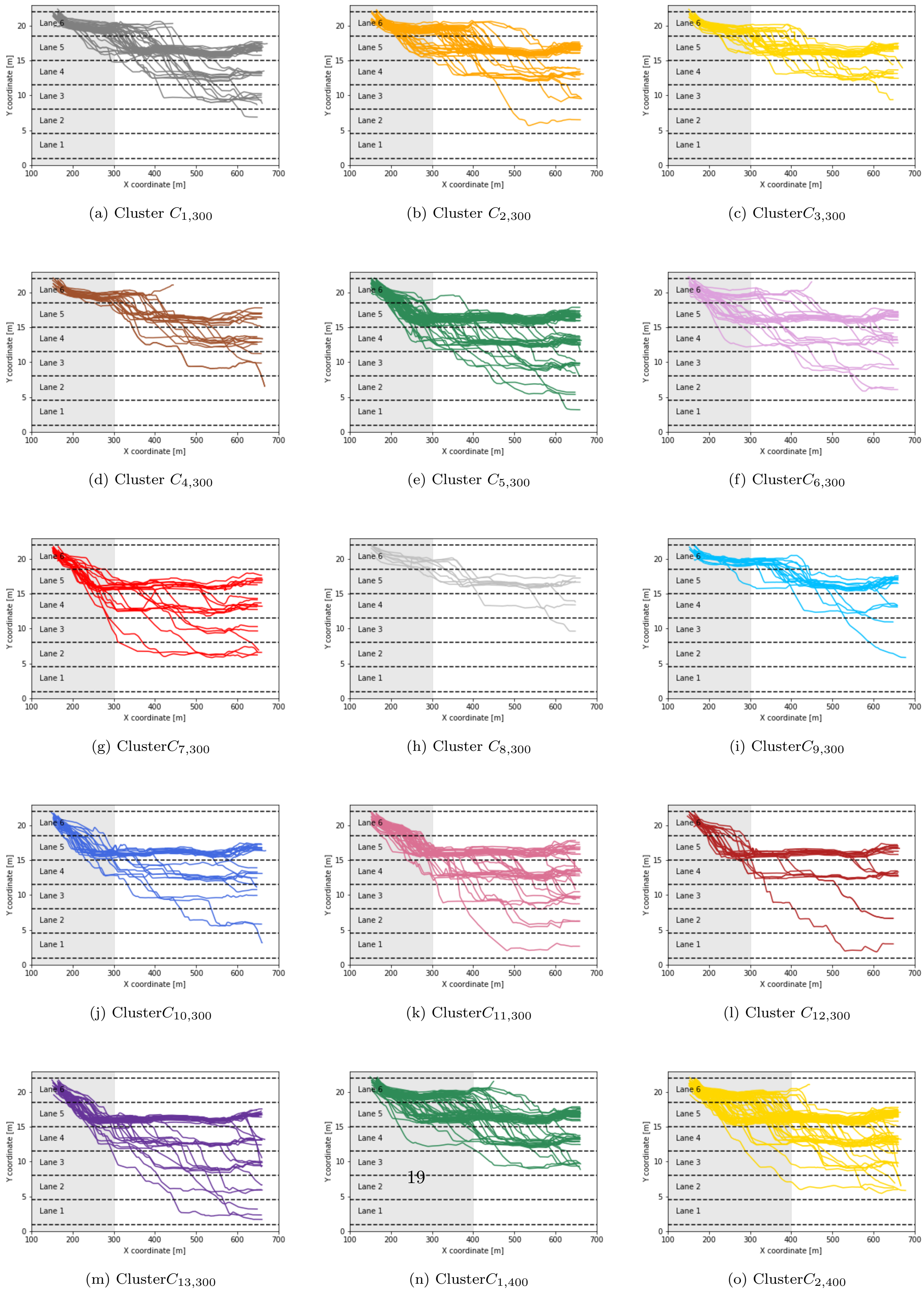
(c) SC (Spectral clustering)

(d) DB index (Spectral clustering)

Figure 8: The relationship between the number of clusters and evaluation metrics at cutting point $p = 600m$.

5.3. Driving Styles Switches

With the above different cutting points, we found a total of 25 static clusters (13, 7, 3 and 2 clusters for each cutting point p respectively) and a set of 136 cluster sequences (ζ) based on the dataset studied. Note that the set of all possible sequences to which vehicles may be assigned could change (e.g. up to $546 = 13 \times 7 \times 3 \times 2$ sequences given the above static clusters) depending on the dataset chosen. In other words, there are no transition links between some clusters across different cutting points identified in this case study. For instance, cluster 6 at the cutting point p=400 meters does not have transition links with clusters 1 and 2 at the cutting point p=500 meters. Using the proposed dynamic clustering approach, it allows us to find more complex driving patterns as it accounts for not only the cluster membership at a specific point of the highway but also the transitions between different driving styles. To better illustrate that, Figure 10 shows all transitions between clusters. We can see that as p changes, some interesting patterns can be visually identified. For instance, we can find some transitions are more likely to occur than others. As an example, a driver of cluster $C_{3,400}$ is more likely to change to $C_{1,500}$ or $C_{2,500}$ than to



(a) Cluster $C_{1,300}$

(b) Cluster $C_{2,300}$

(c) Cluster$C_{3,300}$

(d) Cluster $C_{4,300}$

(e) Cluster $C_{5,300}$

(f) Cluster$C_{6,300}$

(g) Cluster$C_{7,300}$

(h) Cluster $C_{8,300}$

(i) Cluster$C_{9,300}$

(j) Cluster$C_{10,300}$

(k) Cluster$C_{11,300}$

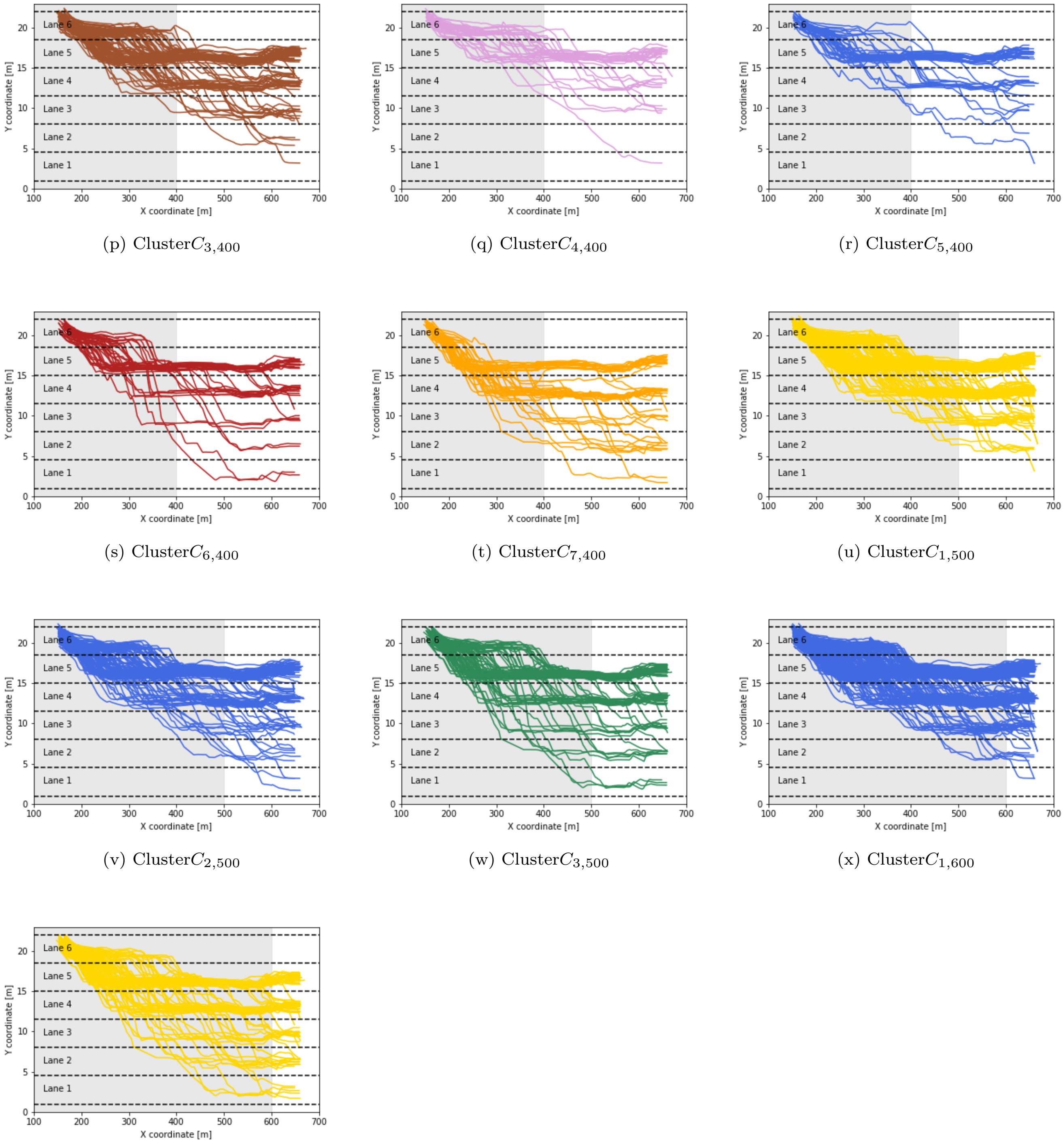
(l) Cluster $C_{12,300}$

(m) Cluster$C_{13,300}$

(n) Cluster$C_{1,400}$

(o) Cluster$C_{2,400}$

Figure 9: Vehicle trajectories by cluster at different cutting points for lane change dynamics understanding. $C_{k,p}$ represents k-th cluster of cutting point position p and disperse trajectories indicate more lane changes in that cluster.



(p) Cluster$C_{3,400}$ (q) Cluster$C_{4,400}$ (r) Cluster$C_{5,400}$

(s) Cluster$C_{6,400}$ (t) Cluster$C_{7,400}$ (u) Cluster$C_{1,500}$

(v) Cluster$C_{2,500}$ (w) Cluster$C_{3,500}$ (x) Cluster$C_{1,600}$

(y) Cluster$C_{2,600}$

Figure 9: Vehicle trajectories (continued) by cluster at different cutting points for lane change dynamics understanding. $C_{k,p}$ represents k-th cluster of cutting point position p and disperse trajectories indicate more lane changes in that cluster.

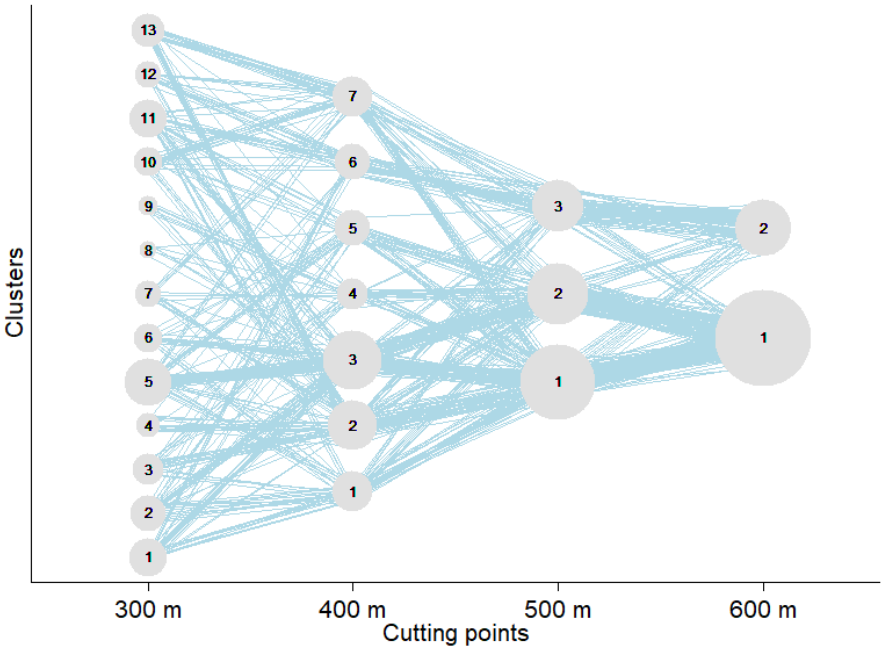


Figure 10: Clusters transitions and driving styles switches.

$C_{3,500}$. In addition, we also find that transitions from highly cautious driving styles to highly aggressive ones are less likely than transitions from e.g. cautious to moderate driving styles.

As can be seen from the above, when we include more information on the flat section of the highway (i.e. with greater cutting points), the optimal number of clusters decreases and driving styles converge into more general driving behaviours. This could imply that drivers change their driving styles depending on surrounding driving environment. On the other hand, the fact that different number of clusters are found by varying p implies in certain situations some defining characteristics are unobservable from the data. Therefore, a dynamic clustering based approach, which can capture dynamic driving patterns, becomes more important and necessary.

5.4. Further Results and Discussion

In this subsection, we include further results based on Fuzzy C-means. Different from K-means and spectral clustering, Fuzzy C-means is a soft clustering algorithm where each observation can belong to multiple clusters with corresponding membership coefficients [24]. Such a clustering algorithm has been used in existing intelligent vehicles studies e.g. [25]. To allow for a direct comparison with other clustering algorithms based on the evaluation metrics considered in the subsection 4.5, we made a modification for the clustering evaluation of Fuzzy C-means by assigning each observation to the cluster with the maximum membership coefficient [26]. As such, SC and DB index can be computed for the Fuzzy C-means. We follow the same experimental process as in the subsection 5.1 to investigate the influence of the number of clusters on the evaluation metrics for Fuzzy C-means where the results can be found in the Figure 11. For cutting point $p = 300m$, the results show that the optimal cluster number based on Fuzzy C-means is either 2 or 3, which is very different from those found by K-means (K=13) and spectral clustering (K= 12). By comparing the SC and DB index corresponding to optimal cluster numbers, it is noted that K-means and spectral clustering achieve better clustering performance. Similar findings apply for the cutting point $p = 400m$. For the cutting point $p = 500m$, both SC and DB index determine the optimal number of cluster to be 3, which aligns with the conclusion of K-means and spectral clustering. For the cutting point $p = 600m$, SC indicates the optimal number of clusters is two whereas the DB index gives a much different choice (K= 15). The inconsistency of evaluation metrics for this cutting point is also observed in the results of K-means clustering and spectral clustering.

Although it is preferable to have more datasets under different situations (e.g. rainy days or night time) to investigate the adaptability of the proposed approach, despite an extensive search we believe such datasets are either not publicly available or simply do not exist. This might be explained by the current data collection practice for intelligent vehicles research which generally falls into two categories: on-board data collection via e.g. sensors, GPS on the vehicle; data collection through external sensors e.g. cameras mounted on the building near the study area. For the former, it is time consuming and expensive, often resulting into small-scale and private data. For the latter, it usually needs extensive post-processing to transform the data e.g. from videos to numerical data. As such, external conditions such as weather will significantly affect the data accuracy and therefore often limit such data collection activities to some particular conditions (e.g. sunny/ daytime) [27]. To mitigate the lack of public datasets under particular situations, a hybrid approach combining real world vehicle trajectory data and simulation data could be developed. Moreover, with the increasing development and deployment of intelligent and connected vehicles, such vehicle trajectory data should be much easier to obtain in the future. Motivated by the above analysis, in the future we plan to conduct a dedicated, systematic and comparative experimental and simulation study by including more benchmarking clustering algorithms and more (and different types of) datasets.

6. Conclusions

In this paper, we propose a dynamic clustering based approach for driving styles identification. We first define the dynamic cluster structure where a cluster sequence rather than a single cluster is used. Second, we consider different cluster algorithms to conduct the clustering by using different cutting points. The proposed dynamic clustering approach, which has the capability to capture dynamic driving behaviours and identify which driving style transitions are more likely to happen given different surrounding environments, is evaluated based on a real world dataset. The proposed approach is particularly important in the context of mixed driving environment consisting of human driven vehicles and autonomous/self-driving vehicles. More specifically, from the perspective of autonomous/self-driving vehicles, the proposed approach will enable the dynamic assessment of driving behaviours of surrounding vehicles (especially the human driven vehicles) to improve the performance of autonomous/self-driving systems. From the perspective of human driven vehicles, the proposed approach embedded with the on-board driving assistance system will provide enhanced driving safety by better understanding the surrounding environment. One potential further work is to use the sequences of clusters to determine the probabilities of adopting particular driving styles given the current and past cluster memberships, which can be combined with relevant decision making methods (e.g. game-theoretic methods [28]) in order to make better decisions for intelligent vehicles based on the perception of surrounding vehicles.

References

[1] R. Kala, K. Warwick, Motion planning of autonomous vehicles in a non-autonomous vehicle environment without speed lanes, Engineering Applications of Artificial Intelligence 26 (5-6) (2013) 1588–1601.

[2] S. Liu, K. Zheng, L. Zhao, P. Fan, A driving intention prediction method based on hidden markov model for autonomous driving, Computer Communications.

[3] C. Hubmann, J. Schulz, M. Becker, D. Althoff, C. Stiller, Automated driving in uncertain environments: Planning with interaction and uncertain maneuver prediction, IEEE Transactions on Intelligent Vehicles 3 (1) (2018) 5–17. doi:10.1109/TIV.2017.2788208.

[4] W. Wang, J. Xi, A. Chong, L. Li, Driving style classification using a semi-supervised support vector machine, IEEE Transactions on Human-Machine Systems 47. doi:10.1109/THMS.2017.2736948.

[5] W. Wang, J. Xi, A rapid pattern-recognition method for driving styles using clustering-based support vector machines, in: 2016 American Control Conference (ACC), 2016, pp. 5270–5275. doi:10.1109/ACC.2016.7526495.

[6] G. Ren, Y. Zhang, H. Liu, K. Zang, Y. Hu, A new lane-changing model with consideration of driving style, International Journal of Intelligent Transportation Systems Research 17 (181). doi:https://doi.org/10.1007/s13177-019-00180-7.

[7] J. Wang, W. Xu, Y. Gong, Real-time driving danger-level prediction, Engineering Applications of Artificial Intelligence 23 (8) (2010) 1247–1254.

[8] W. Zhan, A. L. de Fortelle, Y. Chen, C. Chan, M. Tomizuka, Probabilistic prediction from planning perspective: Problem formulation, representation simplification and evaluation metric, in: 2018 IEEE Intelligent Vehicles Symposium (IV), 2018, pp. 1150–1156. `doi:10.1109/IVS.2018.8500697`.

[9] U.S. Department of Transportation Federal Highway Administration, Next generation simulation (ngsim) vehicle trajectories and supporting data, uS-101-LosAngeles-CA. Provided by ITS DataHub through Data.transportation.gov. Accessed 2019-09-01 from `https://catalog.data.gov/dataset/next-generation-simulation-ngsim-vehicle-trajectories` (2016).

[10] C. Marina Martinez, M. Heucke, F. Wang, B. Gao, D. Cao, Driving style recognition for intelligent vehicle control and advanced driver assistance: A survey, IEEE Transactions on Intelligent Transportation Systems 19 (3) (2018) 666–676. `doi:10.1109/TITS.2017.2706978`.

[11] Z. E. A. Elassad, H. Mousannif, H. Al Moatassime, A. Karkouch, The application of machine learning techniques for driving behavior analysis: A conceptual framework and a systematic literature review, Engineering Applications of Artificial Intelligence 87 (2020) 103312.

[12] Y. Wang, W. Qu, Y. Ge, X. Sun, K. Zhang, Effect of personality traits on driving style: Psychometric adaption of the multidimensional driving style inventory in a chinese sample, PLOS ONE 13 (9) (2018) 1–17.

[13] J. MacQueen, Some methods for classification and analysis of multivariate observations, in: Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability, Volume 1: Statistics, University of California Press, Berkeley, Calif., 1967, pp. 281–297.

[14] D. Arthur, S. Vassilvitskii, k-means++: The advantages of careful seeding, Tech. rep., Stanford (2006).

[15] A. Y. Ng, M. I. Jordan, Y. Weiss, On spectral clustering: Analysis and an algorithm, in: Advances in neural information processing systems, 2002, pp. 849–856.

[16] M. Ester, H.-P. Kriegel, J. Sander, X. Xu, A density-based algorithm for discovering clusters a density-based algorithm for discovering clusters in large spatial databases with noise, in: Proceedings of the Second International Conference on Knowledge Discovery and Data Mining, KDD'96, 1996, pp. 226–231.

[17] A. Bouchachia, Dynamic clustering, Evolving Systems 3 (3) (2012) 133–134.

[18] S. Aghabozorgi, A. S. Shirkhorshidi, T. Y. Wah, Time-series clustering - a decade review, Information Systems 53 (2015) 16 – 38.

[19] A. Fowler, V. Menon, N. A. Heard, Dynamic bayesian clustering, Journal of Bioinformatics and Computational Biology 11 (05) (2013) 1342001.

[20] K. I. Ahmed, Modeling drivers' acceleration and lane changing behavior, Ph.D. thesis, Massachusetts Institute of Technology (1999).

[21] A. van Beinum, H. Farah, F. Wegman, S. Hoogendoorn, Driving behaviour at motorway ramps and weaving segments based on empirical trajectory data, Transportation Research Part C: Emerging Technologies 92 (2018) 426–441.

[22] C. Thiemann, M. Treiber, A. Kesting, Estimating acceleration and lane-changing dynamics from next generation simulation trajectory data, Transportation Research Record 2088 (1) (2008) 90–101.

[23] Y. Liu, Z. Li, H. Xiong, X. Gao, J. Wu, Understanding of internal clustering validation measures, in: 2010 IEEE International Conference on Data Mining, IEEE, 2010, pp. 911–916.

[24] J. C. Bezdek, Pattern recognition with fuzzy objective function algorithms, Springer Science & Business Media, 2013.

[25] D. Yi, J. Su, C. Liu, W.-H. Chen, Data-driven situation awareness algorithm for vehicle lane change, in: 2016 IEEE 19th International Conference on Intelligent Transportation Systems (ITSC), IEEE, 2016, pp. 998–1003.

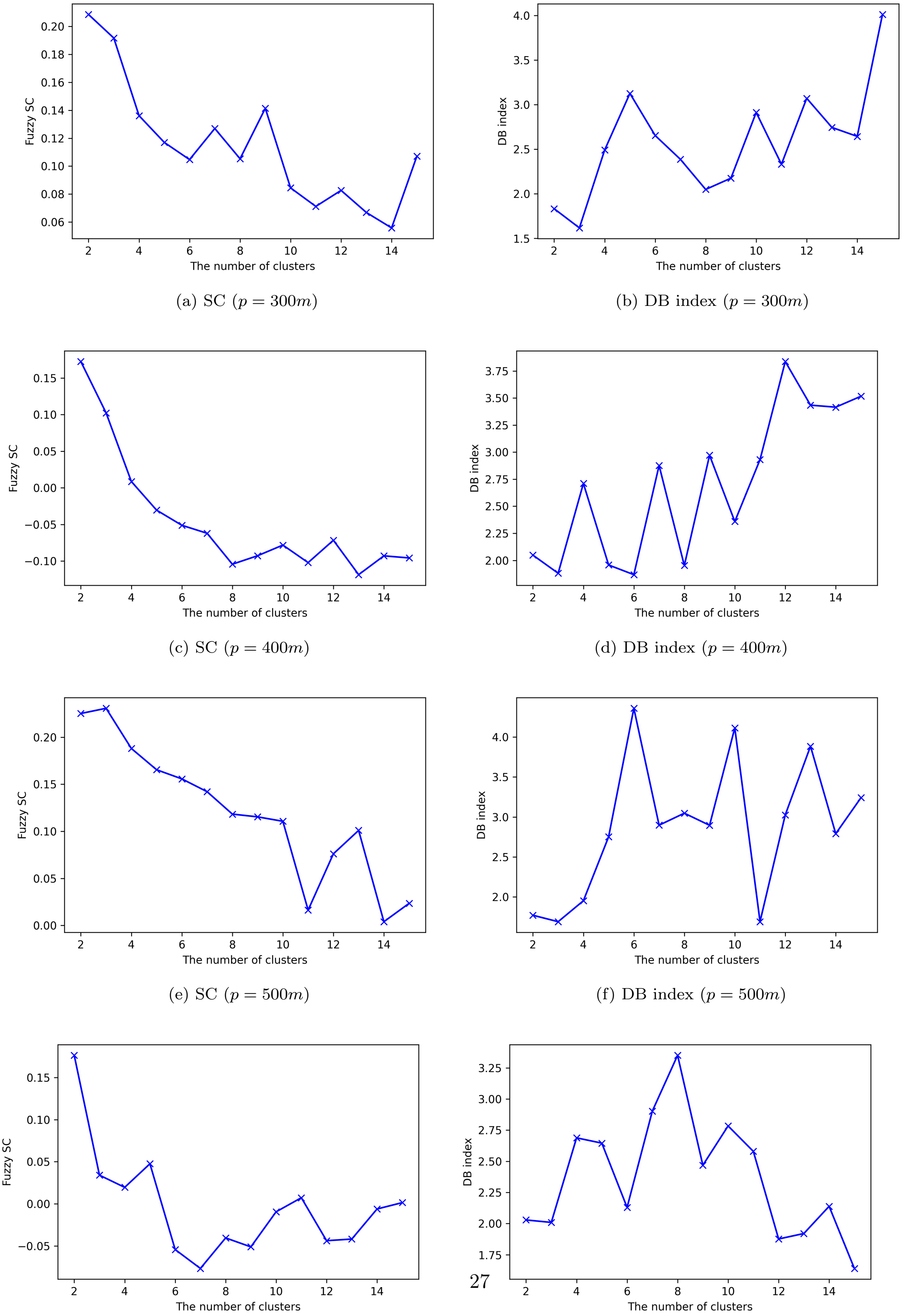
[26] R. J. Campello, E. R. Hruschka, A fuzzy extension of the silhouette width criterion for cluster analysis, Fuzzy Sets and Systems 157 (21) (2006) 2858–2875.

[27] R. Krajewski, J. Bock, L. Kloeker, 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), IEEE, 2018, pp. 2118–2125.

[28] F. Meng, J. Su, C. Liu, W.-H. Chen, Dynamic decision making in lane change: Game theory with receding horizon, in: 2016 UKACC 11th International Conference on Control (CONTROL), IEEE, 2016, pp. 1–6.

Table 4: Clusters Summary Statistics.

	p=300 m.				p=400 m.				p=500 m.				p=600 m.			
Cluster	Mean Vel [km/hr]	Var. Vel [(km /hr)^2]	Mean Space Hw [m]	Mean DLC [#]	Mean Vel [km/hr]	Var. Vel [(km /hr)^2]	Mean Space Hw [m]	Mean DLC [#]	Mean Vel [km/hr]	Var. Vel [(km /hr)^2]	Mean Space Hw [m]	Mean DLC [#]	Mean Vel [km/hr]	Var. Vel [(km /hr)^2]	Mean Space Hw [m]	Mean DLC [#]
1	51.17 (8.89)	14.56 (13.00)	42.25 (17.23)	0.00 (0.00)	53.03 (9.64)	30.92 (37.26)	30.90 (10.07)	0.05 (0.22)	44.63 (7.74)	36.11 (33.00)	22.77 (8.93)	0.45 (0.67)	46.82 (8.55)	47.73 (37.19)	24.07 (9.14)	0.54 (0.77)
2	51.09 (9.68)	15.91 (22.90)	38.24 (16.27)	0.02 (0.15)	45.32 (7.88)	40.86 (37.31)	12.04 (6.14)	0.21 (0.53)	48.89 (9.02)	50.41 (43.00)	25.26 (11.09)	0.52 (0.76)	31.14 (9.32)	176.75 (72.88)	18.24 (4.66)	1.15 (1.23)
3	54.99 (8.03)	8.04 (8.61)	12.89 (6.00)	0.00 (0.00)	46.00 (8.45)	32.14 (36.02)	28.87 (8.00)	0.26 (0.46)	29.72 (8.30)	162.92 (80.37)	17.26 (4.75)	0.72 (1.02)				
4	52.25 (9.44)	12.46 (21.45)	12.90 (3.97)	0.00 (0.00)	51.49 (7.81)	25.88 (24.25)	11.79 (5.54)	0.10 (0.30)								
5	46.4 (6.34)	8.33 (9.38)	30.18 (10.52)	0.10 (0.30)	43.52 (7.48)	41.69 (37.85)	31.45 (8.93)	0.30 (0.57)								
6	54.34 (8.77)	8.19 (8.90)	37.25 (13.19)	0.04 (0.2)	23.21 (5.40)	198.97 (86.81)	18.11 (6.13)	0.33 (0.72)								
7	45.92 (4.79)	7.68 (8.72)	12.06 (5.93)	0.30 (0.57)	36.79 (7.91)	40.11 (26.11)	16.54 (4.73)	0.83 (0.82)								
8	58.16 (10.22)	12.98 (8.90)	39.29 (13.66)	0.00 (0.00)												
9	51.28 (12.04)	36.66 (33.84)	8.58 (6.27)	0.00 (0.00)												
10	40.34 (6.16)	21.14 (19.66)	16.84 (7.46)	0.08 (0.28)												
11	39.99 (10.33)	49.99 (40.96)	31.18 (10.76)	0.10 (0.30)												
12	27.37 (10.60)	122.95 (84.23)	11.26 (3.98)	0.11 (0.32)												
13	34.73 (8.45)	27.45 (10.60)	13.04 (3.72)	0.16 (0.37)												



(a) SC ($p = 300m$)

(b) DB index ($p = 300m$)

(c) SC ($p = 400m$)

(d) DB index ($p = 400m$)

(e) SC ($p = 500m$)

(f) DB index ($p = 500m$)

(g) SC ($p = 600m$)

(h) DB index ($p = 600m$)

Figure 11: The relationship between the number of clusters and evaluation metrics for Fuzzy C-means at different cutting points p.