Driving Style Recognition: Literature Review and Application of Machine Learning

Fahrstilerkennung: Forschungsüberblick and Anwendung von machinellem Lernen
Master-Thesis von Sabina Kruk aus Krakau, Polen
Tag der Einreichung:

1. Gutachten: Prof. Dr. Johannes Fürnkranz
2. Gutachten: M.Sc. Quoc Hien Dang

TECHNISCHE UNIVERSITÄT DARMSTADT

Knowledge Engineering Group
Computer Science Department

Driving Style Recognition: Literature Review and Application of Machine Learning
Fahrstilerkennung: Forschungsüberblick and Anwendung von machinellem Lernen

Vorgelegte Master-Thesis von Sabina Kruk aus Krakau, Polen

1. Gutachten: Prof. Dr. Johannes Fürnkranz
2. Gutachten: M.Sc. Quoc Hien Dang

Tag der Einreichung:

Erklärung zur Master-Thesis

Hiermit versichere ich, die vorliegende Master-Thesis ohne Hilfe Dritter nur mit den angegebenen Quellen und Hilfsmitteln angefertigt zu haben. Alle Stellen, die aus Quellen entnommen wurden, sind als solche kenntlich gemacht. Diese Arbeit hat in gleicher oder ähnlicher Form noch keiner Prüfungsbehörde vorgelegen.

Darmstadt, den 1.5.2017

____

(Sabina Kruk)

Abstract

Intelligent assistance systems for drivers is a wide topic of research nowadays. The human component, driving behavior, is the main source of accidents on the road. These assistance systems aim at reducing the driving workload to increase safety and increase the quality of the driving experience. To develop such intelligent systems which will adjust to the driver, it is crucial to investigate the driver's behavior and habitual driving patterns, also referred to as driving style. A lot of research, therefore, focuses on determining the existing driver styles. It is crucial to keep in mind, that in order to use the term driving style, the observed traits must be independent of the traffic situation.
One of the main methods to determine driving styles is the use of machine learning, especially clustering. This thesis provides a review of the approaches and makes use of the findings to cluster drivers from a provided dataset, according to their driving style. The investigation focuses on the bahavior of drivers at intersections. The results are validated against results of driving behavior investigation.

Contents

1 Introduction

Nowadays, motor vehicles are equipped with an increasing amount of electronics and sensors to build advanced driver support systems. These assistance systems are supposed to increase the quality of the driving experience and reduce the driving workload to increase safety. Examples of such developments are Adaptive Cruise Control (ACC) [72] [25], which performs the longitudinal following control task for a driver, within limited acceleration ranges, Collision Warning (CW) [34] [51] or Lane Departure Warning (LDW) [69] [45]. The performance of such driver assistance systems depends strongly on the current driving context. It is crucial to develop the support system taking into account different traffic situations and environments as well as the drivers' behavior. Such holistic approaches are the basis to provide the system driver vehicle with adequate support which can ultimately result in a comfort gain for the driver, increase in reliability under changing circumstances or an increase of the security of the whole system. For this reason, building models of driver behavior and developing methods to describe driving habits is a huge field of research. The motivation of this thesis firstly, is to provide an overview of the approaches developed to describe and classify driver behavior. Secondly, the aim is to develop a method relying on these findings, that will help to cluster a given dataset. The clusters should represent different types/styles of driver behavior, that are coherent across various traffic contexts.

1.1 Driver Behavior Definition

This thesis focuses on approaches to access drivers' styles alias types. Before getting further into the topic, it is necessary to explain what is meant by a driver's style. There is still no fixed definition of driver style across literature. The concept often involves other constructs such as driver state, driver condition and driver behavior in general. Some state-of-the-art definitions include the one of [16]

> style concerns the way individuals choose to drive, or driving habits that have become established over a period of years.

and [76]

> Driving style concerns individual driving habits - that is, the way a driver chooses to drive

They are very similar to each other and agree that driver style is the habitual way drivers choose to drive or driving preferences they have developed over time.

In [63], the author dealt with the review of different definitions and concluded that a driver style is:

> ... a habitual way of driving which means that it represents relatively stable aspect of driving behavior.

and must differ across individuals or between groups of individuals.

This thesis adapts this definition. Thus, a driving style must be habitual and differ across individuals, not the driving conditions or driving environments. The motivation for this thesis is to determine whether it is possible to group drivers based on a data set recorded during a driving simulation. The literature research is the basis and guide line for the development of the approach.

1.2 Thesis Outline

The thesis is structured as follows. Chapter 2 focuses on the driver styles/categories described in literature and the motivation behind a specific classification. It must be noted that not only approaches regarding driver classification are invoked, but also methods of analyzing and examining driver behavior. They are a valuable source of preprocessing and feature extraction methods and actually make up the biggest part of the literature review. The chapter also contains an overview of the input parameters and features used to determine the driver types or driving behavior. The criteria for determining a specific category are laid down.

Chapter 3 presents the provided data set, including anomalies and observations that might be crucial for developing the right clustering algorithm. This is followed by Chapter 4, which presents the pipeline used to analyze the given dataset in order to search for patterns and clusters.

The pipeline is then applied on the dataset in Chapter 7 and the results presented. Chapter 8 makes an attempt at interpreting the observations made from the experiments. An outlook on future work and conclusions are added in Chapter 9.

2 Driver Categories

In the course of this thesis drivers should be grouped according to their driving styles. The obvious question that arises is how to judge whether the obtained groups are useful and meaningful and how many of them seem a reasonable number.

For this reason it is important to have an overview of the current approaches to driver typification in terms of what are the types introduced so far and what is the motive for this particular choice.

Approaches for determining the driver categories can be split into two kinds, according to whether the categorization is based on subjective assumptions and assessments or is a result of some classification of measurement data. The objective approaches are presented in section 2.1, followed by the subjective approaches in section 2.2.

2.1 Objective Approaches

The author of [57] strives for improving safety on signalized intersections by studying the factors influencing it. The author cites [22] and [47] who state that safety on signalized intersections can be assessed by observing the driver behavior in the so called dilemma zone (DZ). It is an area near the intersection where drivers traveling at the legal speed limit can neither stop nor clear the intersection successfully. One of the traits of a well designed signalized intersection, must be the elimination of any DZ for the drivers who travel at the legal speed limit or slightly above it. For speeds lower than the speed limit, an option zone (OZ) should be created, i.e. an area where either stopping or crossing can be exercised successfully. [57] believes that examining the decision of drivers who find themselves in the dilemma zone as well as their approaching speed is crucial for determining the intersection's safety record and making the right adjustments to the intersection design to make crossing it as safe as possible. The categories proposed by [57] are normal, aggressive and conservative. They are obtained through a two-step classification process. In the first step it is checked whether the driver's approaching speed exceeds a defined speed limit by some percentage. If so, the driver is automatically put in the aggressive group, otherwise considered as nonaggressive. The second step uses the behavior of the drivers when they face a yellow signal as the criterion for classification. The driver can either choose to stop before the crossing or continue driving. Taking this into account and keeping in mind that the driver might happen to be in the option or dilemma zone, three categories are presented, namely conservative, normal and aggressive. Conservative is the driver who stops, even when she could safely exit the intersection. Normal is the driver who acts as expected by the intersection design, either in the case of a dilemma zone or in the case of an option zone. So, in the case of being caught in a DZ stopping should be a normal action and crossing considered an aggressive action. Finally, aggressive drives are those who speed up and cross the intersection when they find themselves in the DZ zone. This type of classification is therefore a process-based classification. Classifications based on these two criteria, i.e. the approaching speed and the choice to stop or not are inter-related. By combining the results of both classification steps, a final classification of drivers is possible.

The authors of [33] concern themselves with dual-power vehicles such as hybrid electric cars. Their goal is to develop a driver type classifier which could be used to determine the optimal strategy for shifting from one of the hybrid power sources to the other and estimate the available reserves of energy. The authors also set for three categories of drivers, namely aggressive, moderate and conservative. The classes are determined empirically.

The authors of [10] use clustering to identify six different driving styles. The features used for the clustering are reduced to components using Principal Components Analysis. Combinations of these components are then referred to as aggressiveness, speed, accelerating and braking. Their manifestations in turn establish six clusters, representing six driver categories.

The work of [4] presents a ranking approach, which arranges drivers according to values for a number of sensor inputs recorded during their driving. The author makes a differentiation of three driving styles, each one representing a part of the ranking scale. Together with data recorded during road tests, the categorization of drivers was a foundation for the classification of driving style.

In comparison [40] came up with six different driving types throughout their research but limited the number of driver styles. Four rules are used to create six distinctive representative driving patterns which are then combined to three driver types, namely: low power demand, medium power demand and high power demand. Each class represents different types of standard deviation in power demand. The first class for example represents typical urban driving patterns where the average power level is low but the variation in power is large due to frequent stop-and-go traffic conditions. The class six, on the other hand, resembles suburban driving patterns where the average power level is high and the standard deviation in power is relatively small.

The article of [23] examines drivers according to the level of fuel economy they pursuit. The authors chose three classes according to a dynamic factor. It is generally acknowledged that the depth of the accelerator pedal reflects the demand of the driver for the current vehicle driving force. Moreover, pedal change rate reflects the demand for fierce change degree of vehicle driving force. Economical driving would manifest itself in medium-sized or small throttle opening and operating the accelerator pedal smoothly. The speed would also be kept at a level making the fuel consumption throughout the drive as small as possible. On the contrary, large throttle opening and changing the position of the accelerator pedal intensively, as well as high speed would contribute to big fuel consumption and thus, uneconomical driving. Considering this, the author developed a dynamic factor ranging from 0 to 1 and reflecting the economy of the driver. 0 represents economical driving and 1 driving for dynamic performance rather than fuel economy. Three driver styles were picked to represent the whole of the dynamic factor scale, namely sports, moderate and economical. Dynamic drivers fall into the category of sports driving style where as attentive drivers frequently tend to demonstrate economy driving style.

In [53], the authors extract driver categories according to their rate of change in acceleration or deceleration, alias jerk. The authors believe that jerk is a more effective feature in driver style classification than just acceleration. They argue, that while an acceleration profile shows how a driver speeds up and slows down, a jerk profile shows how a driver accelerated and decelerated, which is more important in determining the driver's aggressiveness. The classification is carried out as follows: if the standard deviation of the jerk exceeds the average jerk of the road-type the driver is on, then the driver will be classified as aggressive. On the contrary, if the standard deviation of the jerk is much lower than the average jerk of a normal driver; it will be classified as calm, unless the velocity is zero. The reason for using the jerk of the normal driving style on the road-type is currently on as a feature in the classification is that the author believe a driver's driving style is strongly influenced by the roadway type and traffic congestion level the driver is on. Three categories are picked: calm driving, normal driving and aggressive driving.

The work of [30] shows that driver safety can be deduced from driver style if it is defined in terms of the driver's predictibility and the level in which the way they execute manoeuvers complies with what is observed as average and standard. Only sensors provided by mobile phones are used to infer the driver style and two categories are established. Either the driver is aggressive (non-typical) or not (typical).

The authors of [13] also support the claim that that the most typical classification of driver styles confines to the two categories, they refer to as aggressive and nonaggressive, depending on their predictability and consistency in driving behavior. They argue that choosing merely two categories preserves generality.

In [35] an objective as well as a subjective approach is used to determine the driver style. Concerning the objective method, three classes are assumed in the project, sports, normal and calm. The three different driver styles are in accordance with the work of [8]. The specific driver styles are acquired mainly through the observation of maximal longitudinal and lateral acceleration alias maximal longitudinal lag. Authors of [35] claim that having so many classes lacks generality and that three driver style categories are enough to influence the characteristics of a driver assistance system.

In his paper, [48] refers to the S.A.N.T.O.S [35] project. He points out that the classification of drivers into as many as five or six categories is inconvenient and invokes the S.A.N.T.O.S report in this context. For this reason, [48] decides to restrict himself only to three categories, namely sports, normal and calm.

Heike Sacher [62] also uses a combination of different approaches to define the driver styles. Again, in this section we will concentrate on the objective methods. The drivers completed four experimental trips, during which four coefficients were computed. Each coefficient was assigned a weight value to emphasize its contribution in describing a driver style. Considering previous literature, the author chose lateral acceleration to be the most influential factor, thus denoting it highest weight. The choice of three driver styles was also made according to the findings of previous research. The categories are again sports, normal and relaxed.

The study of [61] investigated the effect of driving style on the energy consumption and the potential to reduce the consumption using driver assistance in trucks. Four driver styles are presented in the paper: dynamic, effective driver style, undynamic, ineffective driving style, undynamic, effective driving style and dynamic, ineffective driving style. Dynamic was determined by the longitudinal and lateral acceleration and the vehicle's velocity, whereas effectiveness by the motor and shift-up speed. The explanation for this choice is that all these aspect are influenced by the driver himself.

At this point it is pretty clear that throughout the literature, when dealing with measurement vehicle data, researchers tend to classify the drivers into three categories. Even though the criteria may vary, the categories seem to describe similar tendencies across literature. The review of [75] already summarized that driving styles are usually clustered into three categories; "mild" drivers (calm driving or economical driving style), "normal" drivers (medium driving style), and "aggressive" drivers (dynamic driving style).

There are however opinions that two categories is enough, because it preserves generality, as seen in the case of [13] or [30].

Works of [50] (smooth, aggressive), [2] (aggressive, moderate), [9] (calm, aggressive), [42] (aggressive, moderate), comply with this idea and also propose two categories.

2.2 Subjective Approaches

The goal of this thesis is to find a method to cluster drivers according to their driver styles, given measurement data. It is still interesting however to take a look at approaches where self-report measures of driving behavior are considered. Knowing how the drivers judge themselves, a part of evaluation of the clustering approach developed in this thesis, could be the comparison between the drivers' own assessment and the cluster they belong to according to their actual measurement data. Among these self-reported measures, the easiest and most common way to determine drivers' driving style is with the help of a questionnaire.

The paper of [27] includes reviews of several driving style questionnaires that have been presented in the literature. The first one is the Driver Style Questionnaire (DSQ) introduced by [21]. Analyzing the answers for 15 driving style questions lead to the discovery of six dimensions labeled speed, calmness, planning, focus, social resistance (advise) and deviance. Together they describe the relation between driving style, decision-making style and accident liability .

The authors of [28] suggested a Driving Style Questionnaire with eighteen questions separated over eight components. The assumption underlying the questions was that driving style is an attitude, orientation and way of thinking for daily driving. Each component included questions that were supposed to reveal a particular aspect of the driving style. These included confidence in driving, hesitation for driving, impatience in driving, methodical driving, preparatory manoeuvres at traffic signals, importance of automobile for self-expression, moodiness in driving and anxiety about traffic accidents. The questionnaire was validated by analyzing car following behavior at low speed. An example of an evaluation finding is that there is a positive correlation between confidence in driving skill and the use of the gas pedal.

A very widely used inventory is the Multidimensional Driving Style Inventory (MDSI) [71]. Four general driving styles were distinguished [27]:

- Reckless and careless driving, which is correlated with violations and thrill seeking while driving, characterized by, for example, higher speed.
- Anxious driving, referring to feelings of alertness and tension.
- Angry and hostile driving, characterized by more use of the horn and flash functionality.
- Patient and careful driving, reflecting a well-adjusted driving style.

These four styles were the basis to create the 44 items for the MDSI.

The items focused on accessing the drivers' their feelings, thoughts, and behaviors on a 6-point scale ranging from 1 (not at all) to 6 (very much).Each driver's responses on the relevant scales were averaged to produce four driving style scores, with a higher score indicating a higher level of the particular style.

Analysis of the questionnaire answers revealed eight main factors influencing the assignment to a particular style [27]:

- Dissociative driving, in which people are easily distracted and dissociated during driving.
- Anxious driving, in which people show signs of anxiety and lack of confidence.
- Risky driving, in which people seek for sensation and more risky driving.
- Angry driving, in which people tend to be hostile and aggressive.
- High-velocity driving, in which people tend to drive faster and are more time driven.
- Distress-reduction driving, in which people engage in relaxing activities to reduce stress.
- Patient driving, in which people are polite to other road users and have no pressure of time.
- Careful driving, in which people drive carefully and structured.

To note is, that the MDSI [16] includes, among others, items from the previously mentioned inventories, making it a very popular means of self-reported driving style assemessment.

Some approches develop their own self-reporting questionnaires to complement their findings made through data measurements.

This is for example the case for the project S.A.N.T.O.S [35]. In his work [14], the author mentiones the self-reporting methods used for classifying the drivers in the project. One of the groups taking part in the S.A.N.T.O.S project used classification instruments by [3] where the drivers were supposed to determine their driving style by filling out a polarity questionnaire. For each question there was a 6-level scala, with opposed statements on the poles of the scala. This grading was supposed to differentiate between the driving styles. [3] claims that three polarities are enough to determine a diver's attitude. The first polarity factor could be described as attitude to the traffic and other traffic participants, the second as personal attitude to the driving activity itself, and the third as assessment of ones own skills. In the next step the group investigated th correlation between the three factors introduced by [3] and the drivers' velocity profiles. The three factors were used to calculate a coefficient, which denoted 1 for defensive and calm drivers and 6 for drivers reffered to as aggressive and dynamic. The second self-reporting tool was a standarized questionnaire based on literaure research. The questionnaire was concipted to classifiy the drivers into five types:

- reckless, sporty use
- dynamic, progressive
- experienced, serene
- conservative, low-key
- anxious, reserved

Another example is the work of Heike Sacher [62]. Three methods were applied to determine the driving style. One of them was a self-assessment method where the probands ere simply supposed to choose which term described their driving behavior best, very dynamic, dynamic, comfort-oriented and very comfort-oriented. There was also a questionnaire, including 10 questions. These questions were to be answered in form of a value on a scale. Drivers were interviewed beforehand to concept the items. A driver was then classified as dynamic or comfort-oriented by calculating the median of the scale points obtained on the questionnaire.

2.3 Correlation between Objective and Subjective Approaches

The scores obtained on questionnaires must correspond to the actually observed driving behavior. Otherwise they do not fulfill their purpose as indicators of driving style [64]. It was also found that there was a correlation between the six driver styles and the profiles of lateral and longitudinal acceleration.

The review of [64] describes some of the approaches to verify whether such correlations exist. The first to mention is [77]. In their work, correlations between observations made by observers sitting in the vehicle and self-reported driving styles of the drivers are studied, The self-reported instrument here is the Driving Style Questionnaire (DSQ) [21]. High correlations were found for speed (Pearson correlations between 0.55 and 0.65) and more moderate for calmness (0.39 – 0.41), attentiveness (0.29) and carefulness (0.38).

The authors of [28] also found significant correlations between some of the factors of their Driving Style Questionnaire and observed driving style. To investigate this, they made a car-following study using an instrumented vehicle. The highest correlations were found with gas and brake pedal operations during deceleration.

In [19] it was found that the high scores on the Multidimensional Driving Style Inventory (MDSI) [71] "angry and hostile driving style" scale were significantly correlated with both higher speed (r=0.32) and shorter passing gaps (r = -0.20).

Also Heike Sacher [62] looks into the correlation between self-reporting and classification based on measurement data. She discovers that the correlation between the assessment based on measurement data and the assessment based on the questionnaire score was significant, namely 0.45.

The project S.A.N.T.O.S [35] mentions the attempts of [46] to establish a link between the types of drivers determined thorugh a self-assessment questionnaire and the types obtained from driver dynamic parameters. The questionnaire used was the one by [3]. It could be shown that the driver described as dynamic in questionnaire, typically manifests more sporty longitudinal and lateral accelerations.

To sum up, the significant associations between objectively measured behaviour and the one reported, implies that self-report instruments can still play a significant role in driving style research [64].

2.4 Context

When trying to classify according to driver style, it is crucial to examine the full context in which driving occurs. The context provides important information that might influence the way the drivers behave and thus, influence or distort the classification results. [49] for instance states that developing effective counter measures for reinforcing safe and smooth operation of an vehicle in traffic, the full context in which driving occurs should be taken into account. Furthermore, the authors go on to define three main components of the overall driving context.

- Environment: including roadway infrastructure and the dynamic climatic situations.
- Vehicle: including ever increasing telematic devices and infotainment gadgetry.
- Driver: an essential part of the human-vehicle system which needs to be maneuvered safely in the environment.

Driving behaviour might vary systematically across different road, traffic and driving conditions, such as traffic density, road geometry, weather, light conditions etc. Drivers may manifest different patterns of behaviour in different conditions or the same patterns because the situation forces them too (for example driving slowly when it is snowing).

Driving style however is supposed to vary systematically between individual drivers or groups, independent of the traffic situation.

Vehicle data depends on two factors: the drivers individual driving style and the driving context in which the vehicle is operating. In order to evaluate driving style properly, we must take the current driving context into account, filtering out the aspects of the vehicle data that is a result of the current driving context and not the driver's driving style.

For this reason, it is crucial to exclude behaviour patterns that are exclusively determined by the driving context from the definition of driving style. The research of [52] outlines the most common types of driving events that manifest different behaviors, independently of the driver's global style. The events are:

- driving a vehicle along right curves
- driving a vehicle along left curves
- turning a vehicle left on intersections with roundabouts
- turning a vehicle right on intersections with roundabouts
- turning a vehicle left on intersections without roundabouts
- turning a vehicle right on intersections without roundabouts
- driving straight across an intersection with a roundabout

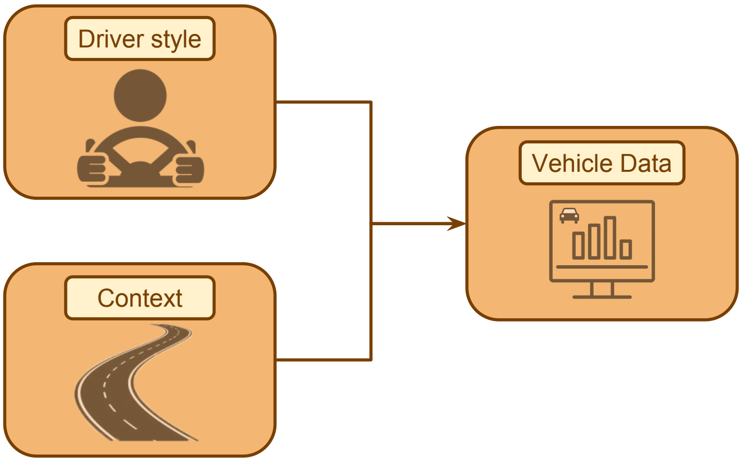


Figure 1: Driver style

2.5 Meaningful inputs

The crucial issue in this thesis is selecting features to use for the clustering. Depending on the driving context, different features might turn out to be relevant. Most driving situation that a driver might encounter may be described by examining the drivers braking, accelerating and turning behavior [56] [73]. The thesis focuses on crossing intersections, where no turning maneouvers take place, so this aspect will be omitted. The ramaining two behaviors, braking and acceleration, can be characterized by a subset of features. These are the features that are broadly and consistently used in research when investigating driver behavior.

The first feature is velocity. It can be found in as an input parameter in most of the approaches, such as [57] [33] [61] [20] [56] [50] [2] [9] [29] [38] [68] [54].

Acceleration is the next straight-forward parameter that is often measured.

The next group of useful features are the pressures applied to the brake and acceleration pedals. Pressure, alias brake torque and throttle is used among others by [49] [37] [33] [20] [56] [73] [18] [55]. Acceleration pedal pressure is considered by [33] [6] [56] [20] [55] [37].

Natrually, many of the mentioned approaches apply more complex feature engineering than just considering these simple parameters. The authors of [33] for instace, present the so called throttle activity index. It denotes the magnitude of the accelerator pedal relative to the frequency of change in the accelerator pedal percentage. Accelerator pedal percentage is the proportion between the driver's pedal position and the position recorded when the pedal is fully depressed.

There are also some opinions that these such a simple description of accaleration and braking are not sufficient. The works of [7] [6] [53] use jerk as their feature instead. The argument is that jerk is a more effective feature in driver style classification than just acceleration. They argue, that while an acceleration profile shows how a driver speeds up and slows down, a jerk profile shows how a driver accelerated and decelerated, which is more important in determining the driver's behavior pattern.

In addition to these features, attributes auch as driver's head and gaze direction ([55] [39]) stand out in the literature and might be relevant for intersection traffic situations.

Steering wheel parameters, on the other hand, even though commonly used as well ([55] [49] [33] [36] [39]), are rather relevant when the drivers must turn. This is not very useful for the provided dataset, where the driver does not turn on any of the intersections.

2.6 Parameter Interpretation

At this point it is pretty clear that aggressiveness is a common term for investigating driver styles. It generally describes maladaptive and risk-related behavior in traffic and is defined by a combination of several behavioral indicators, such as the driving speed, headway, overtaking of other vehicles and the tendency of commiting traffic violations [64].

Aggressive driving is commonly characterized by higher speed, higher acceleration and braking peaks (jerky driving in general), short headway keeing and distance to passing cars, large throttle opening and changing the position of the accelerator pedal intensively [20] [70] [38].

In [78] it is mentioned that visual search patterns are associated with driver experience. Concerning overtaking in this case, experienced drivers tend to allocate their viewpoints more widely in the horizontal plane and farther in the longitudinal direction. Novice drivers, on the contrary, tend to pay more attention to the more narrow scope in front of them.

2.7 Clustering Method

Unsupervised learning is the task of finding a function to model structure in "unlabeled data", meaning data where a categorization or classification of the observations is not given beforehand [31]. Clustering is one of the approaches to unsupervised learning. It aims to group the data observations, so that the observations in the same group (cluster) are more similar (according to some metric) to each other than to those observations assigned to a different group.

As already observed by [41], fuzzy control theory and K-means algorithm are the most common methods to cluster the feature parameters that describe the driver behavior characteristics, in order to achieve classification of the driver behavior characteristics into driver styles [12] [44] [60] [43] [75] [23] [73].

In [1] a good explanation can be found of how the algorithm works.

> The K-means algorithm provides a simple and easy way to classify a given dataset, where the number of clusters k, that should be produced, is set up in advance. The algorithm is executed with the following steps:
>
> 1. Place k points into the space represented by the data samples that are being clustered. These samples represent the initial cluster centroids.
> 2. Assign each data sample to the cluster that has the closest centroid.
> 3. When all data samples have been assigned, recalculate the positions of the k centroids. The new centroids as barycenters of the clusters formed in the initial step.

4. Repeat Steps 2 and 3 until the centroids no longer move. This produces a separation of the objects into groups from which the metric to be minimized can be calculated.

The metric to be minimized is the least within-cluster sum of squares or, the squared error function:

$$\sum_{i=1}^{k} \sum_{x \in S_i} \|x - \varphi_i\|^2 \tag{1}$$

where k is the number of clusters, S_i the subset of data samples belonging to cluster i and φ_i the mean (centroid) of data samples in S_i.

In this thesis the K-means clustering algorithm is used, due to its simplicity, covenience and common use throughtout literature.

The given dataset consists of parameters captured by sensors in a driving simulator. Each of 34 test subjects took the same route twice, resulting in 68 data samples. Each sample consists of data points recorded at a 100 Hz frequency and each data point includes multiple parameters. These parameters can be divided into different categories: simulation parameters such as the length of the covered track or the subject's gender; vehicle parameters such as the vehicle's position in the lane, driver tracking parameters like brake pedal actuation, road features as well as traffic features. Here is a complete overview of the different parameter types [74]:

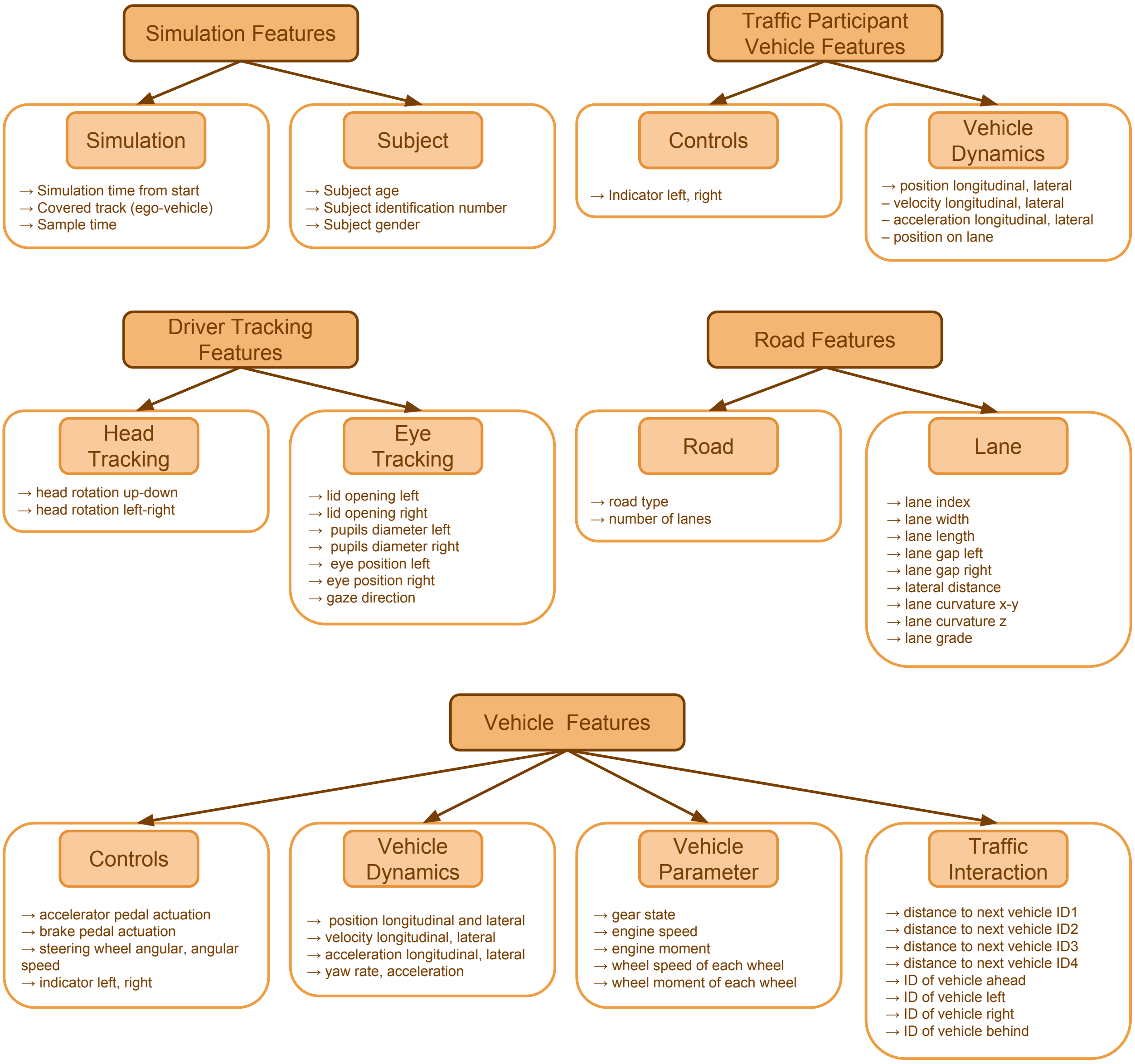


Figure 2: Parameters

The route guidance and the traffic situations are the same for all trips. The route runs through a village, thus the speed limit for the whole route is 50 km/h.

The route is created with focus on a specific traffic situation, namely intersections. There are 25 intersections along the route. Each of the intersections has to be crossed. The intersections along the route are three- or four-way intersections without traffic lights. At 18 intersections the priority of traffic is given by priority to the right, in the other seven cases the driver has the right of way due to traffic signs. At five of the intersections the driver has to give a vehicle from the right the priority in way. Below is an overview of the intersections, starting from the second intersection [74]. The first intersection is left out because it is the point where the drivers start from a rest stop and the data might be noisy. Thus index 0 actually specifies the second encountered intersection.

Intersection	Type	Right of Way	Traffic
0	three-way	priority sign	-
1	three-way	priority sign	oncoming
2	three-way	right over left	oncoming
3	three-way	right over left	-
4	four-way	right over left	-
5	four-way	right over left	left
6	four-way	right over left	right
7	four-way	right over left	oncoming
8	three-way	right over left	-
9	four-way	priority sign	left
10	four-way	priority sign	-
11	four-way	priority sign	oncoming
12	three-way	right over left	oncoming
13	three-way	priority sign	right
14	four-way	right over left	left
15	three-way	right over left	right
16	four-way	right over left	-
17	four-way	right over left	-
18	four-way	right over left	oncoming
19	four-way	priority sign	right
20	three-way	right over left	right
21	three-way	right over left	right
22	four-way	right over left	right
23	three-way	right over left	oncoming

Table 1: Crossings in the simulation route

In order to explore the drivers data and investigate how it can be clustered, multiple, sequenced steps have to be undertaken [5]. Together they build a data processing pipeline which is depicted below. As the first step it is crucial to choose the context, i.e. crossing group that is to be inverstigated and determine potential features that might be relevant in that specific cotext. In the second step the dataset is standarized and any data points considered abnormal are removed for further analysis. Step 3 involves the Principal Component Analysis which helps to reduce the dimension of the feature space, at the same time extracting as much information as possible from the dataset. In step 4 the actual clustering takes place and step 5 assumes interpreting the clustering results. Depending on the conclusions drawn from the interpretation, there might be cues how to change the feature space or the context used, in order to obtain more satisfying results. The following sections will deal with each of the stages in the pipeline in more detail and explain how they were carried out.

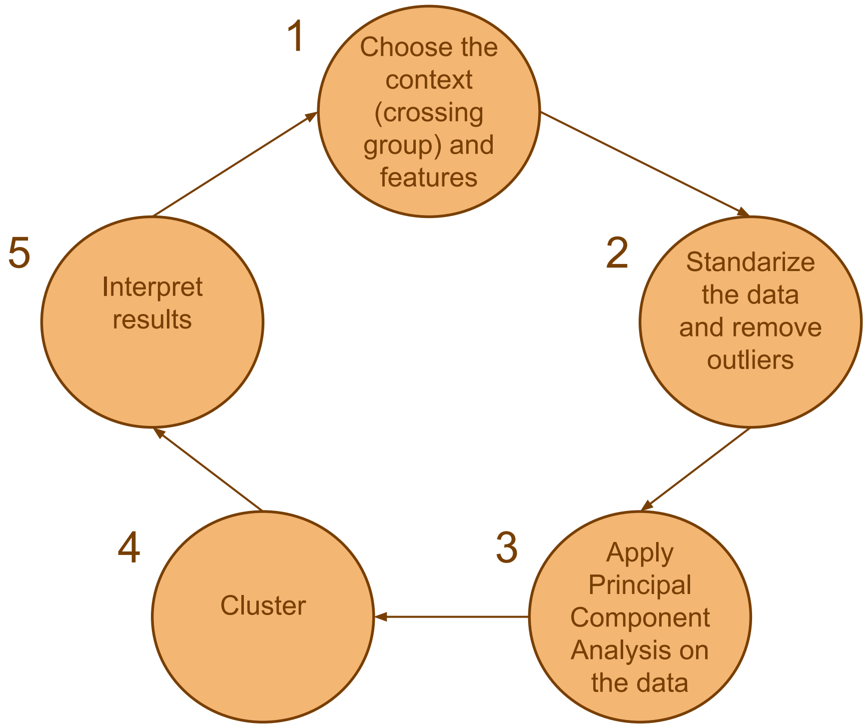


Figure 3: Data Analysis Pipeline

4.1 Choosing the Context and Features

Research proves that the driving context is crucial for exploring driving patterns. Depending on the traffic situation, different features might be relevant and the drivers might manifest varying behaviours depending on the context. The assumption made in this thesis is that the manifestation of features built from the given vehicle data is influenced by the specific traffic situation. Only after obtaining the driving patterns for various traffic situations it will be investigated if there exist distinct characteristics for the drivers, stable across the various traffic situations. Only then, could these characteristics imply a specific driver style (see section 2.4).

Regarding the choice of features, the approach in this thesis is to first determine which features/parameters could be most meaningful based on the findings from the literature and then see how subsets and combinations of them influence the clustering results. Exploring the feature distributions and correlations will help to determine the best feature combination. The clustering performance of for a given set of features might vary depending on the traffic context, making both of these aspects interrelated.

4.1.1 The Context

In the given data set, the intersections are the most distinct traffic situation and the focus will be set on them. The intersections however differ between each other as well. The most prominent distinction is that at some of the intersections the driver has the right of way and at some he does not. Table 1, presenting the intersections, references each one with an index, thus the intersections 0, 1, 9, 10, 11, 13, 19 impose the right of way and the rest do not. First of all the data for all the crossings will be investigated and afterwards, it will be split according to the traffic context, in this case meaning the crossing type. This will enable to determine whether the patterns found are actually stable across different traffic situations and thus indicate a driver style. The Table 2 depicts the various contexts considered.

Case Index	Intersection Type	Traffic	Indices
1	all crossings	all traffic types	range from 0 to 23
2	priority sign crossings	no traffic from the right	0, 1, 9, 10, 11
3	priority sign crossings	traffic from the right	13, 19
4	right over left crossings	no traffic from the right	2, 3, 4, 5, 7, 8, 12, 14, 16, 17, 18, 23
5	right over left crossings	traffic from the right	6, 15, 20, 21, 22

Table 2: Crossing groups

4.1.2 Crossings extraction

There is no parameter that indicated whether the driver is at an intersection on not. The information has to be obtained indirectly from other parameters. For each sample trip it can be well recognized when the driver is crossing an intersection by looking at the time series for lateral distance. When entering an intersection, one can observe a positive peak of the lateral distance which turns to a negative peak when the driver is leaving the intersection. For each driver and each trip, the point in time when the positive and negative peak appeared, are both registered. However, the behaviour of a driver already changes before the exact moment of entering the intersection and also after leaving it. To allow more room for establishing pattern differences, it is necessary to extract a time window not only between the positive and negative lateral distance peak, but some time before the positive one and some time after the negative one. It was decided to observe the driver 100 m before entering the intersection, on the intersection and 50 m after leaving the intersection.

For each driver and each trip, the covered distance is checked at the time of entering and leaving the intersection. This is done by looking up the value of the parameter distance coverage for the time values retrieved for the positive and negative lateral distance peaks. Then, 100 m is substracted from the route coverage before the intersection and 50 m added to the route coverage after leaving the intersection.

Again, the time value corresponding to these two distance coverage values is checked. Now, for every driver and every trip, two time values define the time window relevant for further feature extraction. These time windows will be referred to as crossing time windows throughtout the thesis.

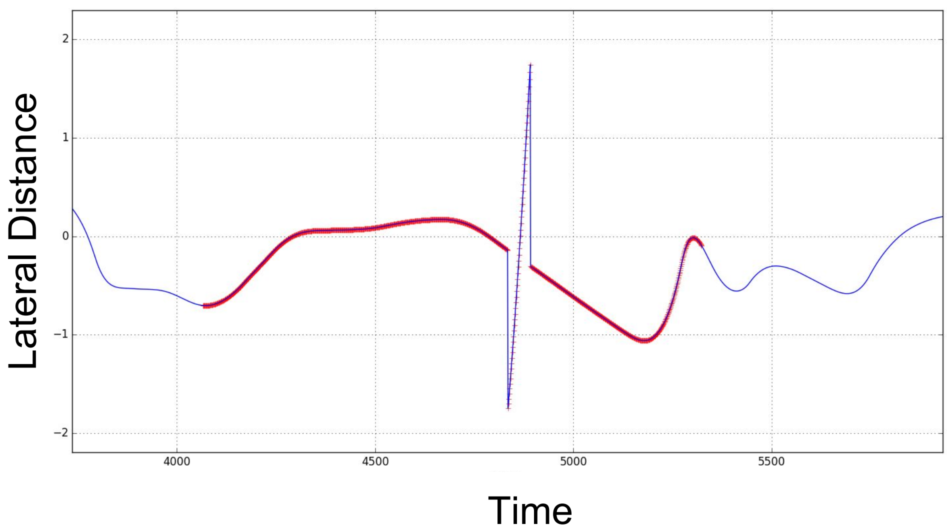


Figure 4: Extracted crossing for one of the trip samples, the red interval corresponds to the extracted crossing window.

4.1.3 Potential Feature Choice

In section 2.5 the most common features the researches worked with and found relevant are presented. Running in line with these findings, the following features are established: jerk standard deviation in longitudinal direction, mean velocity in longitudinal direction, mean of the brake pedal pressure, standard deviation of the brake pedal pressure, mean of the acceleration pedal pressure, standard deviation of the acceleration pedal pressure and last but not least, standard deviation of the head and gaze direction.

The authors of [53] showed that considering only the statistical acceleration properties (such as mean acceleration or variance of the acceleration) was not sufficent to differentiate between drivers and what turned out to be much more successful was observing the change in acceleration. Motivated by the authors idea, for each driver, his both trips and crossings, the longitudinal acceleration is measured for the corresponding crossing time window. Then the first discrete difference is calculated with a period shift of 10 time units (meaning 0.1 s because the data was sampled at a 100 Hz rate). The result is the so called jerk and describes the change of acceleration over the crossing time window. As described in [53], the next step is to compute the standard deviation of the jerk across time. The chosen window length is 30 time steps this time. The actual feature value for the crossing is the mean of this jerk standard deviation. The plot below shows the profile of jerk standard deviation for the two of the crossings (index 0 and 4) and the first two drivers. It is pretty clear that the first driver obtains higher values than the second one, during both trips.

Additionally features concerning braking and accelerating are also created separatly. For this reason, analogously, the standard deviation of braking and acceleration profiles are built. The profiles are based on the corresponding pedal pressure profile. To allow more flexibility in the feature choice, two different

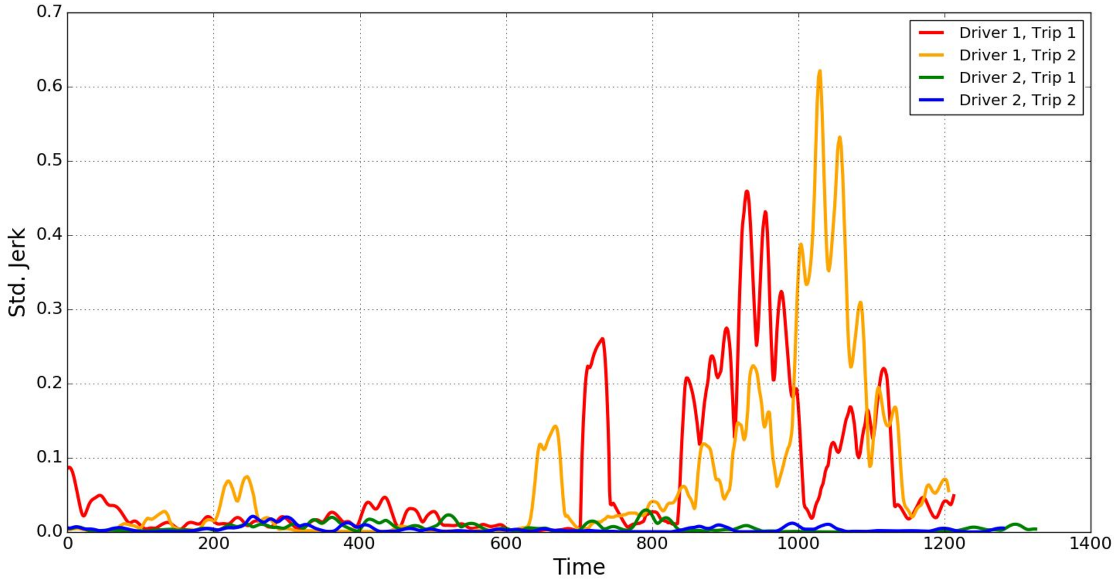


Figure 5: Jerk standard deviation for the crossing with index 0 the first two drivers

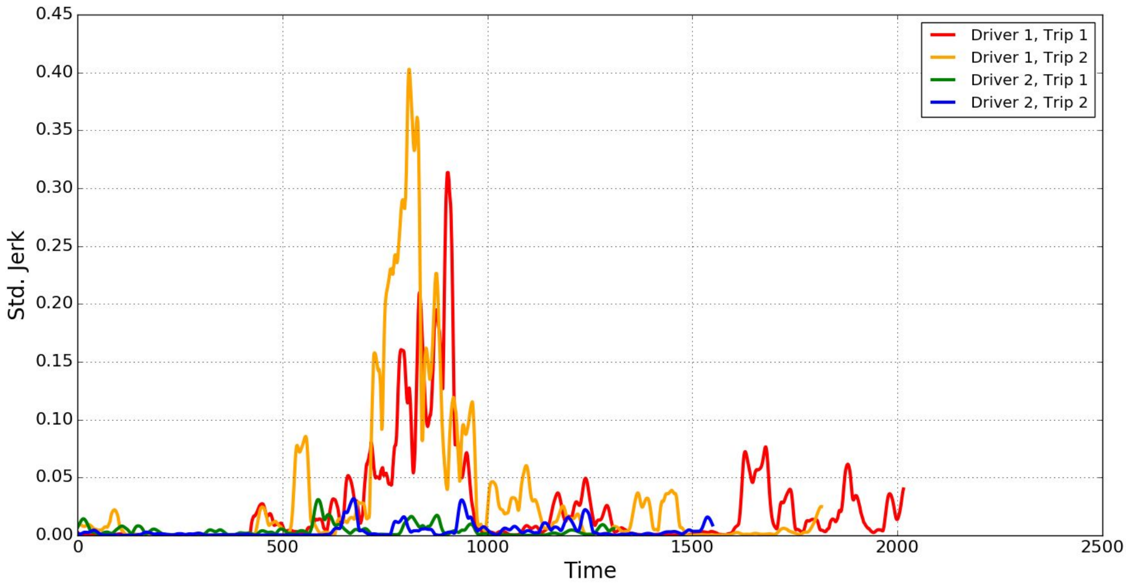


Figure 6: Jerk standard deviation for the crossing with index 4 the first two drivers

windows are picked. The first discrete difference of the acceleration/brake pedal pressure is calculated with a period shift of 5 and also with a period shift of 10. The standard deviation of this difference is then calculated with a window length of 20 and 10, accordingly. Additionally, the potential feature space includes the mean velocity of the driver, as well as the mean pressure of the acceleration and brake pedal.

Last but not least, not only vehicle attributes are considered, but also the ones of the drivers themselves. This includes gaze and head direction.

The changes in head and gaze direction are also added to the potential feature space in the form of mean standard deviation. Again, two pairs of window lenghts are used; 5 and 10 for the first discrete difference, 10 and 20 later for the standard deviation, accordingly.

4.1.4 Sample Formulation

The question that arises at this point is how to construct the feature vectors that are later fed to the clustering algorithm. More specifically, how to build the feature vctors to include the information from multiple crossings for a driver.

The most straight forward variant would be creating a sample for each driver, each trip and each crossing. This yields $n \cdot m \cdot k$ samples, n denoting the number of drivers, m the number of trips each driver has to undertake and k the number of crossings in the route. Each of these samples is a vector containing features extracted for the particular driver, trip and crossing.

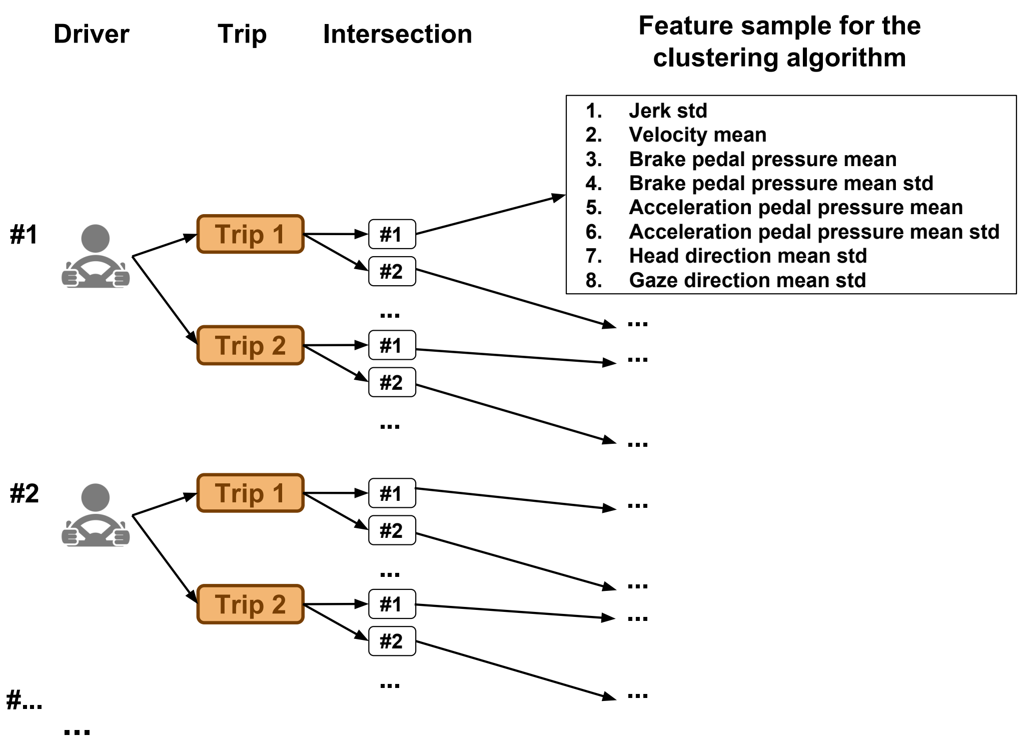


Figure 7: Sample Formulation

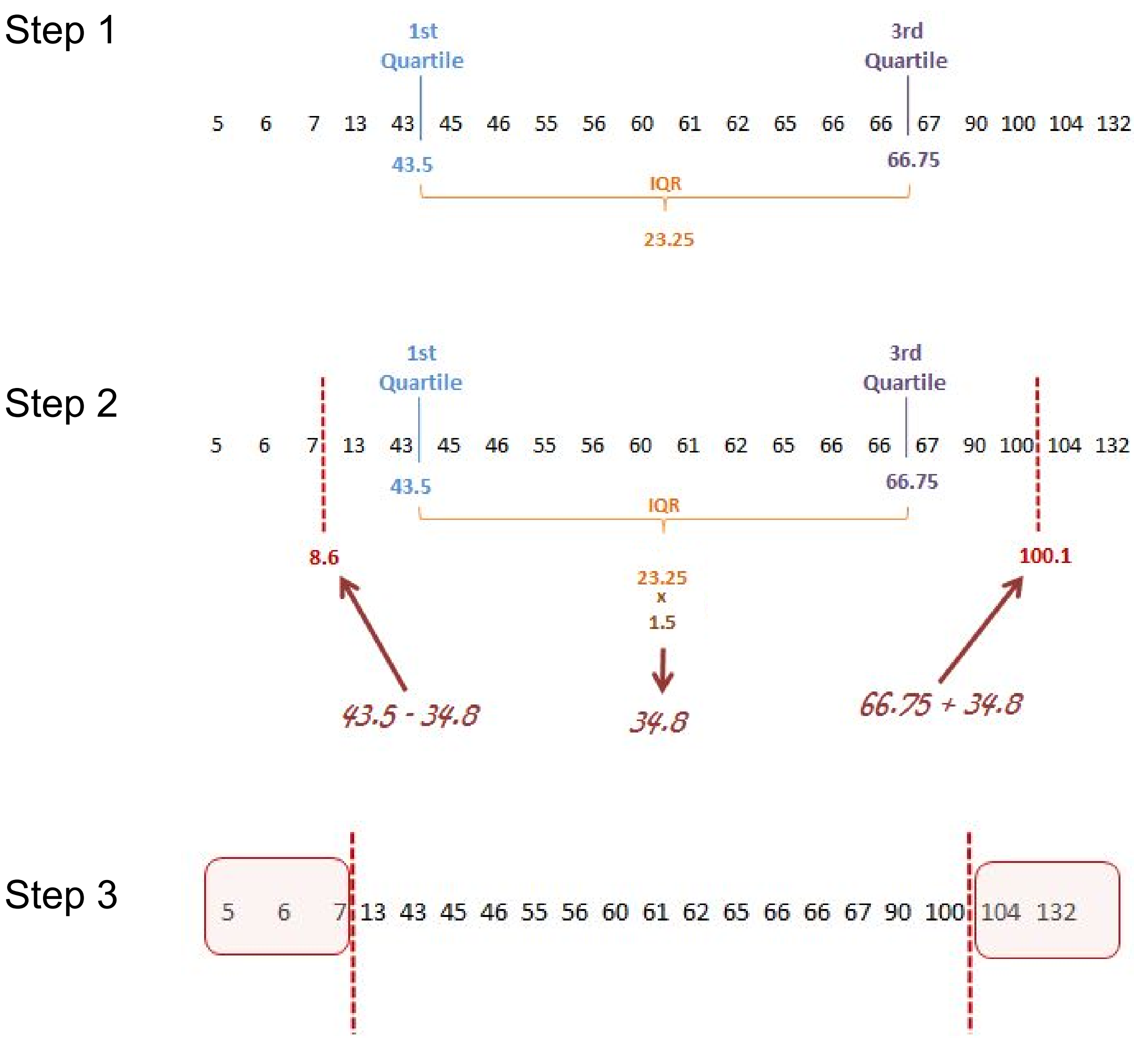
Standardization is especially crucial in order to compare similarities between features based on certain distance measures. The K-means algorithm works with an Euclidean distance measure. This means that when features are on different scales, a feature with values of higher magnitude will effect the clustering result more than the features on smaller scales. Moreover, K-means clustering produce more or less round (rather than elongated) clusters. In this situation leaving variances unequal is equivalent to putting more weight on variables with smaller variance. To assure that all features contribute to the clustering result equally, the data needs to be rescaled. After standarization (or Z-score normalization), the features will be rescaled so that they are centered around 0 with a standard deviation of 1. This assures they have the properties of a normal distribution and influence the clustering result equally [59].

Standarization is also crucial when applying Principal Component Analysis to the dataset. This will be further explained in section 5.

After standarization, there are still data points present, referred to as outliers and seen as noisy observations which do not fit the assumed model that generated the data. They are markedly different from the majority and should be removed in order to make clustering more reliable. Including outliers in the clustering process can skew the results. More specifically, in case of the K-means clustering algorithm, outliers affect the mean of the data points in a cluster. If an outlier will happen to be chosen as an initial seed, then no other point will be assigned to it during the next iterations. This will cause a singleton cluster to be formed (a cluster with only one data point). Moreover, clusters including outliers might have skewed centers and forced to include the outlier, reject other points that could form a tight cluster, without the outliers.

In this thesis Tukey's Method is used for identifying and removing outliers. [24] explains the method as follows. The first step in identifying outliers is to identify the statistical center of the range. To do this, the 1st and 3rd Quartiles are calculated (Step 1, Figure 8). Next, the 3rd Quartile is substracted from the 1st Quartile. This yields an Interquartile Range (IQR). The IQR gives a statistical way of identifying where the main part of the statistical data points (the middle 50%) lies in the range, and how spread out that middle 50% is. Tukey's Method assumes that a data point with a feature beyond 1.5 times the interquartile range (IQR) outside of the IQR is unrepresentative for that feature (Step 2, Figure 8). More specifically, with Tukey's method, outliers are (Step 3, Figure 8):

- values below $(Quartile_1) - (1.5 \cdot IQR)$
- values above $(Quartile_3) + (1.5 \cdot IQR)$



a

Figure 8: Tukey's Method

[a] 24.

In this thesis, for all drivers files and all crossings from the considered crossing group, the crossings/-data points are identified, where any feature lays beyond the calculated outlier step. The data point for this crossing is then removed from the data set. The process is illustrated in Figure 9.

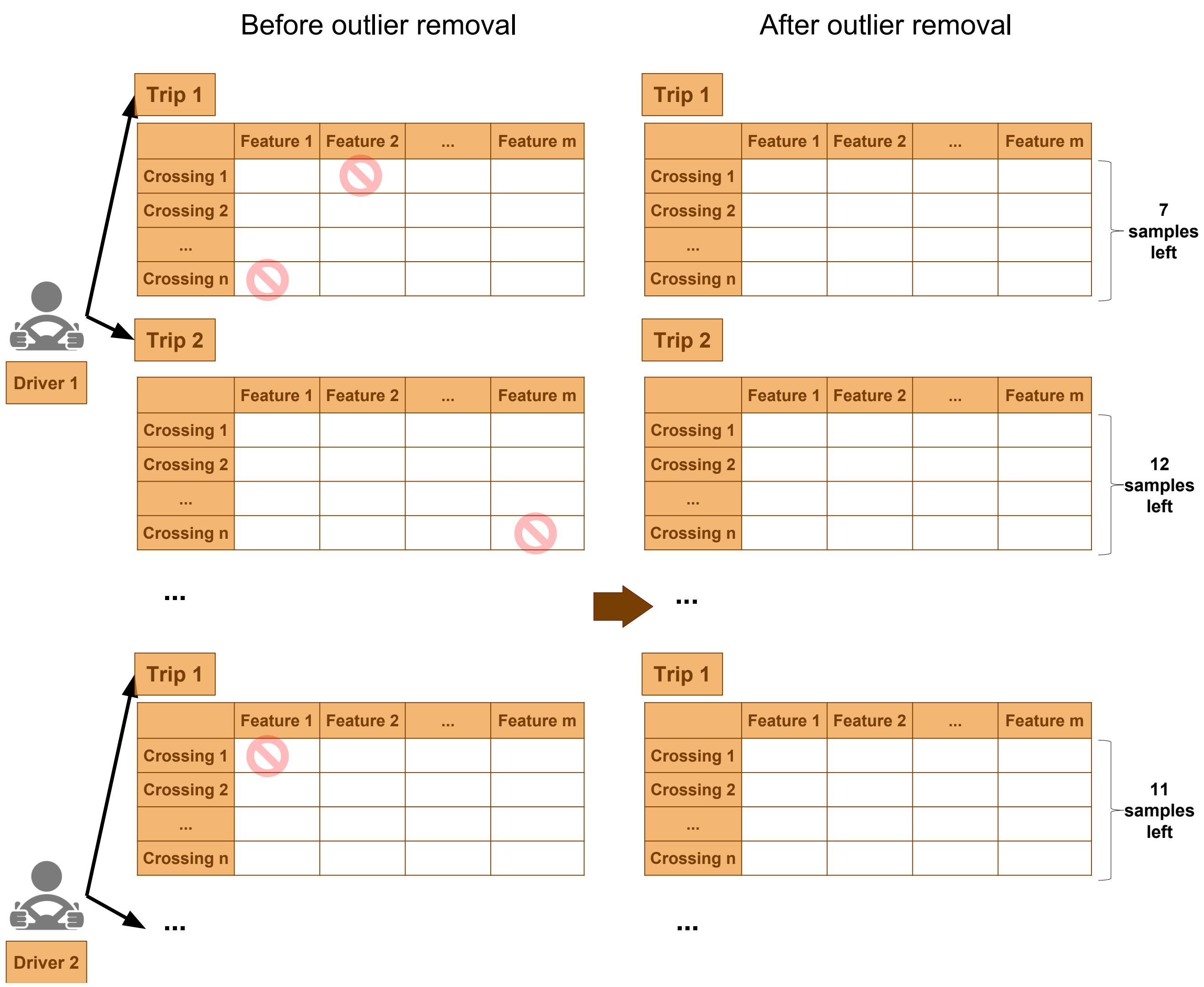


Figure 9: Outlier removal

5 Principal Component Analysis

To reduce the feature space and simplify the graphical representation of clustering results, Principal Component Analysis is carried out on the data. Principal Component Analysis is a way of identifying patterns in data, and expressing the data in such a way as to highlight their similarities and differences. Once these patterns are found, the data is compressed by reducing the number of dimensions, without much loss of information [65]. Specifically speaking,

> Principal Component Analysis (PCA) is a statistical procedure that uses an orthogonal transformation to convert a set of observations of possibly correlated variables into a set of values of linearly uncorrelated variables called principal components. This transformation is defined in such a way that the first principal component has the largest possible variance (that is, accounts for as much of the variability in the data as possible), and each succeeding component in turn has the highest variance under the constraint that it is orthogonal to the preceding components [58].

For PCA to work properly, the data has to standarized so that the mean of the data set is zero. Standarization of features has an effect on the outcome of a PCA. This is because standarization scales the covariance between every pair of variables by the product of the standard deviations of each pair of variables [59].

Now that the data had been scaled to a more normal distribution and has had any necessary outliers removed, PCA can be applied to the dataset to discover the dimensions in which the feature variance is the largest. On top of that, PCA will also provide the explained variance ratio of each dimension, i.e. how much variance within the data is explained by that single dimension. Note that a component (dimension) from PCA can be seen as a new "feature" of the space, however it is a composition of the original features present in the data [5].

6 Clustering

In this thesis, the K-means clustering algorithm is used. The main trait of this algorithm in comparison with other clustering methods, is that the number of clusters is set before clustering occurs. This can turn out to be an advantage or a disadvantage. Setting the number of clusters that should be produced, beforehand prevents the the K-means method from introducing new clusters in case of an anomaly data point. Instead the anomaly data point is sorted to its closest cluster. The main drawback of having fixed the number of clusters in advance, is that it might not be clear how many clusters a dataset might contain. Using an unsuitable k may lead to poor results [17]. In this thesis a number of k values are tried out, ranging from 2 to 5. Considering the research results, the most common cluster number is 2, 3 or 5. Therefore this is also the range of numbers in which good clustering results are expected in this thesis.

6.1 Evaluation metric

Accessing the results of a clustering algorithm implies determining how similar the points in each cluster are and how much the points belonging to different clusters, vary from each other.

A metric used for this purpose is for example the Silhouette Coefficient. It is computed to validate and interpret the consistency within clusters of data. The Silhouette Coefficient is computed for each sample in the data. For its calculation, two parameters are needed; a denoting the intracluster distance for a sample and b, the distance between a sample and the nearest cluster that the sample is not a part of, shortly speaking, nearest-cluster distance. The Silhouett Coefficient for a sample is then equal to:

$$\frac{(b-a)}{\max(a,b)} \tag{2}$$

To validate the clustering result as a whole, the mean Silhouette Coefficient over all samples is determined. This mean is reffered to as the Silhouette Score. The best value is 1 and the worst value is -1. Values near 0 indicate overlapping clusters. Negative values generally indicate that a sample has been assigned to the wrong cluster, as a different cluster is more similar [67]. According to [32], an average silhouette greater than 0.5 suggests reasonable partitioning of data, wheras less than 0.2 indicates that the data do not exhibit cluster structure.

Additionally to this state of the art metric, other metrics are used in this thesis and rely on the characteristics of this specific dataset and application. In order to establish a driver style, the behaviors of a driver must reveal similar characteristics, independent of the traffic situation, alias context. To see if this holds for the behaviors of the given probands, the drivers belonging to a cluster representing a specific behavior profile in one type of traffic context, should also keep the same profile (i.e. belong to a cluster representing a smilar behavior profile) in the second type of traffic situation. It is an indication that the driver's behavior is not a result of coincidence but rather a pattern.

Most importantly, the aspect that every driver completes two trips with the same route is used for evaluating the clustering results. The notion that a driver style classification is found, is reinforced if both trip samples of a driver belong to the same cluster and this also across different traffic situations.

Taking all these aspect into account, the following facets are investigated:

1. Calculate the Silhouette Score for the current clustering trial. If the value exceeds 0.5, carry on with the next steps.
2. Count how many samples belonging to one driver can be found in each cluster.
3. Consider the cluster with the most assignments for this driver. Calculate how much percent of all samples belonging to this driver is assigned to this cluster.
4. Repeat for every driver.
5. If more than 60% of all drivers have a cluster to which more than 75% of their samples are assigned, consider this clustering trial as a possible best result and keep it for fruther analysis.

7 Results

The data analysis was performed on all the crossing groups mentioned in Table 2. This section describes all the steps of the data analysis and presents the outcome of every step.

Starting from step 1, the chosen context is the currently investigated crossing group, and the chosen features are a subset of the potential features presented in section 4.1.3.

Already at the point of extracting crossings, some anomalies are noticed in the data. Three drivers (six driver files, because each driver takes two trips), had to be removed from further investigation. since their The profile of the lateral distance does not clearly indicate at which point each intersection is passed.

The data is then samples to keep only the part corresponding to the currently studied crossing group. The dataset is standardized and outliers removed, according to the scheme described in section 4.2.

To visualize the feature distributions after standardization and outlier removal, a scatter plot matrix is depicted below for all the crossing groups. A scatter plot matrix is defined as follows [66]:

> Given a set of variables $X_1, X_2, ..., X_k$, the scatterplot matrix contains all the pairwise scatter plots of the variables in a matrix format. That is, if there are k variables, the scatterplot matrix will have k rows and k columns and the ith row and jth column of this matrix is a plot of X_i versus X_j.

On the diagonal of the matrix you can see the distributions of the features. Depicting the data in such a way helps to understand the relationship between the attributes.

For example, from the matrices, one can figure out that some of the features are correlated with each other, across all crossing groups. The correlating feature pairs are:

1. $mean\_brake\_pedal$, $mean\_std\_brake\_pedal\_data_1$
2. $mean\_std\_brake\_pedal\_data_2$, $mean\_acc\_pedal\_data$
3. $mean\_std\_acc\_pedal\_data_1$, $mean\_std\_acc\_pedal\_data_2$
4. $mean\_head\_data_1$, $mean\_head\_data_2$
5. $mean\_std\_acc\_pedal\_data_1$, $mean\_std\_gaze\_data_2$
6. $mean\_std\_gaze\_data_1$, $mean\_std\_gaze\_data_1$

The correlations 3, 4 and 6 are straight-forward, since they represent the same aspect. They only differ in the length of sampling time windows.

The first correlation means the more pressure on the brake is applied on average, the higher the standard deviation of the change in pressure. The explanation for this correlation might be the following. A driver who, on average, exerts higher pressure on the brake pedal, probably makes use of a wider range of the possible pressure that can be put on the brake pedal. For this reason the applied pressure also varies more. A driver, who does not brake as severely might keep the pressure on the brake pedal relatively stable.

Correlation 5 implies that the more pressure is applied to the acceleration pedal, the more that gaze direction of the driver varies. This can be easily explained by the fact that before accelerating, the drivers might find it a good idea to look around first to see if it was safe to speed up.

Determining which pairs of features correlate with each other is very important. It suggests that it is sufficient to use one feature from each pair, since the second one will not yield any additional information.

7.1 Clustering

For each crossing group, multiple clustering trials are carried out, with varying number of clusters and varying parameter subset. The chosen number of clusters ranges from 2 and 5 (see section 6) and the used feature subsets included all possible combinations of the features. To decide which of the clustering trials might be meaningful, the introduced metrics had to be computed. Relevant clustering trials have ideally high Silhouette Scores, as well a high value for the second metric presented in section 6.1. What is important however to regard a trial as relevant, is that both of these metric values must lie above the thresholds set and described in section 6.1. More specifically speaking, the Silhouette Score value must be greater than 0.5 and the value for the second metric must be greater than 60%.

7.1.1 Evaluation

The Tables 3 to 9, each present a portion of the clustering results for a specific crossing group. Each table includes the feature subset that was used for a trial, the number of clusters produced and the Silhouette Score. The last column concerns the drivers, who have more than 75% of their samples belonging to one cluster. The values in the last column correspond to these drivers. Each of them is represented by the proportion of samples belonging to the most assigned cluster. Note that the number of drivers can vary due to outlier removal. Trials that stand out because of (a) high metric value(s) are highlighted in green in each table.

ID	Sil. Score	Features	No. of Clusters	2nd Metric Score	Driver Samples Proportion
0	0.66	$mean\_vel$, $mean\_brake\_pedal$	2	70.00	0.77, 0.83, 0.89, 0.80, 0.76, 1.0, 0.8125, 1.0, 0.98, 0.79, 0.925, 0.85, 0.90, 0.83, 0.86, 1.0, 0.91, 0.77, 0.79, 0.76, 0.77
1	0.66	$mean\_vel$, $mean\_std\_brake\_pedal\_data_1$	2	70.00	0.77, 0.83, 0.86, 0.80, 0.76, 1.0, 0.8125, 1.0, 0.98, 0.79, 0.925, 0.85, 0.91, 0.83, 0.87, 1.0, 0.91, 0.77, 0.79, 0.76, 0.77
2	0.62	$mean\_brake\_pedal$, $mean\_std\_brake\_pedal\_data_1$	3	70.00	0.90, 1.0, 0.95, 0.78, 0.76, 0.87, 0.8125, 0.77, 0.83, 0.76, 0.97, 1.0, 0.87, 0.78, 0.88, 0.80, 0.80, 0.92, 0.98, 0.89, 0.89
3	0.72	$mean\_brake\_pedal$, $mean\_std\_brake\_pedal\_data_1$	2	86.67	1.0, 0.92, 0.96, 1.0, 0.98, 0.85, 0.97, 0.96, 1.0, 0.94, 0.95, 0.97, 1.0, 0.93, 1.0, 1.0, 0.93, 0.95, 0.98, 1.0, 0.95, 1.0, 0.80, 0.92, 0.98, 0.79

Table 3: Clustering Results Table for Crossing Group 1

The feature subset yielding the highest metric values for the crossing group 1, which contains all the crossings, is the mean pressure put on the brake pedal and its mean standard deviation. This shows that braking behavior is the best indicator for driver behavior when all crossings are taken into account. The produced number of clusters is two for every trial. Figure 10 presents the scatter plot of the clusters produced for the crossing group 1, when this feature subset is used. All the samples are annotated with a number corresponding to the driver they belong to. The driver indices range from 0 to 29, since 30 drivers are examined after removing three of them in the outlier removal phase.

For the crossing group 2, there are more interesting variants. The first four stand out because of their Silhouette Score and include the combinations:

- $mean\_vel$, $mean\_brake\_pedal$,
- $mean\_vel$, $mean\_std\_brake\_pedal\_data_1$,
- $mean\_brake\_pedal$, $mean\_std\_gaze\_data_2$
- $mean\_std\_brake\_pedal\_data_1$, $mean\_std\_gaze\_data_2$

The features $mean\_brake\_pedal$ and $mean\_std\_brake\_pedal\_data_1$ are correlated with each other, hence combinations which differ only in these two features perform similarly well. What could be gathered from these results is that the velocity and braking behavior seem to produce the best structured clusters. However, after adding some more features, even though the Silhouette Score decreases, the value for the second metric is slightly higher. These combinations include:

- $mean\_vel$, $mean\_std\_brake\_pedal\_data_1$, $mean\_std\_acc\_pedal\_data_1$, $mean\_head\_data_1$, $mean\_std\_gaze\_data_2$
- $mean\_vel$, $mean\_brake\_pedal$, $mean\_std\_acc\_pedal\_data_1$, $mean\_head\_data_2$, $mean\_std\_gaze\_data_2$
- $mean\_vel$, $mean\_std\_brake\_pedal\_data_1$, $mean\_std\_acc\_pedal\_data_1$, $mean\_head\_data_1$, $mean\_std\_gaze\_data_2$

In this case, again, because of feature correlation, it is logical that the combinations will perform similarly well. What can be observed, however, is that adding information about the change of pressure on the acceleration pedal and information about the change of head and gaze direction, slightly improves the clustering result according to the second evaluation metric. In other words, adding these features results in samples belonging to one driver to be more often clustered together.

ID	Sil. Score	Features	No. of Clusters	2nd Metric Score	Driver Samples Proportion
0	0.72	$mean\_vel$, $mean\_brake\_pedal$	2	68.97	1.0, 1.0, 1.0, 1.0, 1.0, 1.0, 1.0, 1.0, 0.88, 0.8, 1.0, 0.89, 0.88, 0.88, 1.0, 1.0, 0.9, 1.0, 1.0, 1.0
1	0.72	$mean\_vel$, $mean\_std\_brake\_pedal\_data_1$	2	68.97	1.0, 1.0, 1.0, 1.0, 1.0, 1.0, 1.0, 1.0, 0.88, 0.8, 1.0, 0.89, 0.88, 0.88, 1.0, 1.0, 0.9, 1.0, 1.0, 1.0
2	0.72	$mean\_vel$, $mean\_brake\_pedal$, $mean\_std\_brake\_pedal\_data_1$	2	68.97	1.0, 1.0, 1.0, 1.0, 1.0, 1.0, 1.0, 1.0, 0.88, 0.8, 1.0, 0.89, 0.88, 0.88, 1.0, 1.0, 0.9, 1.0, 1.0, 1.0
3	0.69	$mean\_brake\_pedal$, $mean\_std\_gaze\_data_2$	2	76.67	1.0, 1.0, 0.78, 1.0, 0.89, 1.0, 1.0, 1.0, 1.0, 1.0, 0.88, 0.83, 1.0, 0.88, 1.0, 1.0, 1.0, 1.0, 0.9, 1.0, 1.0, 0.9, 0.78
4	0.69	$mean\_std\_brake\_pedal\_data_1$, $mean\_std\_gaze\_data_2$	2	76.67	1.0, 1.0, 0.78, 1.0, 0.89, 1.0, 1.0, 1.0, 1.0, 1.0, 0.88, 0.83, 1.0, 0.88, 1.0, 1.0, 1.0, 1.0, 0.9, 1.0, 1.0, 0.9, 0.78
5	0.69	$mean\_brake\_pedal$, $mean\_std\_brake\_pedal\_data_1$, $mean\_std\_gaze\_data_2$	2	76.67	1.0, 1.0, 0.78, 1.0, 0.89, 1.0, 1.0, 1.0, 1.0, 1.0, 0.88, 0.83, 1.0, 0.88, 1.0, 1.0, 1.0, 1.0, 0.9, 1.0, 1.0, 0.9, 0.78
6	0.67	$mean\_brake\_pedal$, $mean\_std\_acc\_pedal\_data_1$	2	66.67	0.89, 0.78, 1.0, 0.78, 0.78, 1.0, 1.0, 0.89, 0.9, 1.0, 1.0, 1.0, 0.83, 1.0, 1.0, 1.0, 1.0, 0.9, 1.0, 1.0
...	...	...	...	...	...
41	0.61	$mean\_vel$, $mean\_std\_brake\_pedal\_data_1$, $mean\_std\_acc\_pedal\_data_2$, $mean\_head\_data_1$, $mean\_std\_gaze\_data_2$	2	75.86	1.0, 0.78, 1.0, 1.0, 1.0, 1.0, 1.0, 0.8, 1.0, 1.0, 1.0, 0.83, 1.0, 0.88, 0.83, 1.0, 1.0, 1.0, 0.9, 1.0, 1.0, 0.9
42	0.61	$mean\_vel$, $mean\_std\_brake\_pedal\_data_1$, $mean\_std\_acc\_pedal\_data_1$, $mean\_head\_data_1$, $mean\_std\_gaze\_data_2$	2	79.31	1.0, 0.78, 1.0, 0.78, 1.0, 1.0, 1.0, 1.0, 0.8, 1.0, 1.0, 1.0, 0.83, 1.0, 0.88, 0.83, 1.0, 1.0, 1.0, 0.9, 1.0, 1.0, 0.9
43	0.60	$mean\_vel$, $mean\_brake\_pedal$, $mean\_std\_acc\_pedal\_data_2$, $mean\_head\_data_2$, $mean\_std\_gaze\_data_2$	2	75.86	1.0, 0.78, 1.0, 1.0, 1.0, 1.0, 1.0, 0.8, 1.0, 1.0, 1.0, 0.83, 1.0, 0.88, 0.86, 1.0, 1.0, 1.0, 0.9, 1.0, 1.0, 0.9

Table 4: Clustering Results Table for Crossing Group 2

Because some of the feature combinations correlate with each other, only the clusters for one variant of each correlating subset are plotted, namely $mean\_vel$, $mean\_brake\_pedal$, $mean\_brake\_pedal$, $mean\_std\_gaze\_data_2$ and $mean\_vel$, $mean\_std\_brake\_pedal\_data_1$, $mean\_std\_acc\_pedal\_data_1$, $mean\_head\_data_1$, $mean\_std\_gaze\_data_2$. The produced number of clusters for every trial, is again, two.

ID	Sil. Score	Features	No. of Clusters	2nd Metric Score	Driver Samples Proportion
44	0.60	$mean\_vel$, $mean\_std\_brake\_pedal\_data_1$, $mean\_std\_acc\_pedal\_data_2$, $mean\_head\_data_2$, $mean\_std\_gaze\_data_2$	2	75.86	1.0, 0.78, 1.0, 1.0, 1.0, 1.0, 1.0, 0.8, 1.0, 1.0, 1.0, 0.83, 1.0, 0.88, 0.86, 1.0, 1.0, 1.0, 0.9, 1.0, 1.0, 0.9
45	0.60	$mean\_vel$, $mean\_std\_acc\_pedal\_data_2$, $mean\_head\_data_1$, $mean\_std\_gaze\_data_2$	2	68.97	1.0, 1.0, 0.78, 1.0, 1.0, 1.0, 1.0, 0.8, 1.0, 1.0, 1.0, 0.8, 1.0, 0.8, 1.0, 1.0, 0.9, 1.0, 1.0, 0.9
46	0.60	$mean\_vel$, $mean\_brake\_pedal$, $mean\_std\_acc\_pedal\_data_1$, $mean\_head\_data_2$, $mean\_std\_gaze\_data_2$	2	79.31	1.0, 0.78, 1.0, 0.78, 1.0, 1.0, 1.0, 1.0, 0.8, 1.0, 1.0, 1.0, 0.83, 1.0, 0.88, 0.86, 1.0, 1.0, 1.0, 0.9, 1.0, 1.0, 0.9
47	0.60	$mean\_vel$, $mean\_std\_brake\_pedal\_data_1$, $mean\_std\_acc\_pedal\_data_1$, $mean\_head\_data_2$, $mean\_std\_gaze\_data_2$	2	79.31	1.0, 0.78, 1.0, 0.78, 1.0, 1.0, 1.0, 1.0, 0.8, 1.0, 1.0, 1.0, 0.83, 1.0, 0.88, 0.86, 1.0, 1.0, 1.0, 0.9, 1.0, 1.0, 0.9]

Table 5: Clustering Results Table for Crossing Group 2

The crossing group 3, seems to be best described by the mean velocity and braking behavior, like in the case of crossing group 2. Adding the change in head direction also increases the value of the second metric, just like the addition of change in gaze and head direction increased the value for crossing group 2.

ID	Sil. Score	Features	No. of Clusters	2nd Metric Score	Driver Samples Proportion
0	0.67	$mean\_vel$, $mean\_brake\_pedal$	2	76.67	1.0, 1.0, 1.0, 1.0, 1.0, 1.0, 1.0, 1.0, 1.0, 1.0, 1.0, 1.0, 1.0, 1.0, 1.0, 1.0, 1.0, 1.0, 1.0, 1.0, 1.0, 1.0, 1.0
1	0.67	$mean\_vel$, $mean\_std\_brake\_pedal\_data_1$	2	76.67	1.0, 1.0, 1.0, 1.0, 1.0, 1.0, 1.0, 1.0, 1.0, 1.0, 1.0, 1.0, 1.0, 1.0, 1.0, 1.0, 1.0, 1.0, 1.0, 1.0, 1.0, 1.0, 1.0
2	0.67	$mean\_vel$, $mean\_brake\_pedal$, $mean\_std\_brake\_pedal\_data_1$	2	76.67	1.0, 1.0, 1.0, 1.0, 1.0, 1.0, 1.0, 1.0, 1.0, 1.0, 1.0, 1.0, 1.0, 1.0, 1.0, 1.0, 1.0, 1.0, 1.0, 1.0, 1.0, 1.0, 1.0

Table 6: Clustering Results Table for Crossing Group 3

ID	Sil. Score	Features	No. of Clusters	2nd Metric Score	Driver Samples Proportion
3	0.65	$mean\_brake\_pedal$, $mean\_std\_brake\_pedal\_data_1$	2	73.33	1.0, 1.0, 1.0, 1.0, 1.0, 1.0, 1.0, 1.0, 1.0, 1.0, 1.0, 1.0, 1.0, 1.0, 1.0, 1.0, 1.0, 1.0, 1.0, 1.0, 1.0, 1.0
4	0.59	$mean\_brake\_pedal$, $mean\_std\_gaze\_data_1$	2	66.67	1.0, 1.0, 1.0, 1.0, 1.0, 1.0, 1.0, 1.0, 1.0, 1.0, 1.0, 1.0, 1.0, 1.0, 1.0, 1.0, 1.0, 1.0, 1.0, 1.0
...	...	...	...	...	...
18	0.51	$mean\_vel$, $mean\_std\_brake\_pedal\_data_1$, $mean\_head\_data_1$	2	79.31	1.0, 1.0, 1.0, 1.0, 1.0, 1.0, 1.0, 1.0, 1.0, 1.0, 1.0, 1.0, 1.0, 1.0, 1.0, 1.0, 1.0, 1.0, 1.0, 1.0, 1.0, 1.0, 1.0
19	0.51	$mean\_vel$, $mean\_brake\_pedal$, $mean\_head\_data_1$	2	79.31	1.0, 1.0, 1.0, 1.0, 1.0, 1.0, 1.0, 1.0, 1.0, 1.0, 1.0, 1.0, 1.0, 1.0, 1.0, 1.0, 1.0, 1.0, 1.0, 1.0, 1.0, 1.0, 1.0

Table 7: Clustering Results Table for Crossing Group 3

Again, since $mean\_brake\_pedal$ and $mean\_std\_brake\_pedal\_data_1$ correlate, only the clusters for the combination with $mean\_brake\_pedal$ are plotted from the two possibilities; $mean\_vel$, $mean\_brake\_pedal$ and $mean\_vel$, $mean\_std\_brake\_pedal\_data_1$.

In case of the crossing group 4, the braking behavior produces the best structured clusters. However, adding velocity to the braking information, increases the number of drivers whose samples are often (in more than 75% cases), clustered together.

ID	Sil. Score	Features	No. of Clusters	2nd Metric Score	Driver Samples Proportion
0	0.74	$mean\_brake\_pedal$, $mean\_std\_brake\_pedal\_data_1$	2	83.33	1.0, 0.96, 0.96, 1.0, 1.0, 1.0, 0.92, 0.95, 0.96, 1.0, 0.92, 1.0, 1.0, 1.0, 1.0, 0.87, 1.0, 0.96, 1.0, 0.78, 0.95, 1.0, 0.94, 0.92, 0.86
1	0.65	$mean\_vel$, $mean\_std\_brake\_pedal\_data_1$	2	70.00	1.0, 1.0, 0.91, 0.96, 1.0, 0.91, 1.0, 0.88, 0.96, 0.77, 0.91, 1.0, 0.88, 0.91, 0.84, 0.96, 1.0, 0.88, 0.88, 0.89, 0.88
2	0.65	$mean\_vel$, $mean\_brake\_pedal$	2	70.00	1.0, 1.0, 0.91, 0.96, 1.0, 0.91, 1.0, 0.88, 0.96, 0.77, 0.91, 1.0, 0.83, 0.91, 0.85, 0.96, 1.0, 0.88, 0.88, 0.89, 0.88

Table 8: Clustering Results Table for Crossing Group 4

Equivalently, only the clusters for one of the combinations $mean\_vel$, $mean\_std\_brake\_pedal\_data_1$ and $mean\_vel$, $mean\_brake\_pedal$ are plotted. Just like in all previous cases, the number of clusters produced for all trials is two.

ID	Sil. Score	Features	No. of Clusters	2nd Metric Score	Driver Samples Proportion
0	0.83	$mean\_brake\_pedal$, $mean\_std\_brake\_pedal\_data_1$	2	83.33	1.0, 0.80, 1.0, 1.0, 1.0, 1.0, 1.0, 1.0, 1.0, 1.0, 0.89, 1.0, 1.0, 1.0, 0.90, 0.80, 1.0, 0.89, 0.78, 1.0, 1.0, 0.90, 1.0, 0.90, 0.90

Table 9: Clustering Results Table for Crossing Group 5

The crossing group 5 is best clustered with the feature combination $mean\_brake\_pedal$, $mean\_std\_brake\_pedal\_data_1$. No additional features boost this performance.

8 Interpretation

To better study the results, for each crossing group a plot is created, depicting the characteristics of the produced clusters. For each relevant feature combination and the clusters created for it, a plot of two histograms is visualized. Since in every case, only two clusters are produced, the cluster labels are always 0 and 1. The first histogram in each plot visualizes the distribution of the feature for drivers belonging to the cluster with index 0, the second histogram, analogously, the cluster with index 1.

The first cluster result to interpret belongs to the crossing group 1 and feature combination $mean\_brake\_pedal$, $mean\_std\_brake\_pedal\_data_1$ Figure 24. From the plot for feature $mean\_brake\_pedal$, it is clear that cluster 0 represents drivers who apply less pressure on the braking pedal and the range of the pressure is not as wide as in case of drivers belonging to the cluster with index 1. The second feature of the combination, $mean\_std\_brake\_pedal\_data_1$ correlates with $mean\_brake\_pedal$, so it is clear that the distributions will have a similar histogram profile. Drivers assigned to cluster 0, have a lower deviation in the pressure they apply to the brake pedal.

According to the notions presented in section 2.6, drivers with stable braking behavior are regarded as calm drivers, whereas frequent changes in brake operation and intensive brake operation indicate an aggressive or dynamic driver. Thus, cluster 0 will correspond in this case to the cluster of calm drivers.

When it comes to the crossing group 2, three relevant clustering trials are observed. The histograms for the feature combination $mean\_brake\_pedal$, $mean\_std\_gaze\_data_2$ are depicted in Figure 25. One can see, that even the histograms of applied pressure on the brake pedal almost overlap with each other, the one for cluster 0 being set slightly more to the left. There is however a substantial difference in the velocity profiles. Drivers belonging to cluster 0 tend to driver slower than drivers assigned to cluster 1. Thus for this particular feature combination, again cluster 0 is regarded as the one including calmer drivers and 1 the dynamic ones.

Next in turn for crossing group 2, is the feature combination $mean\_brake\_pedal$, $mean\_std\_gaze\_data_2$. The histograms for the mean brake pedal pressure almost overlap again for both clusters, but drivers from cluster 0 tend to change their gaze direction far less abruptly than drivers in cluster 1 Figure 26. Invoking the findings described in section section 2.6 , cluster 0 is here again seen as the one of the calm drivers.

Finally, the histograms for the combination $mean\_vel$, $mean\_std\_brake\_pedal\_data_1$, $mean\_std\_acc\_pedal\_data_1$, $mean\_head\_data_1$, $mean\_std\_gaze\_data_2$ indicate that this time cluster 1 is the one in which drivers tend to reach lower velocity, less deviation in the pressure applied to the acceleration pedal and less deviation in gaze direction 27. The histograms for braking and head direction deviation overlap, indicating the behavior seems similar in both clusters. The braking deviation values are set only slightly more left for cluster 1. In this case, therefore, cluster 1 is the one assigned to calmer drivers, whereas cluster 0 includes the more dynamic ones 2.6.

Crossing group 3 with feature combination $mean\_vel$, $mean\_brake\_pedal$ cluster 1 manifests lower velocity values and its histogram for the mean brake pedal pressure is set slightly more to the left, than the histogram for cluster 0. This indicates that that less pressure is applied by the drivers in cluster 1. Thus, analogously, 0 represents the cluster for dynamic drivers, and 1 for the calm ones.

The feature combination $mean\_vel$, $mean\_brake\_pedal$, $mean\_head\_data_1$ produces the cluster 0, with drivers putting slightly less pressure on the brake pedal and driving with lower velocity. The head direction changes are similar for both clusters, probably to the fact that unexpected traffic comes from the right in this crossing situation. Cluster 0 is therefore regarded as the cluster with calm drivers.

Crossing group 4, analogously assigns the drivers with traits attributed to calm driving 2.6, to cluster 0. These traits are lower velocity values and the braking histogram representing a smaller value range and this on the lower value side, too.

Finally, for crossing group 5, cluster 0 assumes more brake pressure, more deviation in the pressure, and wider range of the values, indicating more unstable braking than for drivers in cluster 1. Hence, in this case, cluster 1 represents the calmer drivers.

To visualize whether the drivers were assigned to clusters with similar profiles (i.e. the 'calm' or 'dynamic' cluster), across different crossing situations, Table 10 is created.

The columns of the table represent indices of the drivers, ranging from 0 to 29, since 30 drivers are considered. The rows represent the different relevant clustering cases in terms of the feature combination and crossing group.

For each driver and each clustering trial, it is checked which cluster label is assigned to the driver. The cluster labels can be 0 or 1. The two clusters represent the calm and dynamic drivers. Keep in mind that the label can represent a different type of driver, depending on the the particular clustering trial.

If the label corresponds to a cluster of calm drivers for the particular trial, the cell is highlighted in green. Analogously, if the label corresponds to the cluster with dynamic drivers, the cell is highlighted in pink.

Cells with no label indicate that for this particular trial the driver had less than 75% of the samples assigned to one cluster and is therefore not taken into account.

What can be noticed is that drivers with the majority of "calm" assignments across the different crossings, in fact belong to the "calm" cluster for all the intersection situations. It is also quiet prominent that drivers assigned to the "dynamic" cluster in some situations, do not keep the assignment consistently. They also happen to have undefined assignments in many cases, meaning that less than 75% (exclusively) of the driver's samples were assigned to one cluster. Another apparent observation is that in case of

dynamic driver
calm driver

Crossing Group	Features	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29
1	$mean\_brake\_pedal$, $mean\_std\_brake\_pedal\_data_1$	0	0	-	0	0	0	0	0	0	0	-	0	0	0	0	0	0	0	0	0	0	-	0	0	0	0	-	0	0	0
2	$mean\_vel$, $mean\_brake\_pedal$	1	0	-	-	-	0	0	0	-	0	0	-	-	0	0	0	0	1	0	-	-	1	-	0	0	0	-	0	0	0
2	$mean\_brake\_pedal$, $mean\_std\_gaze\_data_2$	1	0	1	0	-	0	0	0	0	0	0	-	0	0	0	0	-	1	0	-	0	-	1	0	0	0	-	0	0	-
2	$mean\_vel$, $mean\_std\_acc\_pedal\_data_1$, $mean\_std\_brake\_pedal\_data_1$, $mean\_head\_data_1$, $mean\_std\_gaze\_data_2$	0	1	-	1	-	1	1	1	-	1	1	-	1	1	1	1	-	0	1	-	1	0	0	1	1	1	-	1	1	1
3	$mean\_vel$, $mean\_brake\_pedal$	0	-	0	1	1	1	1	-	1	1	1	0	1	1	1	1	1	1	-	1	-	-	-	1	1	-	0	1	1	1
3	$mean\_vel$, $mean\_brake\_pedal$, $mean\_head\_data_1$	0	-	-	0	0	0	0	-	0	0	0	1	0	0	0	0	0	0	-	0	-	-	0	0	0	-	1	0	0	0
4	$mean\_vel$, $mean\_brake\_pedal$	1	0	1	0	-	-	0	0	-	1	-	1	0	0	0	0	0	-	-	1	-	1	0	0	-	0	-	0	0	0
5	$mean\_brake\_pedal$, $mean\_std\_brake\_pedal\_data_1$	-	1	-	1	1	-	1	1	1	1	1	1	1	1	1	0	1	1	1	1	1	-	1	1	1	1	-	1	1	1

Table 10: Clustering results for all drivers and relevant clustering trials

the crossing group 5, with traffic coming from the right side, the drivers tend to be classified as calm, independent of the assignments for other situations. The same unified assignments are found for the crossing group 1, which includes all the encountered intersections. The case of crossing group 5 may be one of the situations where driver behavior depends mostly on the driving context. According to the assumptions made in section 2.4, such behaviors should be disregarded when investigating driver style. The fact that taking into account all crossings does not lead to differentiation between drivers might also be explained by the fact that driving behavior strongly depends on the context and each situation (in this case intersection type) must be viewed separately (section 2.4). If these crossing groups are removed, the drivers, who display dynamic behavior in some of the crossing groups, will show it in the majority of crossing situations, without the consideration of unknown. This indicates that if the threshold for having a cluster assigned was lowered (to less than 75%), the assignments would turn out to be of dynamic character, as well. In this case one might speak of a driving style classification with two styles; calm and dynamic.

9 Conclusion and Future Work

The thesis is devoted to the subject of clustering driver simulation data according to the drivers' behaviour on intersections. The K-means clustering algorithm used for this purpose is one of the most prominent clustering algorithms in the driver style research field. Assuming that driving behaviour is dependent on the traffic context, the data were clustered separately for different types of intersection and for all the intersections, for comparison. Two clusters were found that corresponds to the number of categories used for driver style classification described in literature. The drivers in each cluster manifest behaviour patterns that can be regarded as "aggressive" (dynamic) and "calm" driving style. Thus, the clusters are named "dynamic" and "calm". Dynamic drivers exhibited more varying usage of the brake and acceleration pedal, higher velocity and less variety in gaze direction. The drivers were also almost uniquely clustered in case of the crossing group where the drivers had to let traffic from the right pass first and there was actually traffic coming from the right. All of them had to slow down before the intersection to let the traffic through and they were expecting to do so because they had no right of way. Therefore, the velocity was relatively low and the braking not as severe because the drivers expecting it. All the crossing types did not provide a precise differentiation of driving behaviour. However, there is an indication that two styles could be established if certain conditions were met. The first condition is that the uninformative crossing situation with traffic coming from the right is discarded. The second, that a milder criteria is set for determining the specific cluster affiliation in a crossing situation. In the thesis a relatively strict constraint was assumed. The clustering results that drivers are best distinguished according to their braking behaviour stands in agreement with the findings reported in literature. Moreover, cognitive features such as gaze and head direction did boost the performance of the clustering. To determine more precisely if one can speak of driver style should be continued. The common traits of drivers at a specific type of intersection comes out as another topic for further investigation.

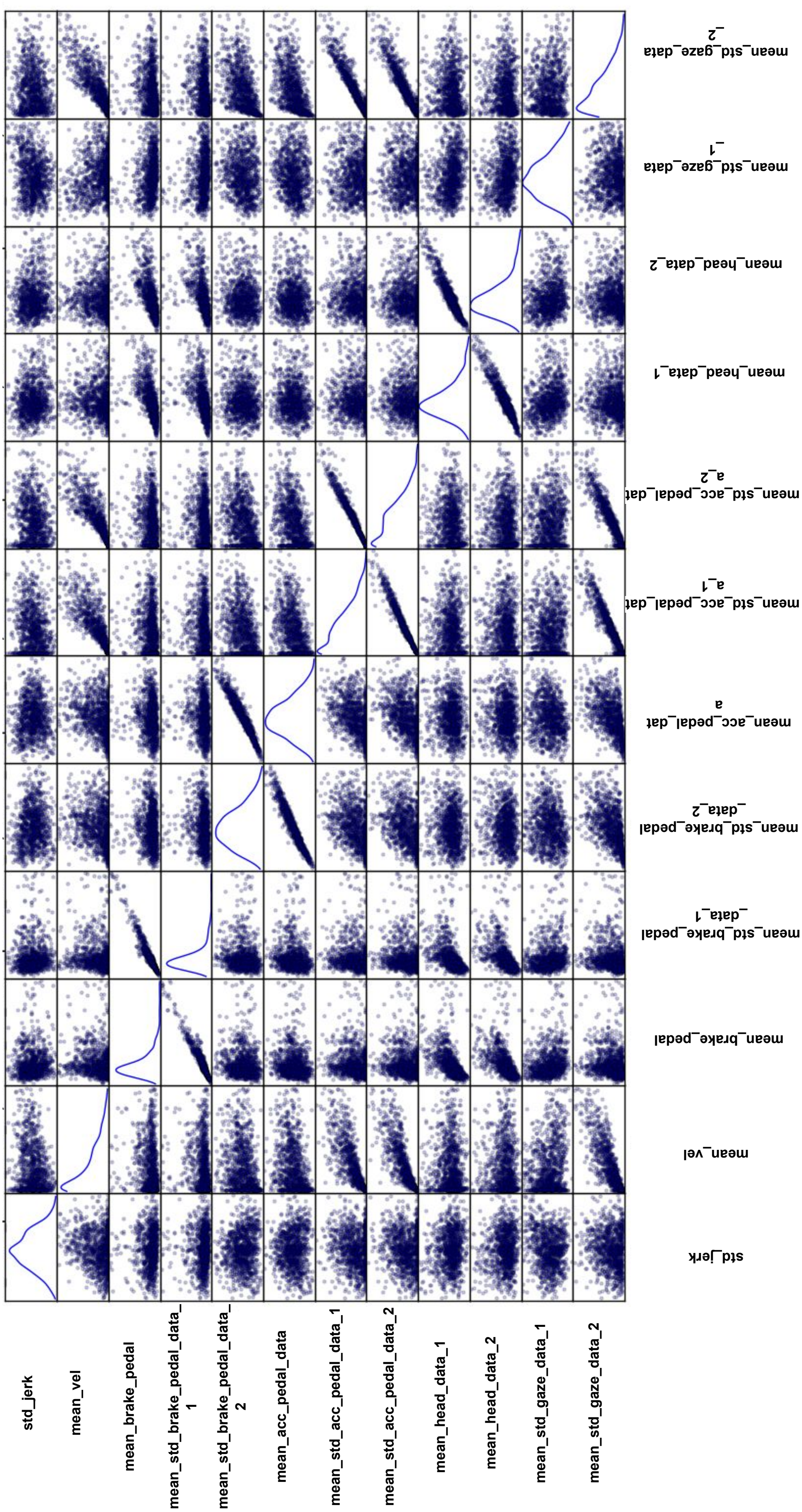


Figure 10: Scatter plot matrix for crossing group 1

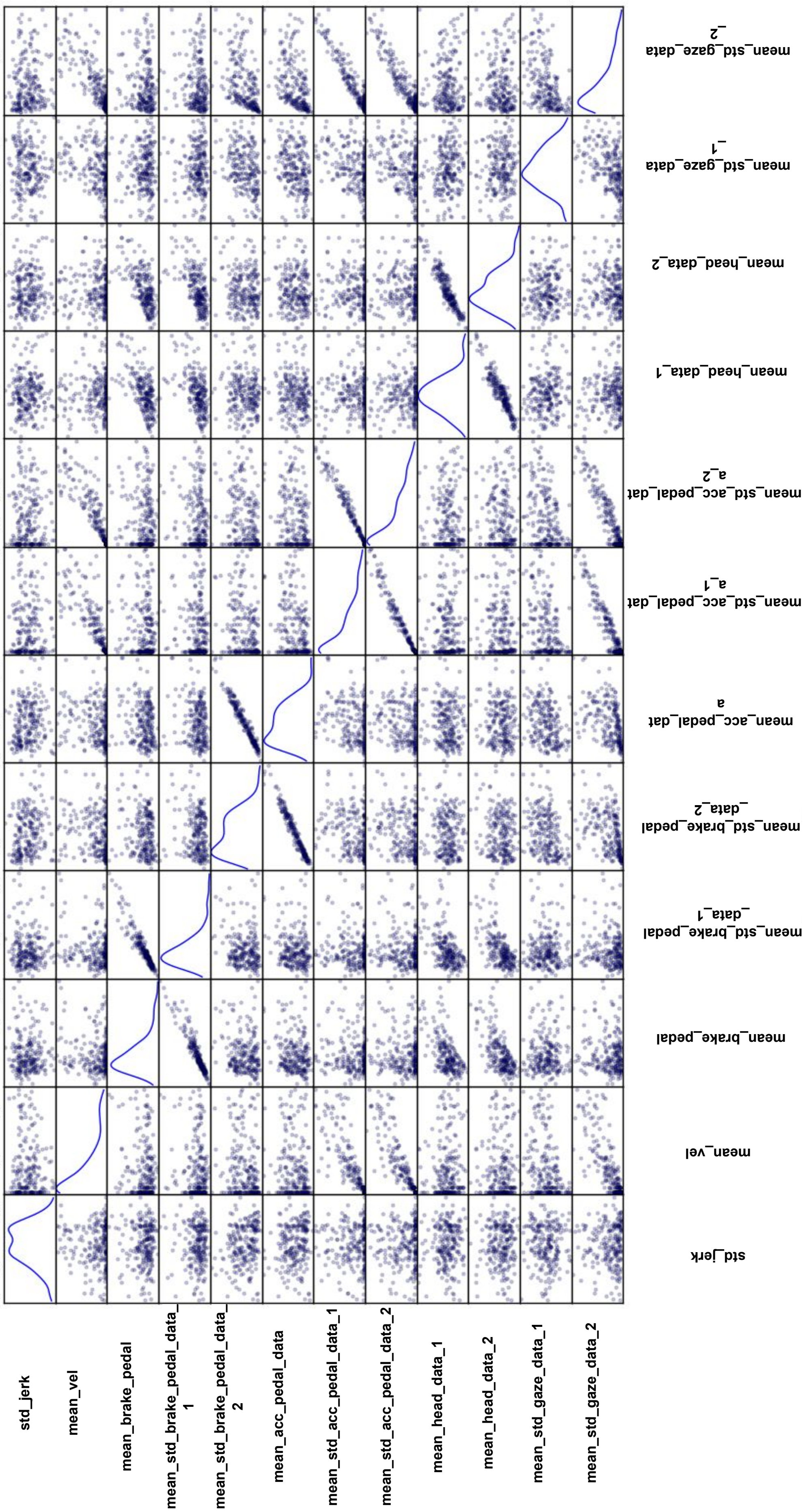


Figure 11: Scatter plot matrix for crossing group 2

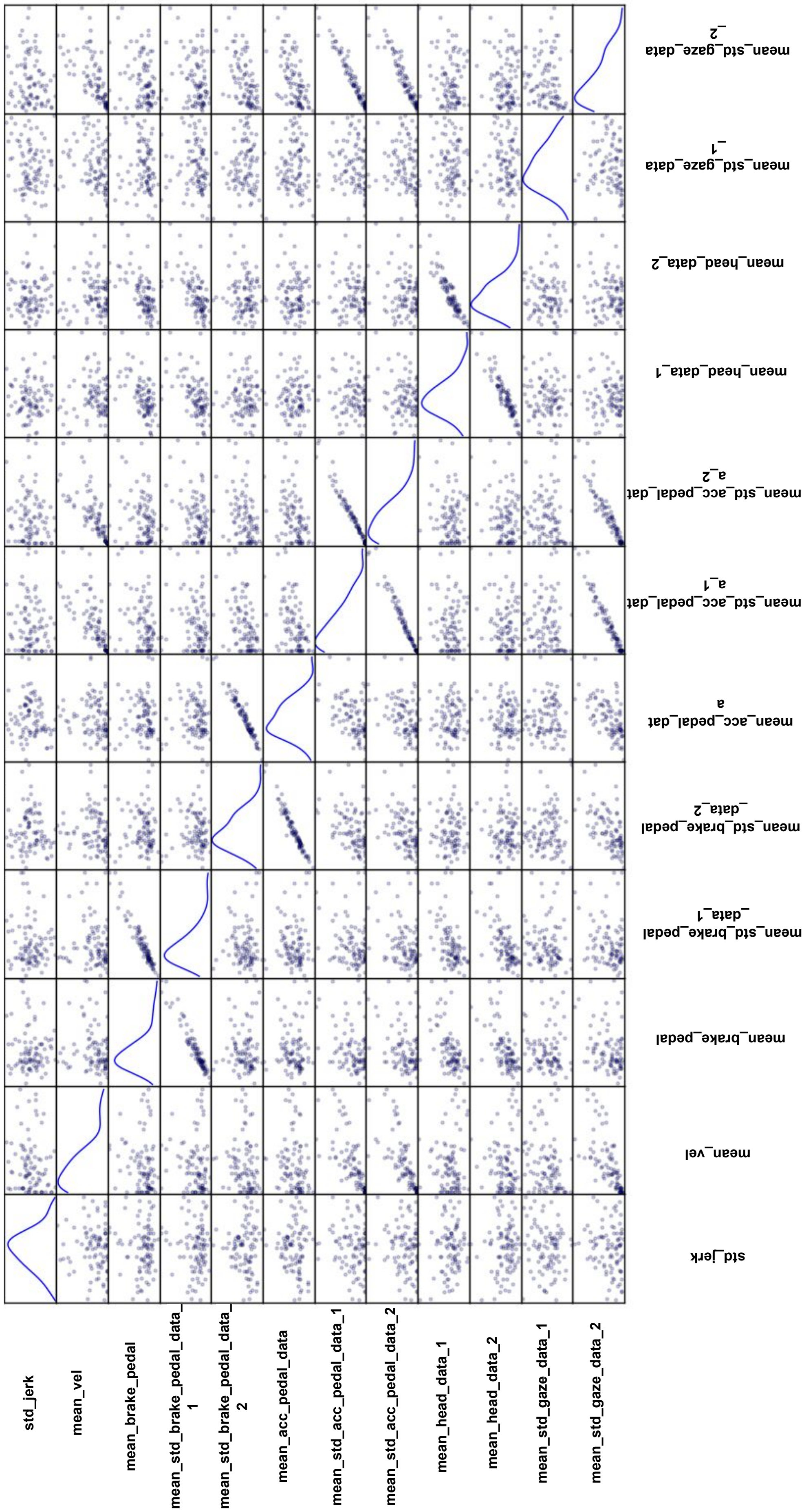


Figure 12: Scatter plot matrix for crossing group 3

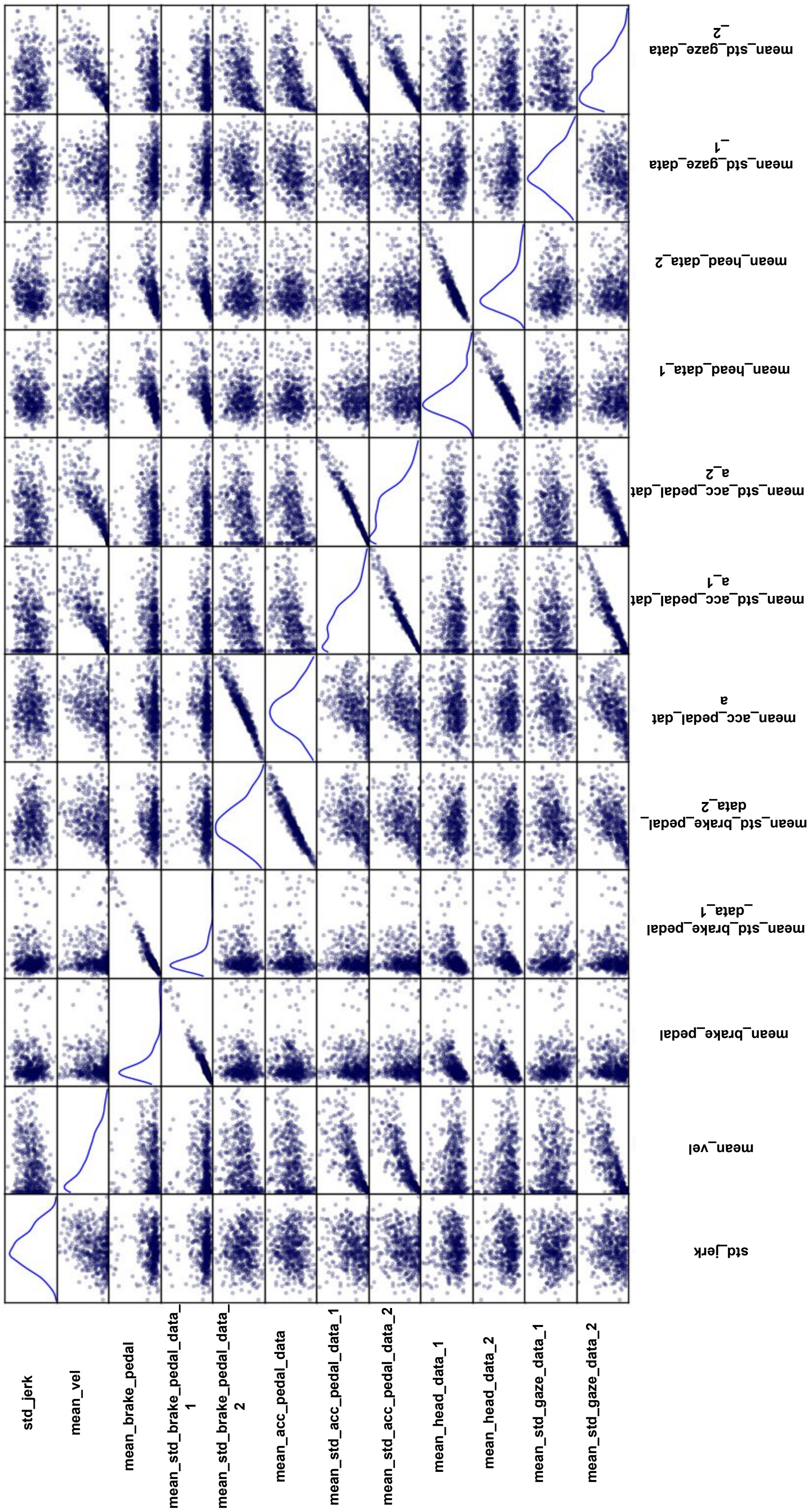


Figure 13: Scatter plot matrix for crossing group 4

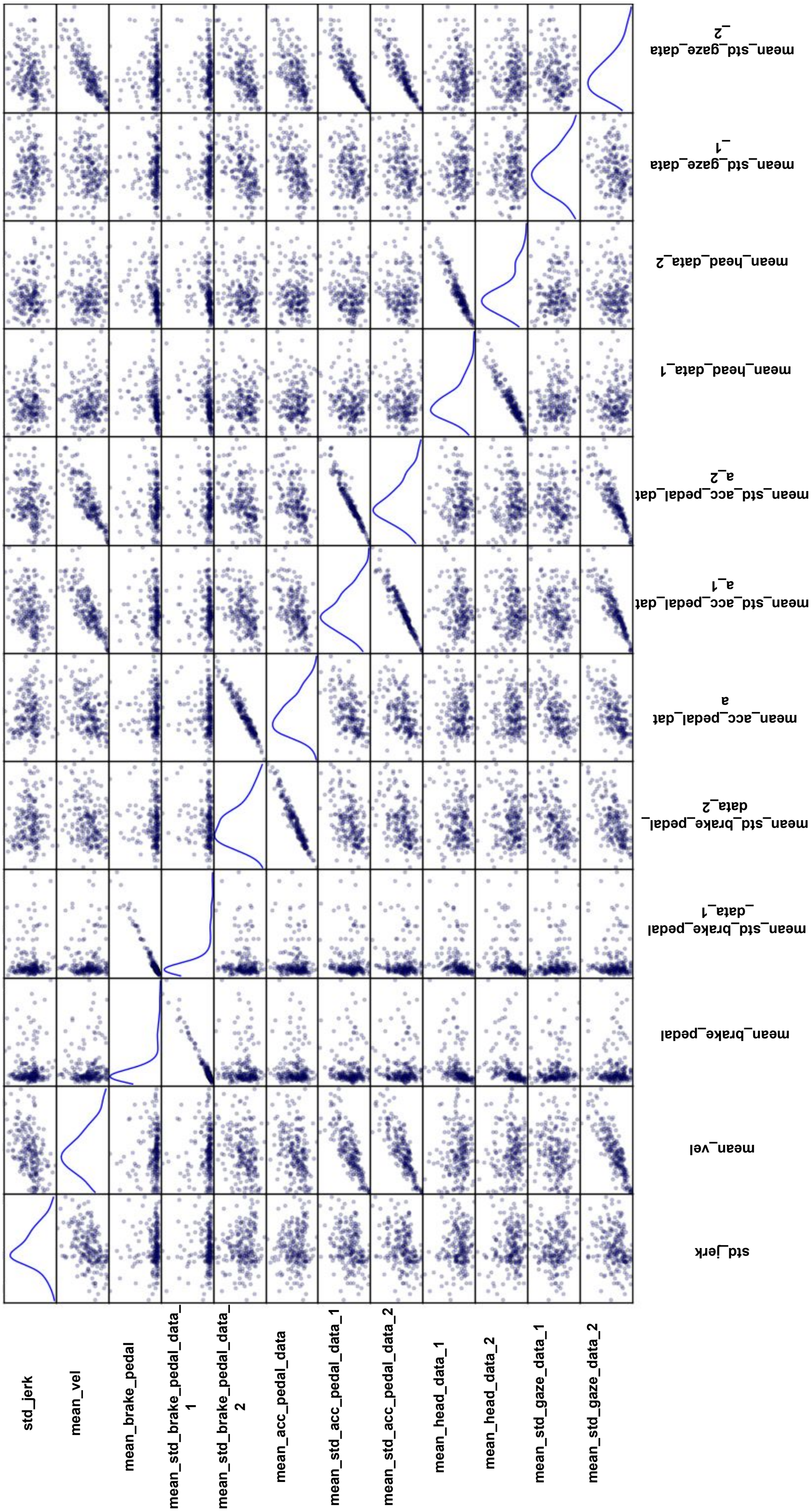


Figure 14: Scatter plot matrix for crossing group 5

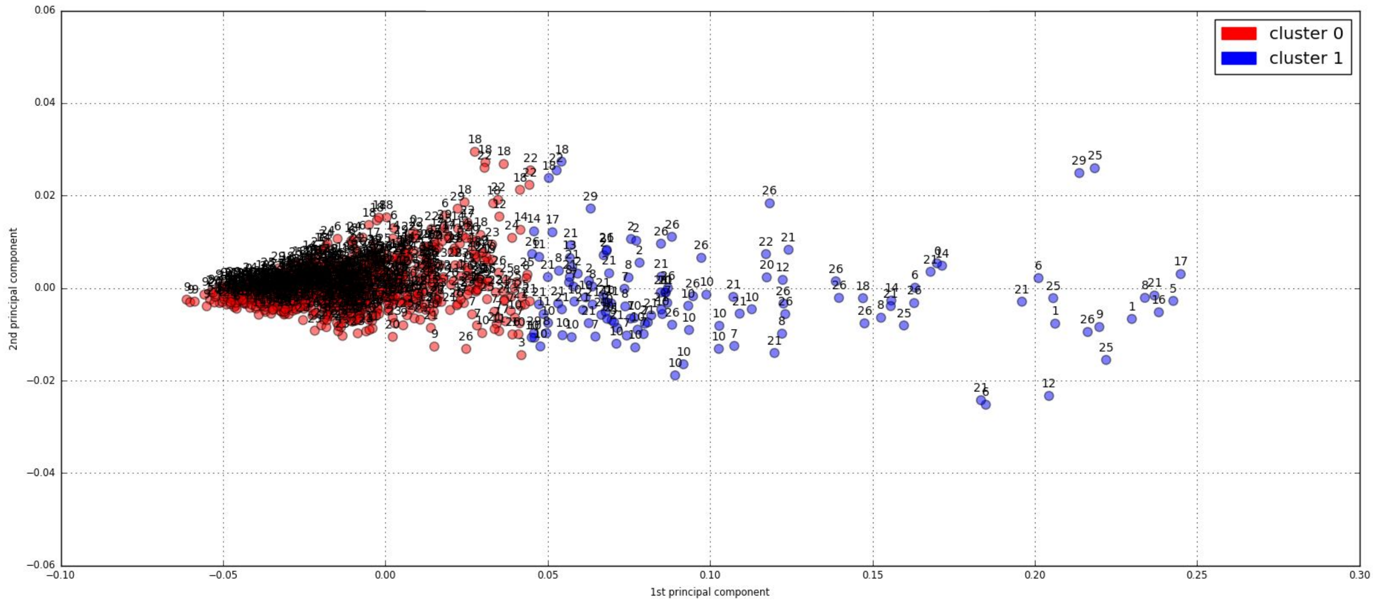


Figure 15: Scatter plot for crossing group 1, features: $mean\_brake\_pedal$, $mean\_std\_brake\_pedal\_data_1$

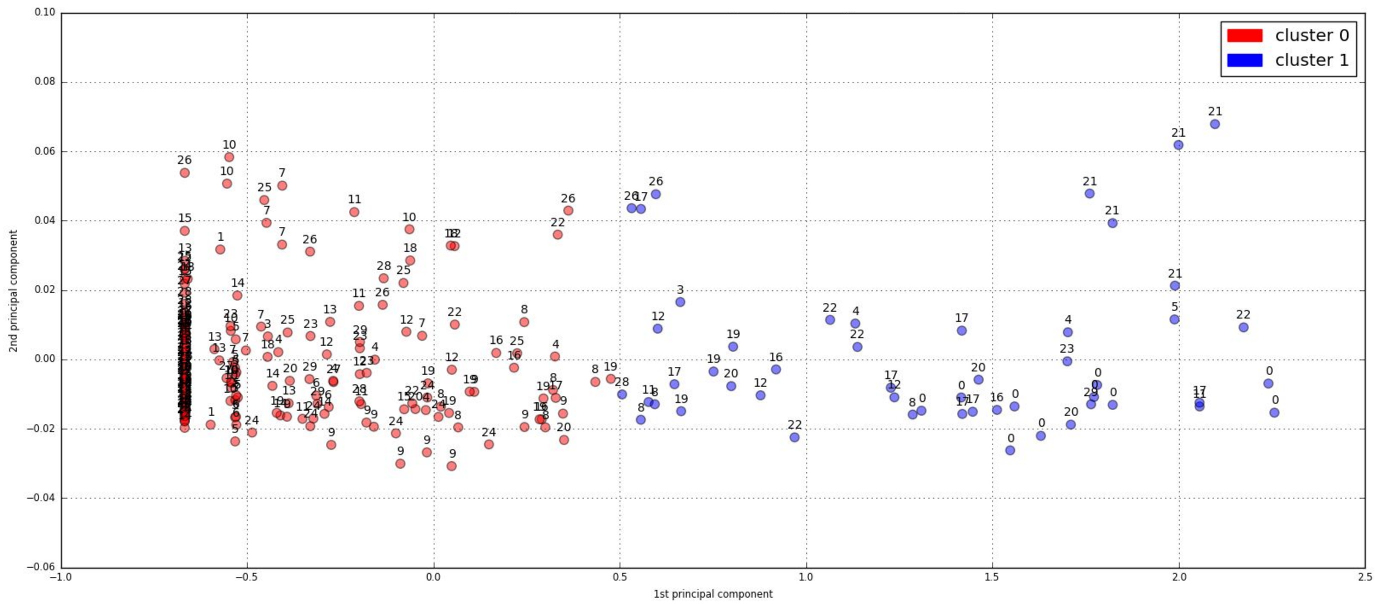


Figure 16: Scatter plot for crossing group 2, features: $mean\_vel$, $mean\_brake\_pedal$

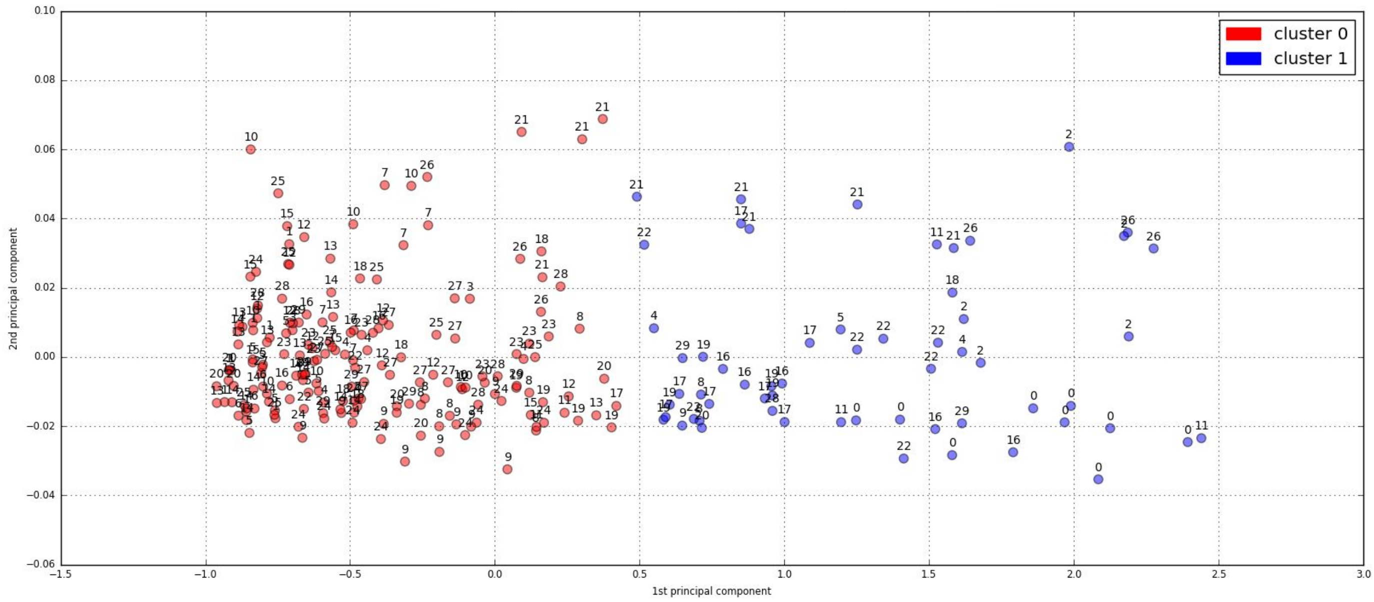


Figure 17: Scatter plot for crossing group 2, features: $mean\_brake\_pedal$, $mean\_std\_gaze\_data_2$

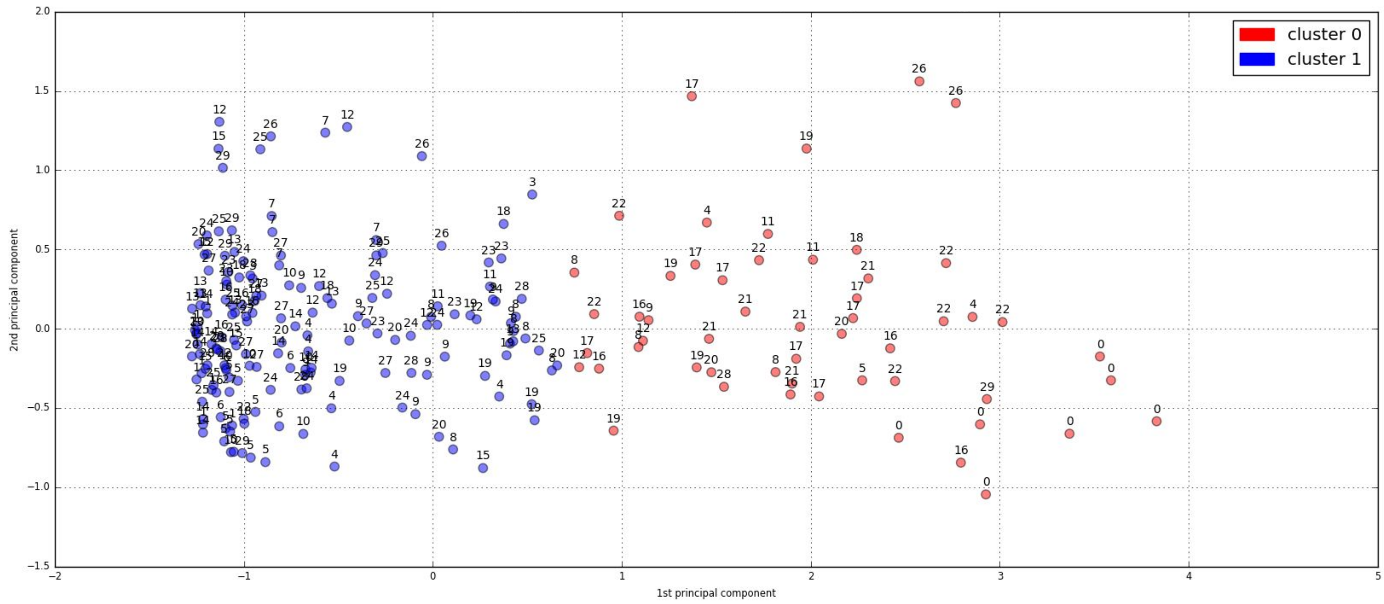


Figure 18: Scatter plot for crossing group 2, features: $mean\_vel$, $mean\_std\_brake\_pedal\_data_1$, $mean\_std\_acc\_pedal\_data_1$, $mean\_head\_data_1$, $mean\_std\_gaze\_data_2$

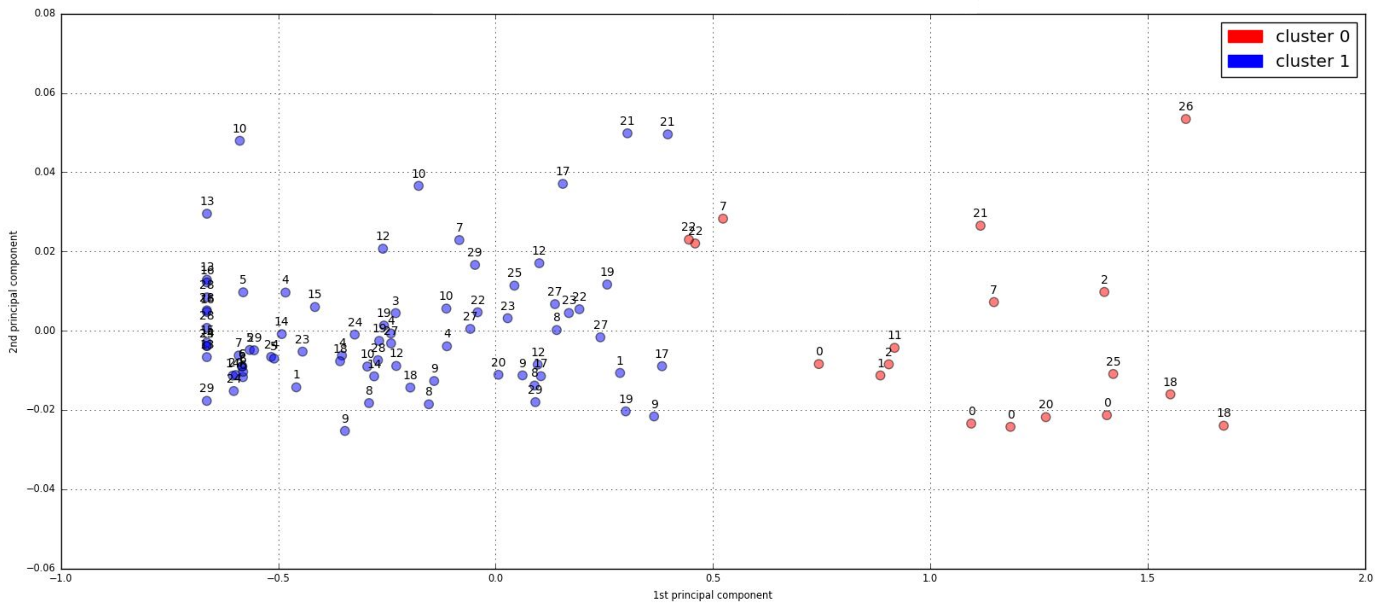


Figure 19: Scatter plot for crossing group 3, features: $mean\_vel$, $mean\_brake\_pedal$

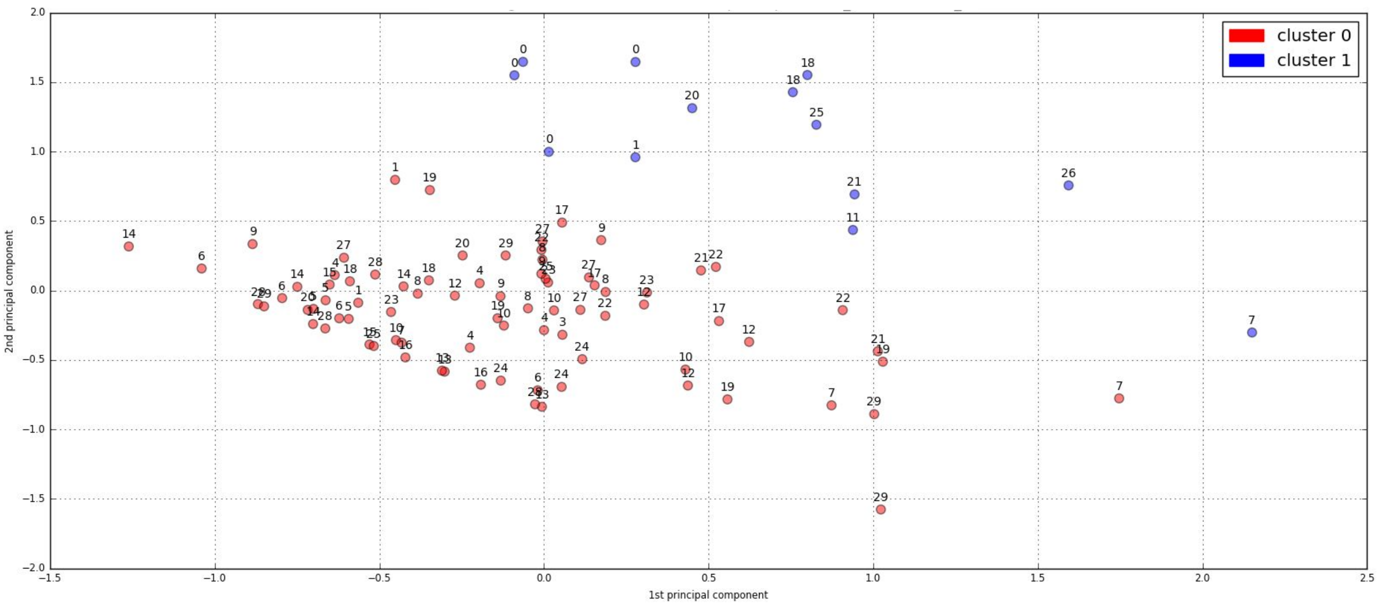


Figure 20: Scatter plot for crossing group 3, features: $mean\_vel$, $mean\_brake\_pedal$, $mean\_head\_data_1$

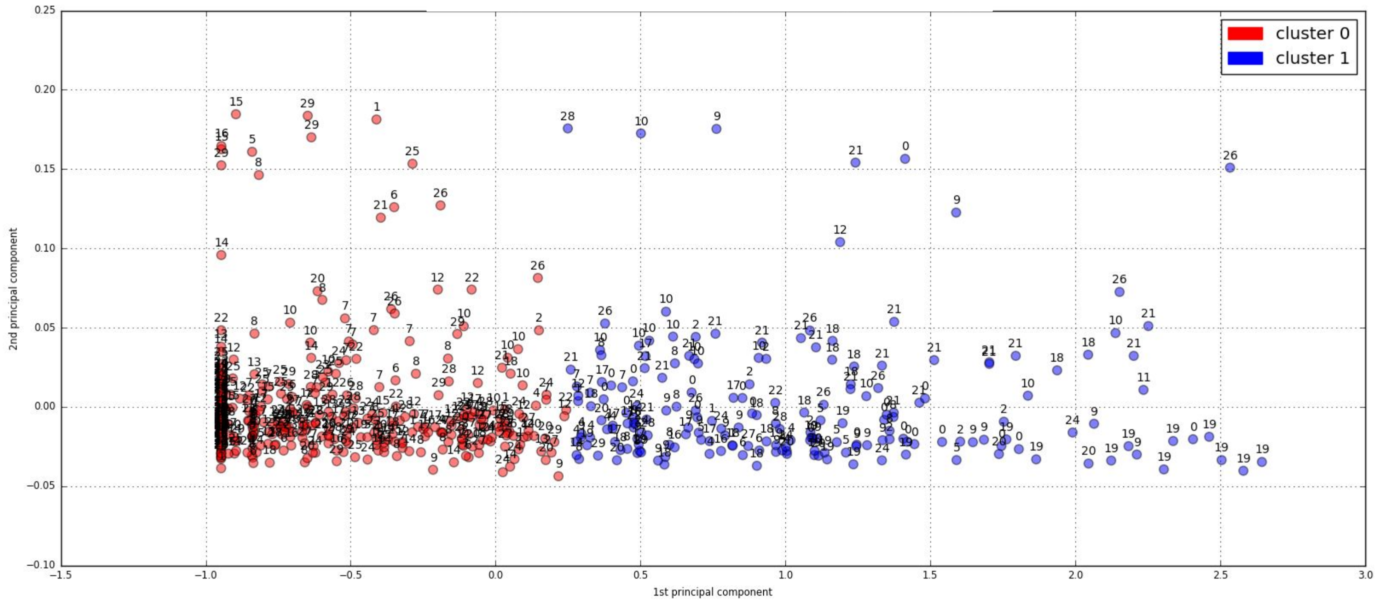


Figure 21: Scatter plot for crossing group 4, features: features: $mean\_brake\_pedal$, $mean\_vel$

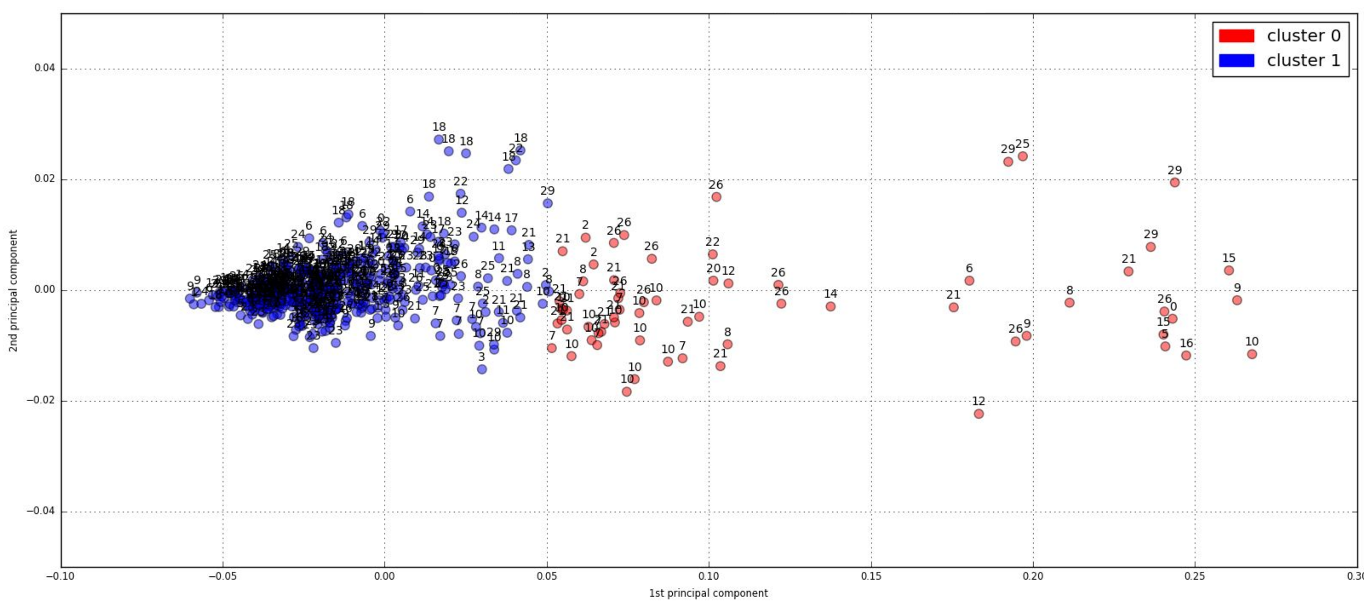


Figure 22: Scatter plot for crossing group 4, features: $mean\_brake\_pedal$, $mean\_std\_brake\_pedal\_data_1$

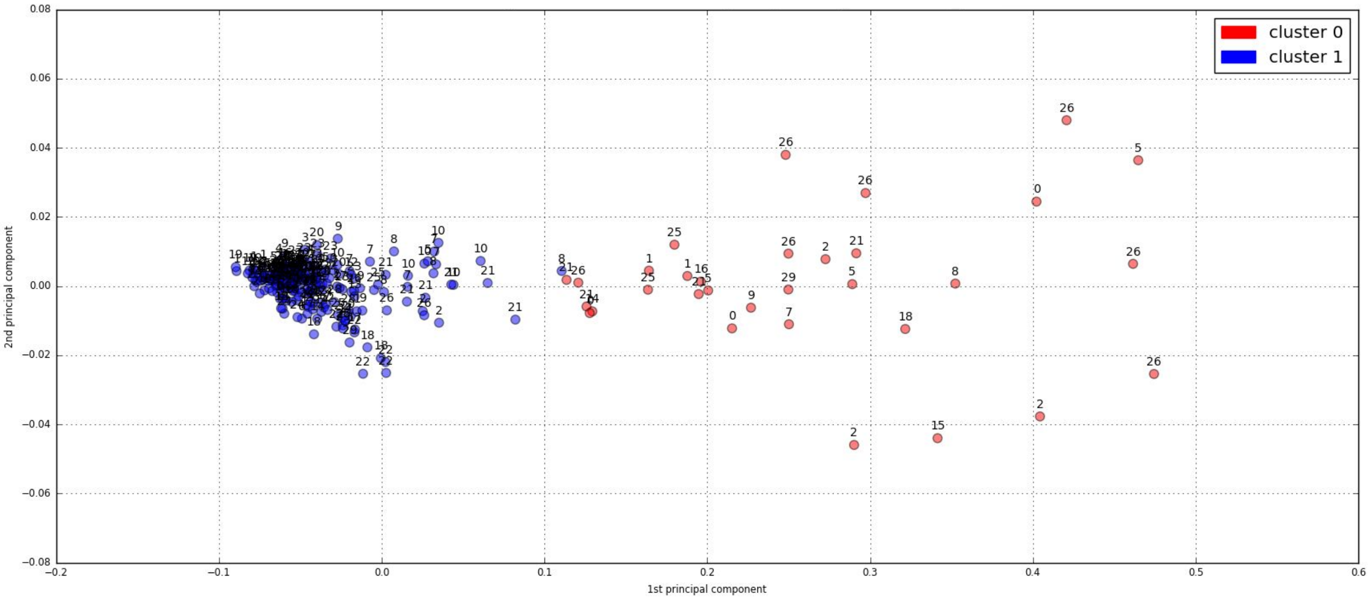


Figure 23: Scatter plot for crossing group 5, features: $mean\_brake\_pedal$, $mean\_std\_brake\_pedal\_data_1$

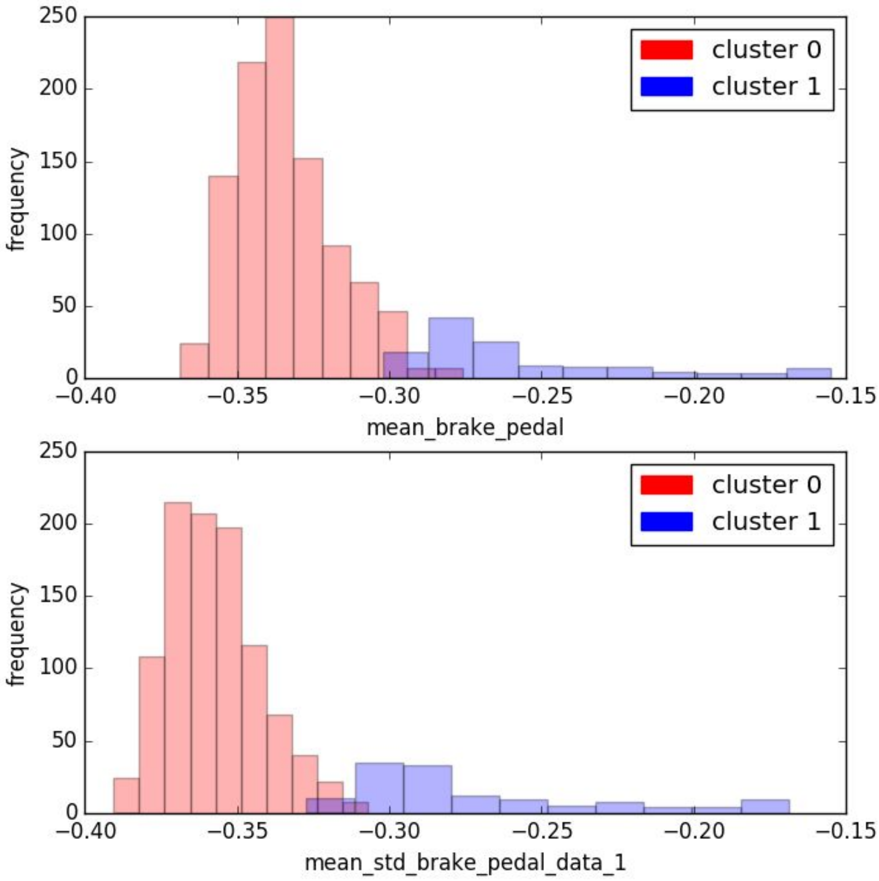


Figure 24: Feature distribution in clusters for crossing group 1, features:$mean\_brake\_pedal$, $mean\_std\_brake\_pedal\_data_1$

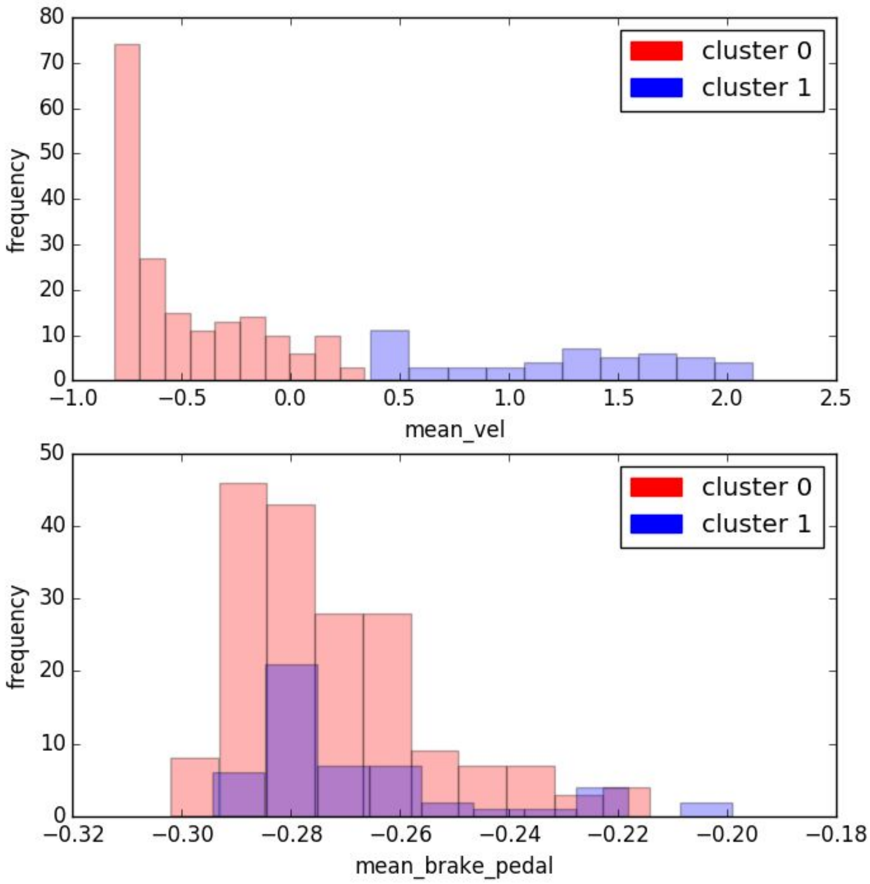


Figure 25: Feature distribution in clusters for crossing group 2, features: $mean\_vel$, $mean\_brake\_pedal$

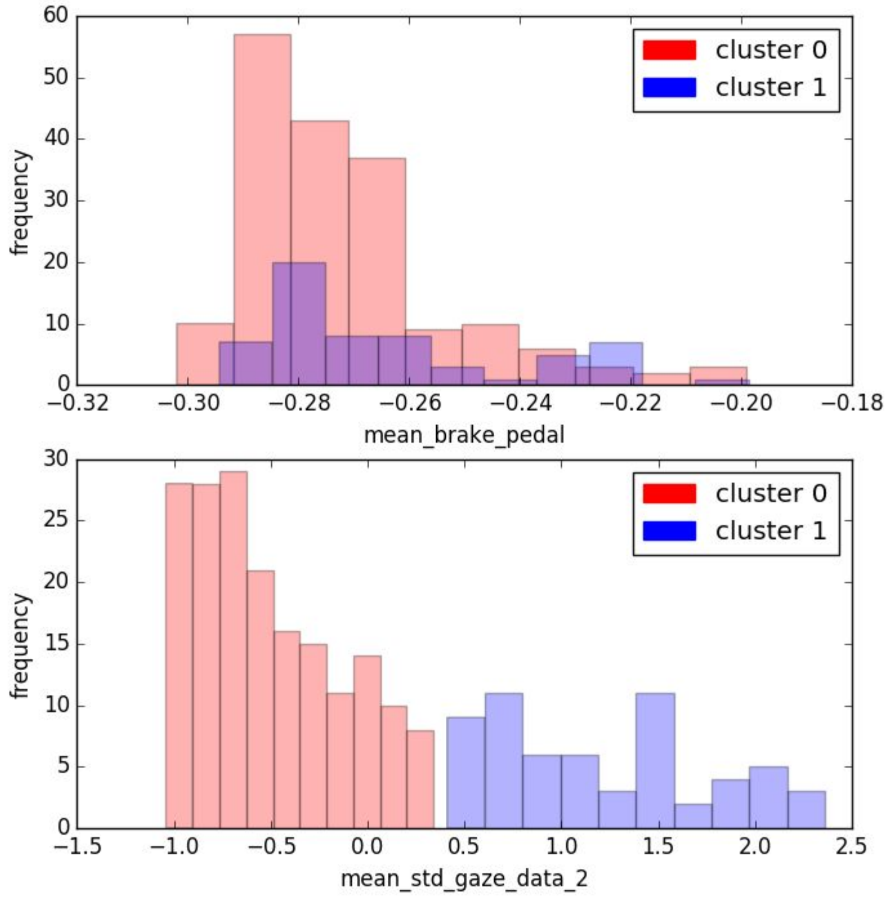


Figure 26: Feature distribution in clusters for crossing group 2, features: $mean\_brake\_pedal$, $mean\_std\_gaze\_data_2$

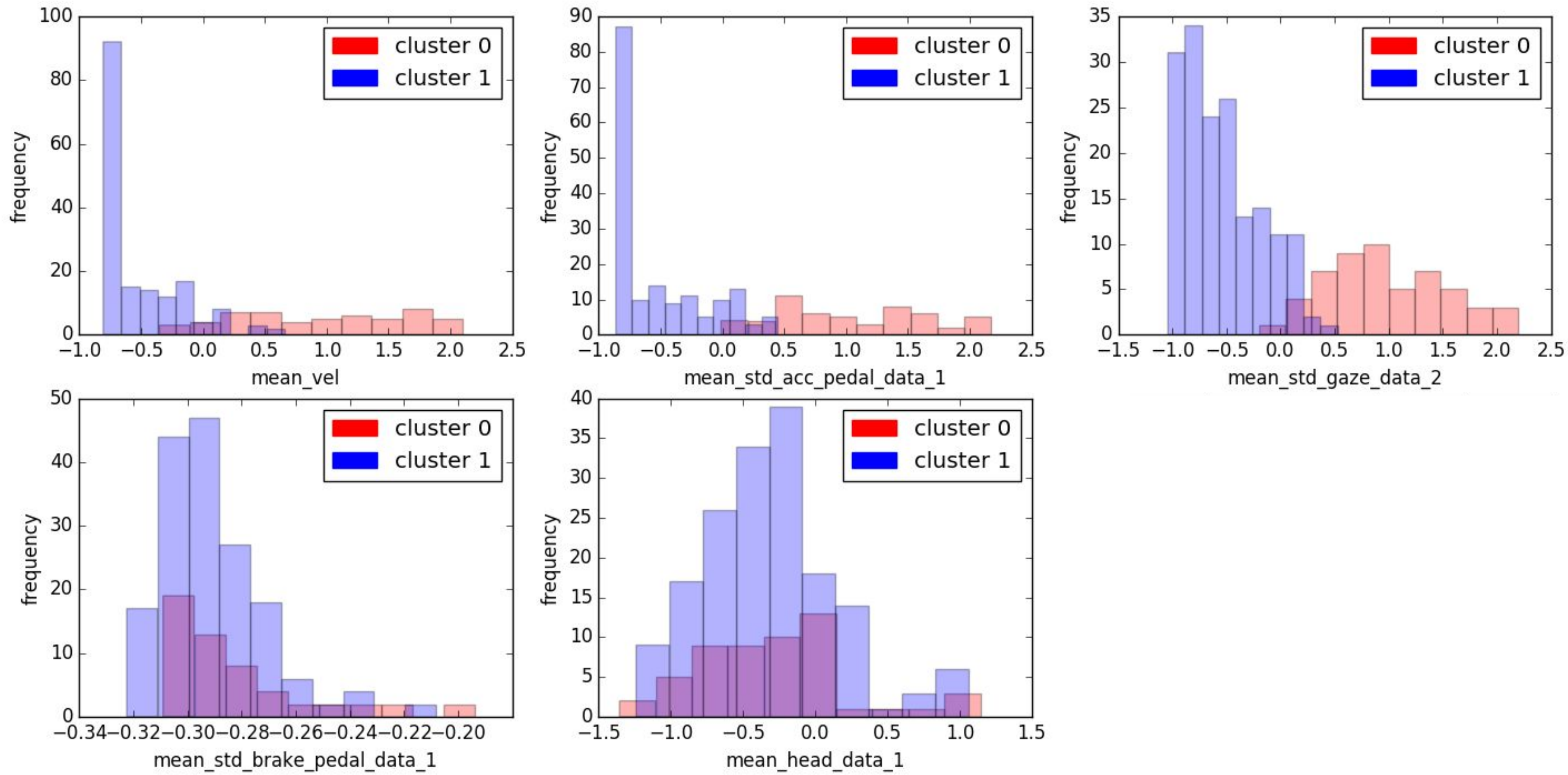


Figure 27: Feature distribution in clusters for crossing group 2, features: $mean\_vel$, $mean\_std\_brake\_pedal\_data_1$, $mean\_std\_acc\_pedal\_data_1$, $mean\_head\_data_1$, $mean\_std\_gaze\_data_2$

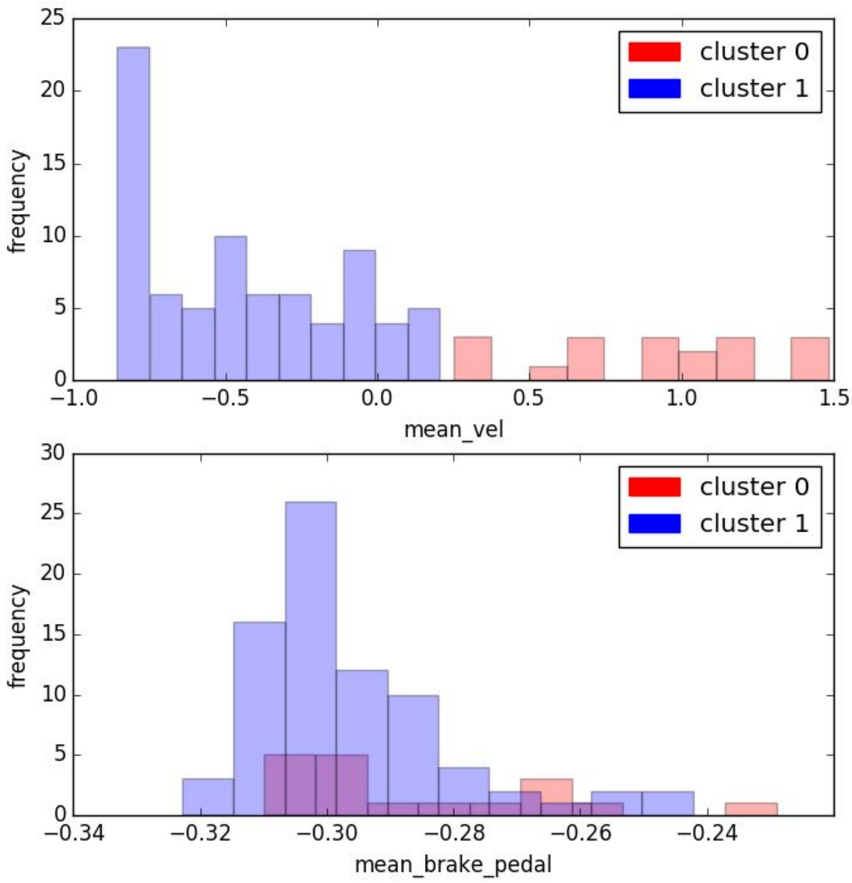


Figure 28: Feature distribution in clusters for crossing group 3, features: $mean\_vel$, $mean\_brake\_pedal$

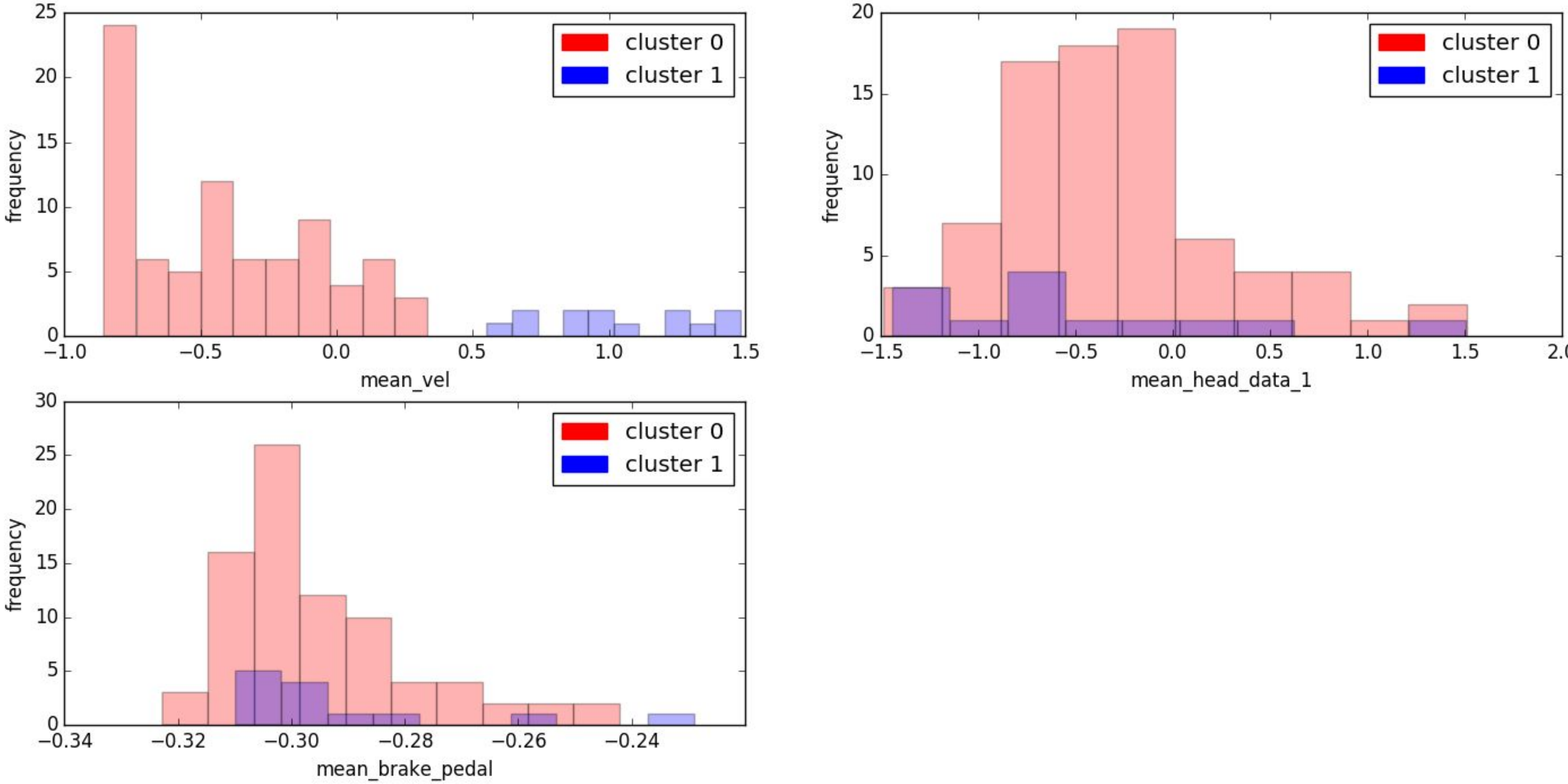


Figure 29: Feature distribution in clusters for crossing group 3, features: $mean\_vel$, $mean\_brake\_pedal$,$mean\_head\_data_1$

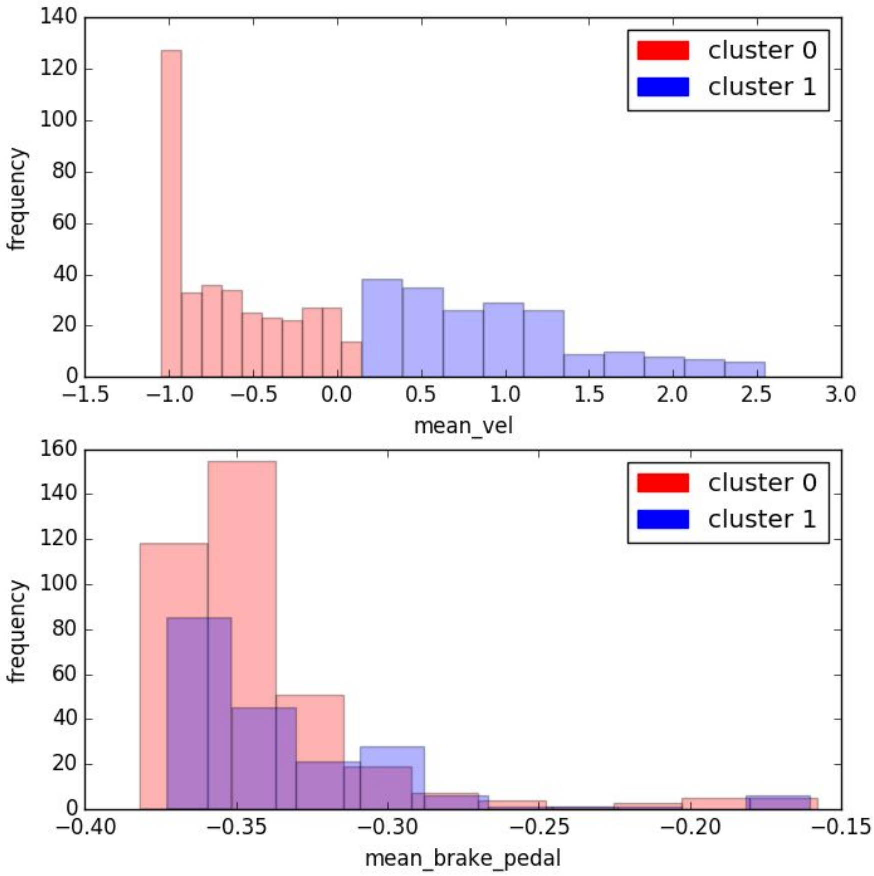


Figure 30: Feature distribution in clusters for crossing group 4, features: $mean\_brake\_pedal$, $mean\_vel$

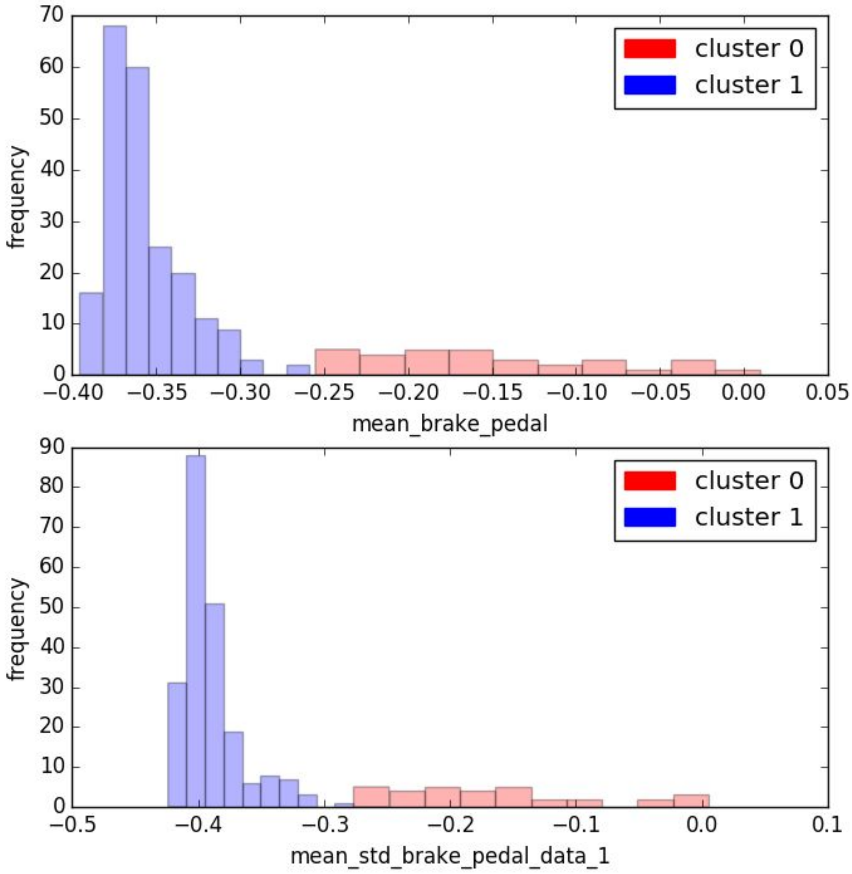


Figure 31: Feature distribution in clusters for crossing group 5, features: $mean\_brake\_pedal$, $mean\_std\_brake\_pedal\_data_1$

References

[1] A Tutorial on Clustering Algorithms, K-Means Clustering. URL: https://home.deib.polimi.it/matteucc/Clustering/tutorial_html/kmeans.html.

[2] G Andria et al. Development of an automotive data acquisition platform for analysis of driving behavior. In: Measurement 93 (2016), pp. 278–287.

[3] Ernst Assmann. Untersuchung über den Einfluss einer Bremsweganzeige auf das Fahrverhalten. na, 1985.

[4] ANDRZEJ Augustynowicz. Preliminary classification of driving style with objective rank method. In: International journal of automotive technology 10.5 (2009), pp. 607–610.

[5] bebenbang. mini Project: Creating Customer Segments. URL: https://benbenbang.github.io/machine-learning/customer_segments_portfolio/.

[6] Hanna Bellem et al. Objective metrics of comfort: developing a driving style for highly automated vehicles. In: Transportation research part F: traffic psychology and behaviour 41 (2016), pp. 45–54.

[7] Michael Breuß et al. Highly Robust Clustering of GPS Driver Data for Energy Efficient Driving Style Modelling. In: arXiv preprint arXiv:1610.02815 (2016).

[8] H Bubb, R Marstaller, and O Sträter. Abschlussbericht SANTOS des Lehrstuhles für Ergonomie (lfe). In: Schlussbericht, Technische Universität München (2002).

[9] German Castignani et al. Smartphone-Based Adaptive Driving Maneuver Detection: A Large-Scale Evaluation Study. In: IEEE Transactions on Intelligent Transportation Systems (2017).

[10] Zoran Constantinescu, Cristian Marinoiu, and Monica Vladoiu. Driving style analysis using data mining techniques. In: International Journal of Computers Communications & Control 5.5 (2010), pp. 654–663.

[11] Joseph F Coughlin, Bryan Reimer, and Bruce Mehler. Monitoring, managing, and motivating driver safety and well-being. In: IEEE Pervasive Computing 10.3 (2011), pp. 14–21.

[12] Dominik Dörr, Konstantin D Pandl, and Frank Gauterin. Optimization of system parameters for an online driving style recognition. In: Intelligent Transportation Systems (ITSC), 2016 IEEE 19th International Conference on. IEEE. 2016, pp. 302–307.

[13] Anup Doshi and Mohan M Trivedi. Examining the impact of driving style on the predictability and responsiveness of the driver: Real-world and simulator analysis. In: Intelligent Vehicles Symposium (IV), 2010 IEEE. IEEE. 2010, pp. 232–237.

[14] Dirk Ebersbach. Entwurfstechnische Grundlagen für ein Fahrerassistenzsystem zur Unterstützung des Fahrers bei der Wahl seiner Geschwindigkeit. In: (2005).

[15] Javier Echanobe, Inés del Campo, and M Victoria Martínez. Design and optimization of a Neural Network-based driver recognition system by means of a multiobjective genetic algorithm. In: Neural Networks (IJCNN), 2016 International Joint Conference on. IEEE. 2016, pp. 3745–3750.

[16] James Elander, Robert West, and Davina French. Behavioral correlates of individual differences in road-traffic crash risk: An examination of methods and findings. In: Psychological bulletin 113.2 (1993), p. 279.

[17] PHILIP ELIASSON and NIKLAS ROSÉN. Efficient K-means clustering and the importance of seeding. In: (2013).

[18] J Engstrom and T Victor. Real-time recognition of large-scale driving patterns. In: Intelligent Transportation Systems, 2001. Proceedings. 2001 IEEE. IEEE. 2001, pp. 1018–1023.

[19] Haneen Farah, Shlomo Bekhor, and Abishai Polus. Risk evaluation by modeling of passing behavior on two-lane rural highways. In: Accident Analysis & Prevention 41.4 (2009), pp. 887–894.

[20] Lino Figueiredo et al. Towards the development of intelligent transportation systems. In: Intelligent Transportation Systems, 2001. Proceedings. 2001 IEEE. IEEE. 2001, pp. 1206–1211.

[21] Davina J French et al. Decision-making style, driving style, and self-reported involvement in road traffic accidents. In: Ergonomics 36.6 (1993), pp. 627–644.

[22] Denos Gazis, Robert Herman, and Alexei Maradudin. The problem of the amber signal light in traffic flow. In: Operations Research 8.1 (1960), pp. 112–132.

[23] Anlin Ge et al. Study on automobile intelligent shift architecture. In: China Mechanical Engineering 5.18 (2001), pp. 106–109.

[24] Highlighting Outliers in your Data with the Tukey Method. URL: `http://datapigtechnologies.com/blog/index.php/highlighting-outliers-in-your-data-with-the-tukey-method/`.

[25] Marika Hoedemaeker and Karel A Brookhuis. Behavioural adaptation to driving with an adaptive cruise control (ACC). In: Transportation Research Part F: Traffic Psychology and Behaviour 1.2 (1998), pp. 95–106.

[26] Hui-Huang Hsu and Cheng-Wei Hsieh. Feature Selection via Correlation Coefficient Clustering. In: JSW 5.12 (2010), pp. 1371–1377.

[27] Hanneke Hooft van Huysduynen et al. Measuring driving styles: a validation of the multidimensional driving style inventory. In: Proceedings of the 7th International Conference on Automotive User Interfaces and Interactive Vehicular Applications. ACM. 2015, pp. 257–264.

[28] Motonori Ishibashi et al. Indices for characterizing driving style and their relevance to car following behavior. In: SICE, 2007 Annual Conference. IEEE. 2007, pp. 1132–1137.

[29] Byeong Wook Jeon et al. Development of Smart Shift and Drive Control System Based on the Personal Driving Style Adaptation. Tech. rep. SAE Technical Paper, 2016.

[30] Derick A Johnson and Mohan M Trivedi. Driving style recognition using a smartphone as a sensor platform. In: Intelligent Transportation Systems (ITSC), 2011 14th International IEEE Conference on. IEEE. 2011, pp. 1609–1615.

[31] k-means clustering. URL: `https://en.wikipedia.org/wiki/K-means_clustering`.

[32] Leonard Kaufman and Peter J Rousseeuw. Partitioning around medoids (program pam). In: Finding groups in data: an introduction to cluster analysis (1990), pp. 68–125.

[33] Gurunath Kedar-Dongarkar and Manohar Das. Driver classification for optimization of energy usage in a vehicle. In: Procedia Computer Science 8 (2012), pp. 388–393.

[34] Raymond J Kiefer. Developing a forward collision warning system timing and interface approach by placing drivers in realistic rear-end crash situations. In: Proceedings of the Human Factors and Ergonomics Society Annual Meeting. Vol. 44. 20. SAGE Publications Sage CA: Los Angeles, CA. 2000, pp. 3–308.

[35] W König et al. SANTOS Situations-Angepasste und Nutzer-Typ-zentrierte Optimierung von Systemen zur Fahrerunterstützung. In: ERGONOMIE UND VERKEHRSSICHERHEIT. KONFERENZBEITRAEGE DER HERBSTKONFERENZ 2000. 12.-13. OKTOBER 2000 AN DER TECHNISCHEN UNIVERSITAET MUENCHEN (2000).

[36] Nobuyuki Kuge et al. A driver behavior recognition method based on a driver model framework. Tech. rep. SAE Technical Paper, 2000.

[37] Toru Kumagai and Motoyuki Akamatsu. Prediction of human driving behavior using dynamic Bayesian networks. In: IEICE TRANSACTIONS on Information and Systems 89.2 (2006), pp. 857–860.

[38] Yulong Lei et al. Research on driving style recognition method based on driver's dynamic demand. In: Advances in Mechanical Engineering 8.9 (2016), p. 1687814016670577.

[39] Yulan Liang, Michelle L Reyes, and John D Lee. Real-time detection of driver cognitive distraction using support vector machines. In: IEEE transactions on intelligent transportation systems 8.2 (2007), pp. 340–350.

[40] Chan-Chiao Lin et al. Driving pattern recognition for control of hybrid electric trucks. In: Vehicle System Dynamics 42.1-2 (2004), pp. 41–58.

[41] Na Lin et al. An overview on study of identification of driver behavior characteristics for automotive control. In: Mathematical Problems in Engineering 2014 (2014).

[42] José Onate López, Andrés C Cuervo Pinilla, et al. Driver behavior classification model based on an intelligent driving diagnosis system. In: Intelligent Transportation Systems (ITSC), 2012 15th International IEEE Conference on. IEEE. 2012, pp. 894–899.

[43] Jianbo Lu et al. From vehicle stability control to intelligent personal minder: real-time vehicle handling limit warning and driver style characterization. In: Computational Intelligence in Vehicles and Vehicular Systems, 2009. CIVVS'09. IEEE Workshop on. IEEE. 2009, pp. 43–50.

[44] X Ma and I Andréasson. Classification of car following regimes based on clustering analysis. In: submit to European Transportation Research Arena, Göteborg (2006).

[45] Rashmi N Mahajan and A Patil. Lane departure warning system. In: International Journal of Engineering and Technical Research 3.1 (2015), pp. 120–123.

[46] R Marstaller. Fahrerverhaltensaenderungen bei der fahrerassistierenden Kfz-Steuerung mit aktiven Bedienelementen gegenueber dem konventionellen Bedienkonzept. In: VDI-Berichte 1613 (2001).

[47] Peter T Martin, Vikram C Kalyani, and Aleksander Stavanovic. Evaluation of advance warning signals on high speed signalized intersections. Tech. rep. Mountain-Plains Consortium, 2003.

[48] Christoph Mayser et al. Fahrerassistenzsysteme zur Unterstützung der Längsregelung im ungebundenen Verkehr. In: Conference Aktive Sicherheit durch Fahrerassistenz. 2004.

[49] Joel C McCall and Mohan M Trivedi. Driver behavior and situation aware brake assistance for intelligent vehicles. In: PROCEEDINGS-IEEE 95.2 (2007), p. 374.

[50] Javier E Meseguer et al. DrivingStyles: A mobile platform for driving styles and fuel consumption characterization. In: arXiv preprint arXiv:1611.09065 (2016).

[51] Ronald Miller and Qingfeng Huang. An adaptive peer-to-peer collision warning system. In: Vehicular technology conference, 2002. VTC Spring 2002. IEEE 55th. Vol. 1. IEEE. 2002, pp. 317–321.

[52] Dejan Mitrovic. Reliable method for driving events recognition. In: IEEE transactions on intelligent transportation systems 6.2 (2005), pp. 198–205.

[53] Yi Lu Murphey, Robert Milton, and Leonidas Kiliaris. Driver's style classification using jerk analysis. In: Computational Intelligence in Vehicles and Vehicular Systems, 2009. CIVVS'09. IEEE Workshop on. IEEE. 2009, pp. 23–28.

[54] Vladimir Nikulin. Driving Style Identification with Unsupervised Learning. In: Machine Learning and Data Mining in Pattern Recognition. Springer, 2016, pp. 155–169.

[55] Nuria Oliver and Alex P Pentland. Graphical models for driver behavior recognition in a smartcar. In: Intelligent Vehicles Symposium, 2000. IV 2000. Proceedings of the IEEE. IEEE. 2000, pp. 7–12.

[56] Tarek Ouali et al. Driving Style Identification Algorithm with Real-World Data Based on Statistical Approach. Tech. rep. SAE Technical Paper, 2016.

[57] Panagiotis Papaioannou. Driver behaviour, dilemma zone and safety effects at urban signalised intersections in Greece. In: Accident Analysis & Prevention 39.1 (2007), pp. 147–158.

[58] Principal component analysis. URL: `https://en.wikipedia.org/wiki/Principal_component_analysis`.

[59] Sebastian Raschka. About Feature Scaling and Normalization – and the effect of standardization for machine learning algorithms. URL: http://sebastianraschka.com/Articles/2014_about_feature_scaling.html.

[60] Ofer Raz, Hod Fleishman, and Itamar Mulchadsky. System and method for vehicle driver behavior analysis and evaluation. US Patent 7,389,178. 2008.

[61] Philip Rumbolz. Untersuchung der Fahrereinflüsse auf den Energieverbrauch und die Potentiale von verbrauchsreduzierenden Verzögerungsassistenzfunktionen beim PKW. expert Verlag, 2013.

[62] Heike Sacher. Gesamtheitliche Analyse des Bedienverhaltens von Fahrzeugfunktionen in der täglichen Nutzung. Cuvillier, 2009.

[63] Fridulv Sagberg. Road accidents caused by drivers falling asleep. In: Accident Analysis & Prevention 31.6 (1999), pp. 639–649.

[64] Fridulv Sagberg et al. A review of research on driving styles and road safety. In: Human factors 57.7 (2015), pp. 1248–1275.

[65] Dr. Saed Sayad. PCA Principal Component Analysis. URL: http://chem-eng.utoronto.ca/~datamining/Presentations/PCA.pdf.

[66] Scatterplot Matrix. URL: http://www.itl.nist.gov/div898/handbook/eda/section3/scatterb.htm.

[67] Silhouette Score. URL: http://scikit-learn.org/stable/modules/generated/sklearn.metrics.silhouette_score.html.

[68] J Son et al. Comparative study between Korea and UK: relationship between driving style and real-world fuel consumption. In: International Journal of Automotive Technology 17.1 (2016), pp. 175–181.

[69] Keisuke Suzuki and Håkan Jansson. An analysis of driver's steering behaviour during auditory or haptic warnings for the designing of lane departure warning system. In: JSAE review 24.1 (2003), pp. 65–70.

[70] Mikhail Sysoev, Andrej Kos, and Matevž Pogačnik. Smart Driving: Influence of Context and Behavioral Data on Driving Style. In: International Conference on Next Generation Wired/Wireless Networking. Springer. 2016, pp. 141–151.

[71] Orit Taubman-Ben-Ari, Mario Mikulincer, and Omri Gillath. The multidimensional driving style inventory—scale construct and validation. In: Accident Analysis & Prevention 36.3 (2004), pp. 323–332.

[72] Ardalan Vahidi and Azim Eskandarian. Research advances in intelligent collision avoidance and adaptive cruise control. In: IEEE transactions on intelligent transportation systems 4.3 (2003), pp. 143–153.

[73] Minh Van Ly, Sujitha Martin, and Mohan M Trivedi. Driver classification and driving style recognition using inertial sensors. In: Intelligent Vehicles Symposium (IV), 2013 IEEE. IEEE. 2013, pp. 1040–1045.

[74] Manuel Volk. Ego-Vehicle velocity profile prediction at intersections in urban environment. Tech. rep.

[75] Rui Wang and Srdjan M Lukic. Review of driving conditions prediction and driving style recognition based control algorithms for hybrid electric vehicles. In: Vehicle Power and Propulsion Conference (VPPC), 2011 IEEE. IEEE. 2011, pp. 1–7.

[76] Henriette Wallén Warner et al. Cross-cultural comparison of drivers' tendency to commit different aberrant driving behaviours. In: Transportation research part F: traffic psychology and behaviour 14.5 (2011), pp. 390–399.

[77] Robert West et al. Direct observation of driving, self reports of driver behaviour, and accident involvement. In: Ergonomics 36.5 (1993), pp. 557–567.

[78] Wenhui Zhang et al. Drivers' Visual Search Patterns during Overtaking Maneuvers on Freeway. In: International journal of environmental research and public health 13.11 (2016), p. 1159.