CHALMERS

A natural language processing approach for identifying driving styles in curves

Master's thesis in Applied Mechanics

ERIC MCNABB
MARCUS KALANDER

Department of Applied Mechanics
CHALMERS UNIVERSITY OF TECHNOLOGY
Göteborg, Sweden 2016

MASTER'S THESIS IN APPLIED MECHANICS

A natural language processing approach for identifying driving styles in curves

ERIC MCNABB
MARCUS KALANDER

Department of Applied Mechanics

Division of Vehicle Safety

CHALMERS UNIVERSITY OF TECHNOLOGY

Göteborg, Sweden 2016

A natural language processing approach for identifying driving styles in curves
ERIC MCNABB
MARCUS KALANDER

Master's thesis 2016:05
ISSN 1652-8557
Department of Applied Mechanics
Division of Vehicle Safety
Chalmers University of Technology
SE-412 96 Göteborg
Sweden
Telephone: +46 (0)31-772 1000

Cover:
Wordcloud created with the word frequencies of our thesis draft with a slight modification on a few word frequencies. The image was produced by us using: http://www.wordclouds.com/.

Chalmers Reproservice
Göteborg, Sweden 2016

A natural language processing approach for identifying driving styles in curves
Master's thesis in Applied Mechanics
ERIC MCNABB
MARCUS KALANDER
Department of Applied Mechanics
Division of Vehicle Safety
Chalmers University of Technology

Abstract

A machine able to autonomously recognise driving styles has numerous applications, of which the most straightforward is to recognise risky behaviour. Such knowledge can be used to teach new drivers with the goal of reducing accidents in the future and increasing traffic safety for all road users. Furthermore, insurance companies can incentivise safe driving with lower premiums, which in turn can motivate a more careful driving style. Another application is within the field of autonomous vehicles where learning about driving styles is imperative for autonomous vehicles to be able to interact with other drivers in traffic.

The first step towards identifying different driving styles is being able to recognise and distinguish between them. The aim of this thesis is to identify the indicators of aggressive driving in curves from a large amount of naturalistic driving data. The first step was finding curve sections to analyse within trips and the second step was reducing the data to become more manageable. Symbolic representations were used for the second preprocessing step, which in turn allowed the use of Natural Language Processing techniques for the analysis.

We categorise drivers into different groups depending on their perceived tendency towards aggressive driving styles. This categorisation is used to compare the drivers and their driving style with each other. The tendencies used were Speeding, Braking, Jerky curve handling and Rough curve handling. Some general trends among the analysed drivers are also identified. It is possible to reuse the categorisation to include more drivers in the future or to use what we have learned about the features and drivers for further research.

Keywords: Driving style, naturalistic driving data, Latent Dirichlet Allocation, Symbolic Aggregate approximation, data mining, machine learning

Sammanfattning

En maskin som automatiskt kan identifiera körstilar har många användningsområden. En av de viktigaste av dessa är identifiering av riskfullt beteende i trafiken. Mer kunskap inom området skulle kunna förbättra utbildning av nya förare med syfte att minska antalet olyckor och förbättra säkerheten för alla trafikanter. Dessutom skulle försäkringsbolag kunna ge incitament genom att ge lägre premier till förare med säkrare körstil, vilket skulle kunna motivera säkrare körning. En annan tillämpning är inom området för självkörande bilar där kunskap om körstilar är viktigt för att självkörande fordon ska kunna interagera med andra förare i trafiken.

Första steget till identifiering av körstilar är att kunna känna igen och särskilja mellan dem. Syftet med uppsatsen är att använda en stor mängd vardaglig kördata för att kategorisera förare efter hur aggressiva de är under körning i kurvor. För att kunna uppfylla syftet behövdes förarbete, först särskiljdes kurvorna från övrig kördata och sedan reducerades all data för att kunna analyseras och hanteras enklare. Symboliska representationer användes för datareduktionen vilket i sin tur möjliggjorde användning av Natural Language Processing metoder i analysen.

Vi lyckades kategorisera förarna i olika grupper beroende på deras upplevda tendens till aggressiva körstilar. De olika grupperna används sedan för att jämföra förare och deras körstilar med varandra. Tendenserna som hittades var fortkörning, bromsning, ryckig kurvkörning, samt grov kurvkörning. Vi hittade även några generella trender bland förarna. Det är möjligt att återanvända dessa grupper för att lägga till fler förare i framtiden och att använda vad vi lärt oss om köregenskaperna och förare för framtida forskning.

Acknowledgements

We would like to express our deep appreciation for our supervisor and examiner Dr. Selpi who showed interest and helped us start this project on relatively short notice. She was also very supportive and helpful throughout the project.

We would like to thank SAFER for giving us a productive working environment and expanding our horizon with regards to traffic safety immensely.

An honorable mention to Graham Kemp and Olof Mogren who helped us find both project and supervisor. Finally, we would also like to thank family and friends for supporting us throughout the project and special thanks to those who helped with proofreading this report.

Marcus Kalander, Eric McNabb
June 25, 2016

Contents

1 Introduction

1.1 Background

One of the primary incentives for determining driving styles is to increase traffic safety. Previous research suggests a relationship between crash risk and driving behaviour. Parker et al. (1995) investigated the risk of accident involvement by studying the tendency for traffic violations, lack of thoroughness in decision making and speeding. Their results showed that different behavioural characteristics are associated with proneness for different types of accidents. A more recent study by the American Automobile Association Foundation for Traffic Safety stated; "As many as 56 per cent of deadly crashes involve one or more unsafe driving behaviours typically associated with aggressive driving" (AAA Foundation for Traffic Safety, 2009).

Being able to find and determine people's style of driving can contribute to increased knowledge in the area of safe driving. This knowledge can be used in many different fields, one example being education of new drivers. Improved education can, in turn, lead to improved traffic safety for all road users. This is especially true since education mainly targets young and inexperienced drivers who tend to be involved in accidents to a larger extent compared with more experienced drivers. Being able to determine driving styles could also be used in the insurance industry. The insurance fee for those who are more inclined to dangerous driving behaviours could be set higher than for more careful drivers. The incentive for the insurance companies would be the possibility of increasing profits and fairness, but such a system could also increase public traffic safety. If dangerous driving is punished by higher fees and safer driving styles are encouraged, there might be a decrease in aggressive driving. Another application is within the field of autonomous vehicles where learning about driving styles is imperative for autonomous vehicles to be able to interact with other drivers in traffic.

Currently no existing model or system exists which can determine driving style, though a general scheme has been suggested by Sagberg et al. (2015). Thus far, different research groups have used their own set of variables and methods to determine driving style. These have varied depending on what data have been available at the time and which results have been sought. Hence, there are no standard variables to use, although speed and acceleration have been prominent in most studies.

When looking at the crash risk aspect, horizontal curves are highly relevant since most crash studies indicate a higher risk for serious accidents in these road sections (Othman et al., 2009; Radimsky et al., 2016). Driver behaviour in horizontal curves has been studied before. Most of these studies have not used naturalistic driving data, i.e. they have used data from other sources rather than collected from normal day-to-day driving. There have been studies using on-pavement sensors and radar guns (Bonneson et al., 2006; Passetti and Fambro, 1999), but most have used instrumented vehicles or driving simulators (Altamira et al., 2014; Montella et al., 2015). Another common approach is to use questionnaires and analyse the participants' answers (Ishibashi et al., 2007; Deffenbacher et al., 2002). Only a few have used naturalistic driving data, e.g. Othman et al. (2014), and Palmberg et al. (2015).

The data used in this thesis are naturalistic data from the EuroFOT project, the first European large-scale Field Operational Test on Active Safety Systems (EuroFOT, 2012). The project involved more than 28 organisations and data from more than 1000 vehicles were collected, of which 100 were Swedish Volvo cars. Sensors and logging devices were installed in each vehicle and GPS, speed, acceleration and lateral acceleration, among other features, were recorded. The data also include video of the driver, hence, it is somewhat sensitive due to privacy issues. The data collection period was long, in many cases over a year. Hence, the recorded driver behaviour and vehicle data are without any experimental controls which is an advantage when modeling the data. The downside is that the dataset is large and complex.

The EuroFOT dataset has been used in earlier studies at Chalmers University of Technology. This thesis will partially build upon one of them: Imberg and Palmberg (2015). That study focused on specific types of road curves and their geometry. We aim to build upon the knowledge generated about curves and naturalistic driving data and to explore what is possible with the aid of machine learning.

1.2 Objectives

The purpose of this thesis is to investigate the possibility to train a machine to distinguish different driving styles in curves using naturalistic driving data from the EuroFOT project. It also aims to explore the use of machine learning on the large EuroFOT dataset.

The first objective is to extract curves from the naturalistic driving data. To enable this extraction, an informal definition of what a curve is needs to be made.

The next objective is to use machine learning to obtain groups of driving behaviours. The last objective is to connect the groups of driving behaviours; by doing so, we hope to get indicators of certain driving styles (e.g. aggressive).

1.3 Scope

The thesis has only used data collected from the 100 Swedish Volvo cars that took part in the EuroFOT project. Only data from passenger cars were used, no other vehicles such as trucks or bicycles. Driver factors such as nationality, personality, stress level, etc. are not considered. Neither are any specific vehicle traits, since the vehicles are all of the same model.

Only parts of the data in the dataset are used. Relevant factors have been chosen among the available variables. These include, but are not limited to, speed, lateral and longitudinal acceleration. In contrast, all kinds of roads and weather are utilized. These are only limited by their prevalence in the dataset.

Only curves located in rural areas with a speed over 30 km/h are considered in the analysis. In cities the environment and other road users have a large influence over how a person drives for this initial analysis. The curves are mainly located in the Gothenburg area, but there are also trips in the rest of Sweden as well as a few trips through Europe.

No distinction is made between simple and compound curves or whether curves have spirals or not. Curves too close to each other (within 100 meters) and reverse curves are not considered. Curves where the driver overtakes another vehicle or changes lane are allowed but not treated separately or marked in any way.

1.4 Thesis outline

This thesis is comprised of a literature study in chapter 2, which attempts to give an overview of what has been done before within the field of driving behaviour, both in terms of curves, and in terms of machine learning. The chapter also highlights the benefits of using naturalistic data compared to other data gathering methods.

This chapter is followed by a review in chapter 3 of the methods used in this thesis, which includes a short introduction to the field of machine learning and some of the terms used in this paper which are helpful to understand the thesis. It also describes the data reduction technique used in the thesis, symbolic aggregate approximation and the machine learning technique, Latent Dirichlet Allocation.

Chapter 4 describes most of what was done during the thesis project and how we went about reaching the objectives of the thesis. This includes information about the dataset we used and how we interpreted it, as well as how to extract curves from the data. The later parts of the chapter describe how the data were reduced to a manageable amount and how that data were pre-processed before Latent Dirichlet Allocation was used. The chapter is concluded by explaining how we interpreted the final output and driving styles which could be observed.

The results of the thesis are presented in chapter 5, including the curve dataset, the topics that were created by Latent Dirichlet Allocation, and our analysis of the topics.

Chapter 6 contains a discussion about the strengths and limitations of the thesis and the methods used as well as how we tried to minimise those limitations by various means. It also contains reflections on things that

could be improved upon in the method and results. Following this, our conclusions and suggestions for possible future work are presented in chapters 7 and 8.

2 Literature review

The following chapter highlights recent research in the field of driving styles and curves. The papers presented briefly in this chapter will help the reader understand some of the challenges with determining driving styles and applying machine learning methods to solve this problem.

The chapter is divided in three sections. The first section presents general knowledge about driving styles and which factors that affect driving styles. This is followed by a section in which recent studies of driving styles in curves are summarized and a section presenting studies using machine learning to determine driving style. The last section is further divided into two parts, one part with supervised methods and one part with unsupervised methods.

2.1 Driving styles

Sagberg et al. (2015) define driving style as a "habitual way of driving which is characteristic for a driver or group of drivers". A habitual way of driving refers to behaviours which have become habitual to such a degree as to occur reliably and repeatedly in everyday driving. This definition is used in this thesis. Sagberg et al. (2015) also mention that driving styles are subcategories of driving behaviour. Driving style and behaviour tend to change depending on the environment, e.g. different road, traffic and driving conditions.

Habits are formed by a combination of individual characteristics, sociocultural values and technological factors, such as gender, age, type of car, nationality and personality. A driver's behaviour may be reinforced and become habitual as a driver experience situations which are perceived as similar, eliciting the same response that previously had a positive outcome.

Many different terms are used for different driving styles, with little consensus about their definitions. Sagberg et al. (2015) suggest a theoretical framework which is used in this thesis. For classification of driving styles, they suggest a distinction between global and specific driving styles: Each global driving style, for instance; aggressive, calm, careful, etc. contains a subset of specific driving styles, e.g. speeding, tailgating, inappropriate honking. Some driving styles tend to have some commonalities, for example; an expedient driving style (someone who wants to move forward as fast as possible) and a retaliatory driving style (excessive honking, tailgating, etc.) could both appear to be the seen as aggressive driving but have different causes.

Driving styles are determined either by using questionnaires or surveillance (i.e. a human observer inside vehicle, simulator, instrumented vehicles or naturalistic driving data). However, in some cases it was found that participants repeatedly underestimated or overestimated certain aspects of their driving in questionnaires (Amado et al., 2014). Results from questionnaires can be biased (Sagberg et al., 2015) and caution should be used when drawing conclusions from such results.

Analysis of driving behaviour over many trips and under a longer period of time is preferred, i.e. using naturalistic driving data. This kind of approach has, as of yet, not been used extensively. Instead, driving behaviour in a more controlled environment has been studied. By utilizing these more controlled methods different types of scenarios can be isolated and tested separately. In addition, comparing data between drivers is simple.

Naturalistic driving data are hard to analyse and there are many unknown factors influencing the data which can discourage its use. However, naturalistic data seem to be the only way to avoid the bias that is inherent in other methods. Another factor is the data collection phase which takes a lot of time and effort. In recent years, smartphones have been used to simplify the data collection for naturalistic data (Meseguer et al., 2013; Hong et al., 2014). In this way larger sets of data can be easily acquired. The downside is that the internal signals in the car cannot be collected, only data such as speed, acceleration, etc.

A study investigating correlation between personality and driving behaviours showed that those with positive attitudes towards traffic safety were safer drivers and vice versa (Kong et al., 2013). This exemplifies the difficulty in determining driving styles when factors which affect driving style exist that cannot be measured by

sensors. Just as the current mindset has an influence, stress will also have an effect on driving but will only impact driving style if its a habitual conduct. For example, braking for red lights might be a violation caused by stress or some other factor, but it can also be a mistake by the driver. This is not considered a driving style unless the mistake is made repeatedly.

Gaze and glance behaviour can be an indicator of focus and skill. An inexperienced driver is more likely to look very close in front of the vehicle which would make them less inclined to notice things happening further away. This in turn might make the driver appear to be driving in a reckless way even though he or she is mostly inexperienced. Younger drivers also tend to drive faster than more experienced drivers, as well as exhibiting other risky behaviours such as keeping shorter gap distances and running yellow lights. Hence, there should exist some kind of beginner style that is common among new drivers. However, it does make it harder to identify other styles of driving.

The environment will also influence the driving style. Boyle and Mannering (2004) found in their study that the driving speeds on icy or wet roads were lower compared to bare roads and that speed limits were rarely exceeded on wet and icy roads. However, Fitzpatrick et al. (2000) did a systematic review that indicated that speed does not change much when driving on wet roads compared to bare roads. Although there was a slight decrease in speed it was statistically insignificant. A larger difference was instead noted during heavy rain, with speed differences of 5-10km/h lower than normal. This is plausible considering that heavy rain or snow will affect the road conditions to a larger degree and it will also affect a driver's line of sight. Likewise, another study by Wallman (1998) found that drivers tended to reduce their speed depending on how slippery they perceived the road to be as opposed to the actual friction of the road. They also found (somewhat contradictory to the study by Boyle and Mannering (2004)) that drivers when driving on winter roads chose a speed between 75% and 90% of their speed on bare roads, and that speeds were at least 11km/h lower on winter roads.

There are comparatively fewer studies on the differences between driving during the day and driving at night. The time of day seems to be largely insignificant for the choice of speed on two-way rural highways. Fitzpatrick et al. (2000) state that the majority of drivers do not change their speed on two-way rural highways during nighttime, which Quaium (2010) and Donnell et al. (2006) corroborate.

There have been several studies exploring whether there are any clear differences in driving styles among the different genders. Sagberg et al. (2015), in their review of several papers, found that the relation between driving styles and gender is unclear, with papers reporting mixed results. However, a trend that men are prone to a riskier driving style is visible. This could be connected to biological factors (e.g. testosterone) and sociocultural factors.

Regarding the age factor, Sagberg et al. (2015) and MacAdam et al. (1998) found that young drivers generally have a riskier, more aggressive driving style compared to the average driver. Older drivers tend to adopt a more careful driving style than average. This is believed to be caused by young people lacking experience and not having fully developed cognitive functions. Furthermore, older drivers' driving styles might be affected by medical impairments and sociocultural norms of how older drivers are supposed to drive.

Previous research has also found that drivers who are less experienced are more likely to speed and have a more aggressive driving style (Li et al., 2015; Krahe, 2005). Lajunen and Parker (2001) also found in a study with women that drivers with more mileage had a more careful driving style; for men the result was that mileage did not have any relation to aggressiveness. However, results from Machado-León et al. (2016) indicate that drivers who are experienced (drive more than 10,000 km annually) tend to consider certain risky behaviour (speeding, being distracted, etc.) to be less dangerous, hence, are more likely to drive in an aggressive way. Another study by Krahé and Fenske (2002) found that annual mileage did not affect the aggressiveness of the driver.

As mentioned before; weather, personality, road conditions, etc. will have varying effects on driver behaviour. In the best of worlds, all factors affecting the driving style would be considered but some are very hard to determine from raw traffic data. Driving style is driving behaviour that occurs reliably and repeatedly, which makes it so that while weather and road conditions does affect the driving behaviour, it should not affect the particular driving styles much.

2.2 Driving styles in curves and curve extraction methods

All aspects of driving style and driving behaviour mentioned in Section 2.1 apply not only to general driving but driving in curves as well. Hence, this section will be about the differences that occur in curves as opposed to straight driving, and how these curves can be located in the dataset. The differences in driving styles are mainly an effect of the curve geometry. The section therefore begins by introducing some necessary knowledge about curve geometry. The geometry is not the main area of interest for this study and it will be brief. However, the terminology is needed to comprehend the following sections.

2.2.1 Curve geometry

The straight road sections approaching and leaving curves are labeled tangents. Tangents are one of three components building up a horizontal curve. The other two components are circular curves and spiral transitions. All curves have at least two tangents, one in the beginning and one in the end, but in between there can be any combination of elements. The simpler curves comprise two tangents and a combination of circular curves. There are mainly three combinations to note: simple, compound and reverse curves (Imberg and Palmberg, 2015).

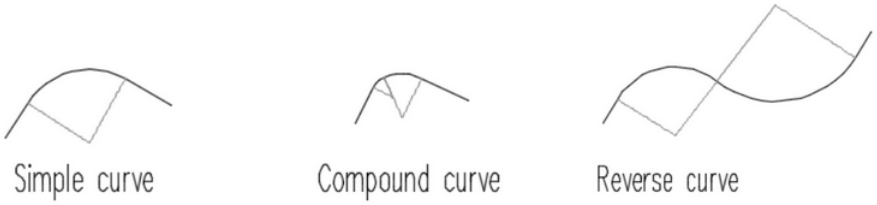


Figure 1: *Three different types of curves formed by two tangents and circular curves.*

A curve with a single constant radius is called a simple curve. If the radius changes inside the curve it can be seen as two or more curves in succession that are turning in the same direction. These are called compound curves. A reverse curve is two simple curves with equal radii turning in opposite directions with a common tangent (Garber and Hoel, 2009).

The last element to build up curves are spiral transitions, also called transition curves. Spiral transitions are commonly placed between tangents and circular curves, or between two adjacent circular curves with considerably different radius. The purpose of using them is to make the curve smoother by having a gradual change in curvature and lateral acceleration (Lannér et al., 2000). Hence, driving comfort will increase, which can explain their wide usage today. Other reasons could be their aesthetically pleasing features and usage as a transition for superelevation (Banks, 2002). Two curves are displayed in Figure 2 to exemplify the usage of spiral transition elements in curves.

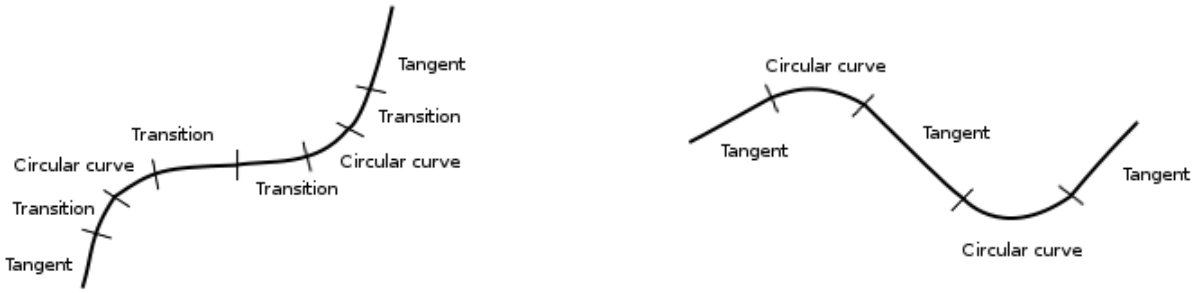


Figure 2: *A comparison between a curve with spiral transitions and a curve without. As can be seen in the figure the curve with spiral transitions is much smoother, hence, increasing driver comfort.*

2.2.2 Previous studies on driving styles in curves

There have been previous studies researching driving styles and driver behaviour specifically in curves. Most of these have used simulators or instrumented vehicles (Altamira et al., 2014; Montella et al., 2015; Pérez-Zuriaga et al., 2013) to obtain data since this facilitates both the data collection and analysis. Usually these kinds of tests are performed only under good conditions, i.e. without snow or rain during the daytime, although there have been exceptions (Hu and Donell, 2010; Quaium, 2010). There have also been tests that employ the use of on-pavement sensors and radar guns (Bonneson et al., 2006; Passetti and Fambro, 1999). Only a few have been using naturalistic driving data, e.g. (Othman et al., 2014; Imberg and Palmberg, 2015). A lot of different approaches and features have been tested. The most common features used in the case of curve sections have been speed, longitudinal acceleration and deceleration, as well as lateral acceleration.

Imberg and Palmberg (2015) investigated the relationship between curve geometry and speed. They found that the factor that influences the speed the most was the radius of the curves. They established that drivers decrease their speed in curves, especially in curves with a small radius. The entering speed and speed inside a curve increased with larger radius in nearly all cases. These results corroborate previous studies done in the area (Othman et al., 2014; Quaium, 2010; Montella et al., 2015). Othman et al. (2014) states that the curve radius is more influential on speed than the posted speed limit.

Imberg and Palmberg (2015) also found that the speed differentials and lateral acceleration were lower for curves with larger radius. This means that the speed before entering curves with larger radius does not significantly decrease. At the same time drivers accelerate less out of curves with larger radius. By combining these results they conclude that larger radius curves have a more consistent speed profile as well as higher overall speeds. Results from Hu and Donell (2010) are similar; they found that when leaving curves, the acceleration was higher for curves with small radius compared to large radius curves.

High speeds can occur on the tangents before a curve if they are long, presumably due to a longer stretch of road to accelerate on (Hu and Donell, 2010). A longer approach tangent will also result in higher speed differentials as the driver needs to decelerate more in the curve (Imberg and Palmberg, 2015). The effect of longer exit tangents is contradicting. Hu and Donell (2010) found that drivers accelerate less on long exit tangents, presumably due to the long stretch to accelerate on. In contrast to this, Imberg and Palmberg (2015) results imply that drivers accelerate more on these road sections.

The presence of long spiral transitions allows for higher speeds when approaching a curve. In these curves, drivers also tend to have a higher speed in the middle. The maximum lateral acceleration is not affected, except for very small curves (Helmers and Törnros, 2006). This indicates a changed trajectory for these curves as either speed needs to be reduced or a trajectory of larger radius must be chosen to reduce lateral acceleration (Boer, 1996).

Previous curves will affect the driver behaviour in a curves although Imberg and Palmberg (2015) state that its influence is low compared to curve radius and spiral length. The previous curve have a higher significance

when the their radius is small. If the previous curve was large, both a higher speed reduction and a decrease of maximum lateral acceleration can be observed. This is a consequence of drivers being able to keep a higher speed in the previous curve and then accelerating to even higher speeds on the tangent.

The direction of travel is also relevant. In a study by Othman et al. (2014) a higher speed was found for inner (right) curves as compared with outer (left) curves. A higher number of lane changes were also found on inner curves compared to outer curves, with 20% more lane changes occurring on inner curves. It was also noted that the number of lane changes significantly increased with increasing radius.

Altamira et al. (2014) found higher deceleration on the tangent approaching a curve than inside the curvature. When the radius increased the necessary length for the deceleration decreased. Montella et al. (2015) found deceleration lengths between 230 and 50 meters depending on the curve radii. Another study by Pérez-Zuriaga et al. (2013) found the point of deceleration to be somewhere between 50 and 100 meters from the beginning of the curve. A t-test showed the point of deceleration to be not significantly different from a point 70 meters from the beginning of the curve, although with quite a high standard deviation. Therefore, to analyze speed patterns in curves, the first 70 to 100 meters should be taken into account.

As mentioned in Section 2.1 the time of day does not seem to affect the speed on two-way rural highways. For other kinds of roads, this is not always true, a simulator study using only small radius curves (Bella et al., 2014) found lower speeds on (long) tangents before curves during the night. However, the speed inside the curve was nearly the same regardless of time of day. The authors gave a possible explanation to this behaviour; that it is due to the lower visibility at nighttime. Quaium (2010) state that in curves with low retroreflectivity levels, drivers may have slowed down significantly at night compared to day.

2.2.3 Curve extraction methods

When analyzing curves, there is a need to isolate the time frames inside the curves from the rest of the data. This can easily be done if direct video observations are available but for larger amount of data this method is unfeasible.

There are many different methods that can be used to estimate start and end points of a curve as well as the radius (Carlson et al., 2005). A common and accurate method is by estimating curve geometry using a map (Imberg and Palmberg, 2015; Nie, 2006; Pérez-Zuriaga et al., 2013) or as-built plan sheets (Carlson et al., 2005; Syed, 2005). These methods are preferred when dealing with a smaller number of curves and where high precision is required.

When the number of curves increases, other methods need to be employed. Hu and Donell (2010) used a specially equipped van, the Federal Highway Administration's Digital Highway Measurement van. The equipment includes inertial navigation unit, differential GPS (enhanced version of GPS with improved accuracy), stereoscopic cameras, lasers, side-looking radars and more. Using the collected data, estimations of curve radii were performed.

Using traditional GPS to estimate start and end points of curves have also been done successfully (Altamira et al., 2014; Othman et al., 2012; Carlson et al., 2005). Altamira et al. (2014) used five heading recordings for each curve and then identified the azimuth variations to compute start and end points. Othman et al. (2012) used naturalistic driving data and the heading variable. Only curves with five or more trips passing through were considered. A mean value of the heading for these trips was calculated to estimate the start and end. The heading change in the curve and curve length were then used to estimate the curve radius. "An accurate path curve radius cannot be found if there are few trips on the curve" (Othman et al., 2012). The use of mean values from multiple trips is preferred when using GPS due to the inaccuracy in the instrument.

Other methods include using steering angle, yaw rate, lateral acceleration or field studies (Carlson et al., 2005; Godthelp, 1986). Estimation using maps or as-built plans are always more accurate. GPS is nearly as accurate when measurements from both inner and outer curves of at least five trips are available. When measurements from only one trip are considered, there will be inaccuracies when using GPS or other measured parameters.

2.3 Machine learning to classify driving styles

There are mainly three different approaches for classification problems: supervised and unsupervised learning or a combination of the two. Depending on the data, one method is more suitable than the other. Supervised methods require labeled data for training and unsupervised methods do not. Supervised methods have been dominating the area of classifying driving styles. A shorter summary of some recent studies relevant to this thesis will be outlined below. For more information on some of the machine learning terms used in this section it is recommended to read chapter 3.

2.3.1 Supervised learning

When using naturalistic driving data, labeled data are rarely available. Hence. studies using supervised methods are commonly using other surveillance methods, e.g. simulators or instrumented vehicles. Generally, the classes of the data are known in advance; for example, drivers have been divided into a group of unskilled drivers and a group of highly skilled drivers. Since this is known beforehand, supervised methods can be applied. However, a drawback of such an approach is that a large bias will be introduced.

MacAdam et al. (1998) used range (distance to the vehicle ahead) as well as range rate (the derivative of range with regards to time) as inputs to an artificial neural network. The output categories represented the speed of the vehicle compared to the vehicle ahead. There were five categories ranging from closing-in rapidly to falling behind rapidly. An "aggressivity index" was then used to compare the drivers, taking into account the percentage of time spent closing-in and falling behind, as well as the time spent following. Combined with the drivers' ages, they showed trends of behaviour in their respective age groups. Younger drivers tended to drive more aggressively, but no noticeable difference could be observed between the middle-aged and older groups.

MacAdam et al. (1998) also tried a similar technique to simulate the accelerator pedal position with range and range rate as inputs to a neural network. However, the results were not very favourable; the prediction did not perform as well as intended. The authors believed a possible explanation could be the limited sensor information used. Other factors could affect the driver's behaviour such as visual distractions, nearby vehicles and slopes. Hence, a more robust model that includes other factors in addition to the two already used could build a more general model.

One of the more popular machine learning techniques used to produce classifiers is Support Vector Machines (SVMs). Two recent studies utilizing SVMS were conducted by Zhang et al. (2010) and Chandrasiri et al. (2012). To capture the temporal characteristics of the data, both studies used some kind of dimensional reduction. Zhang et al. (2010) used discrete fourier transformation and Chandrasiri et al. (2012) used Principal Component Analysis (PCA). PCA especially has been used in several studies because it has high data compression ability, especially for more sparsely distributed data. It achieves this by combining input parameters with high correlation into individual components, while simultaneously removing features that have an insignificant effect on the outcome.

The output from the data reduction is used as input to train the SVM classifiers. The goal of the two research groups was the same, to classify drivers into two groups: experts and typical/low-skilled drivers. Zhang et al. (2010) obtained an overall accuracy of 87.7% (after one often misclassified driver had been removed) by only using steering angle. Chandrasiri et al. (2012) improved on this result by adding more features; speed, longitudinal acceleration, lateral acceleration, yaw rate, etc. They obtained an accuracy of 93.9%, which seems to justify the use of more features in these types classification problems.

Zhang et al. (2010) and Chandrasiri et al. (2012) did not only use SVMs; they both compared this method with other well-known machine learning techniques. In the Zhang et al. (2010) case both Multiple Layer Perceptron Artificial Neural Networks and decision trees were tried. The accuracy of these classifiers were comparable to the trained SVM and only fractionally worse. Chandrasiri et al. (2012) on the other hand used k-Nearest Neighbor (k-NN), a clustering method. For all combinations of features used, SVM performed better than k-NN. The main advantage of k-NN is ease of implementation. Both studies used simulators to gather data, hence. the environments were highly controlled. They state that more complex environments with surrounding vehicles, different road gradients, etc. are preferred.

2.3.2 Unsupervised learning

Previous studies that used unlabeled data are not as common as studies that used labeled data. Most studies using unsupervised methods involve naturalistic driving data, hence, common difficulties are huge datasets and that little knowledge exists on how to interpret the output, e.g. how a machine can infer driving style from raw data. In most studies, some form of dimensionality reduction and/or feature extraction are used to make the data manageable and interpretable. This is followed by a learning algorithm trained with the output from the reduction step.

Constantinescu et al. (2010) used PCA in conjunction with Hierarchical Clustering Analysis (HCA), resulting in six different clusters. The purpose was to use data mining techniques to uncover the possibility of classifying drivers based on their risk-proneness. The results give an idea of possible ways to model a network that use unlabeled data as well as illustrates some of the difficulties working with large datasets. The difficulty due to lack of a formal definition of what aggressive driving entails is likewise evident, i.e. how the data should be interpreted.

The features Constantinescu et al. (2010) used for PCA were speed, positive and negative acceleration, and the mechanical work of the engine. The PCA showed that there were three components that accounted for 92% of the variation in the data. The first component correlated with all types of acceleration whilst the second one correlated with speed values. More components were needed to show the differences in driver's behaviour, therefore rotated components were introduced. These are calculated by the rotation of the principal components using the varimax method, a common post-PCA tool. HCA was used on all of the components (3 regular components and 3 rotated components) because of its properties of minimizing inter-group variation and maximizing outer-group variations. Constantinescu et al. (2010) used Ward's method and Euclidian distance when performing HCA, as this is efficient for clustering.

The results showed that is was possible to categorize the drivers into 5 categories depending on their level of aggressiveness. The scale ranged from very high to very low. Constantinescu et al. (2010) identified a weakness of their study being the low number of features used.

Towfic (2014) went one step further and used an artificial neural network to complement the unsupervised clustering. The features used were the mean and standard deviation for speed and acceleration (both negative and positive), as well as total distance traveled. In the first processing stage, a HCA method variant known as hierarchical agglomerative clustering with Ward's linkage was used to determine which features were useful for certain clusters. In the beginning each observation is in its own cluster, at each step of the algorithm a pair of clusters are merged together. Four classes (clusters) were identified, where each one had a different style of driving.

The classes were used to train a multiple layer perceptron artificial neural network which achieved an accuracy of 95.6% but the sample size merely consisted of 28 observations, making the results less reliable. However, it gives an indication of a good way to train such a network successfully.

PCA was used to learn more of the features used, which resulted in three components that together explained 87.3% of the variation in the data. Each component could take on a range of values depending on its underlying features. The four identified classes were then described by using the three components. Each class was assigned a level of risk depending on the values of the components. E.g. the highest level of risk was shown to be drivers who were speeding often and had high positive accelerations (component 1), frequent braking when driving and shorter braking distances (component 2), as well as a high mean speed (component 3) in their driving.

Piecewise aggregate approximation was used by Lin et al. (2003) to develop Symbolic Aggregate approXimation (SAX), a data reduction technique. SAX divides the data into x equal sized blocks. Each block will be assigned a symbol depending on the mean of the values inside the block. These symbols are in turn used to build up documents. McDonald et al. (2013) showed that SAX could be used to reduce naturalistic driving by 92.6% and still retain the characteristics of the original data and therefore be used for further analysis. For a more detailed explanation of how SAX works see Section 3.2.

McLaurin et al. (2014) used SAX together with a Probabilistic Topic Modeling (PTM) algorithm known as Latent Dirichlet Allocation (LDA). PTM is a group of unsupervised machine learning algorithms which

uses word frequencies in documents to determine the topics of a document. The topics are sets of words determined by the probability distribution over the vocabulary in the documents. LDA is one of the simpler implementations of PTM. A helpful characteristic of LDA is that it automatically removes rare words (i.e. noise). In this way a majority of the data included in the output is relevant.

The purpose of collecting the data was to see whether drivers suffering from obstructive sleep apnea could be distinguished from healthy drivers. Obstructive sleep apnea is a sickness that causes involuntary pauses of breathing during sleep and microsleep episodes during the day. Drivers suffering from the sickness can be identified since they perform small swerves when driving. These indicate a microsleep episode.

In the study, SAX transformed a driving trip worth of data into words which, after using LDA, resulted in 20 topics and a vocabulary of 284 words. To check if topics might distinguish between sick drivers and healthy drivers the topic probabilities were used as features in a random forest machine learning algorithm, which achieved an accuracy of 72.7%.

Bender et al. (2015) used a Bayesian multivariate linear model for pre-processing and dimension reduction. This was followed by an extended LDA algorithm for clustering, which resulted in the classifier dividing trip sections into different colours representing the maneuvers braking, turning, accelerating, coasting and standing still. The authors concluded that the different maneuvers imply an underlying structure of behaviour even though they did not look for behaviours specifically.

3 Machine Learning Methods

The aim of this chapter is to give the reader necessary knowledge to understand the algorithms and machine learning methods used in the thesis. In this chapter, we will first provide a basic introduction to some of the fields within machine learning related to the thesis. The second section describes how Symbolic Aggregate approXimation (SAX) works, whilst the implementation of SAX is described in Section 4.4. The third section describes latent Dirichlet allocation, both its uses and the terminology used, which is useful for understanding the latter half of the thesis.

3.1 Basic introduction to machine learning

Machine learning is a blend between pattern recognition and computational learning theory. The aim of machine learning is that a computer should be able to learn without being specifically programmed for a certain task. Machine learning explores the study and construction of algorithms that can learn from data and make predictions. Many techniques more common in computational statistics and mathematical optimization are used; hence, some knowledge about statistics is required. The following quote summarises the subject well:

"A computer program is said to learn from experience E with respect to some class of tasks T and performance measure P if its performance at tasks in T, as measured by P, improves with experience E." - Mitchell (1997)

Machine learning is used in numerous different computing tasks where designing and programming more explicit algorithms is not feasible. Some of the more common applications are:

Classification
This is the process of identifying which category a new observation belongs to. There must be at least two categories, or more commonly called classes, hence, the process name. A well-known application is the widely used email spam-filter which would categorize incoming email into the two classes; spam or non-spam.

Regression
A statistical process that estimates the relationship between variables. The estimation is done by having one variable fixed and slightly adjust the other variables to observe eventual changes on the fixed variable. Regression is usually used for prediction and forecasting future events, which is in line with most common machine learning techniques. However, compared to classification which works on discrete variables, regression is more concerned with continuous variables.

Clustering
Clustering is similar to classification; however, the groups are not known beforehand. Except for figuring out which class a certain observation belongs to, the clustering algorithm also finds the classes themselves.

Dimensionality reduction
The aim of dimensionality reduction is to simplify data by mapping the input data to some lower-dimensional space, thereby making the data easier to analyse both in terms of time and complexity.

In machine learning the computer learns how to solve a certain task; this is called learning. During the learning phase the computer is trained towards a certain goal. Machine learning algorithms are usually split into two broad categories depending on how they perform their training. These are called supervised learning and unsupervised learning. The key distinction between the two categories is that in supervised learning, the training is performed under supervision, i.e. with existing classes decided by a human or computer. In other words, a sample input and desired output are necessary to train the algorithm. After the training phase, input data can be classified using the accumulated knowledge.

In unsupervised learning, the classes are unknown and the computer must find patterns on its own. This can be useful when looking for patterns that are not easily identifiable or when a network is trained towards a certain set of features. A typical task is to perform a clustering of the inputs by looking at similar features.

3.2 Symbolic Aggregate Approximation

Symbolic Aggregate approXimation (SAX) was developed in order to analyse streaming data, i.e. large amounts of continuous data (Lin et al., 2003). This is a dimensionality reduction technique using symbolic representations, which is necessary for some problems where the original dataset is too large. Hence, many researchers have looked for ways of reducing data and SAX is one of the most successful ways found.

Lin et al. (2003) identified three issues with contemporary symbolic representations. Firstly, the dimensionality of the reduced data tended to be the same as the original data. Secondly, distance measures on the symbolic representation were not the same as in the original. Lastly, they needed access to the entire dataset to make a symbolic representation rather than a partition of the dataset. The whole dataset is not available when using streaming data and hence, contemporary symbolic representations would not work.

To understand how SAX works we begin by denoting the original time series $C = c_1, ..., c_n$ of length n. A reduced time series can be represented in a w-dimensional space by a vector $\overline{C} = \bar{c}_1, ..., \bar{c}_w$ of length w where $w < n$ and normally $w << n$. For each element i in $\overline{C}$ the following equation is calculated:

$$\overline{C}_i = \frac{w}{n} \sum_{j=\frac{d}{w}(i-1)+1}^{\frac{n}{w}i} C_j$$

In other words when C is reduced to $\overline{C}$ (from n dimensions to w dimensions), C is divided into w equal sized "frames". Then the mean value of the values within each frame is calculated, which becomes the reduced representation of the data, $\overline{C}$.

The above step is called Piecewise Aggregate Approximation (PAA). PAA is useful in the simple and intuitive way it reduces the dimensions of the data. In addition, PAA lower bounds the reduced data in such a way that distance measures used on the reduced data will output approximately the same results as a distance measure used on the original data. An example is provided in Figure 3.

The last step of SAX is transforming the reduced representation of the data $\overline{C}$ into a symbolic representation $\widehat{C} = \widehat{c}_1, ..., \widehat{c}_w$. This is done by determining breakpoints in the original data. The breakpoints are a sorted list of numbers that encompasses the range of numbers in the original data. The breakpoints are either determined by looking at the Gaussian distribution of the data and choosing equiprobable values, or by using heuristics.

The set of symbols used for the symbolic representation are called the alphabet, A (e.g. A =[a,b,c,d,e]). The size of the alphabet is determined by the number of breakpoints used in the transformation step. The number of breakpoints must be chosen carefully, because a larger amount of numbers will increase the complexity for later calculations and a smaller amount of numbers will reduce the accuracy. Lin et al. (2003) suggest an alphabet size of between 5 and 8.

Once breakpoints are determined, the value of each PAA segment, $\bar{c}_i$, is compared to the breakpoints. If $\bar{c}_i$ is below the first breakpoint, $\widehat{c}_i$ is assigned the letter a. If $\bar{c}_i$ is below the second breakpoint $\widehat{c}_i$ is assigned the value b, and so on, using a symbol, $s \in A$. The concatenation of symbols is called a *word*, which is the output of SAX.

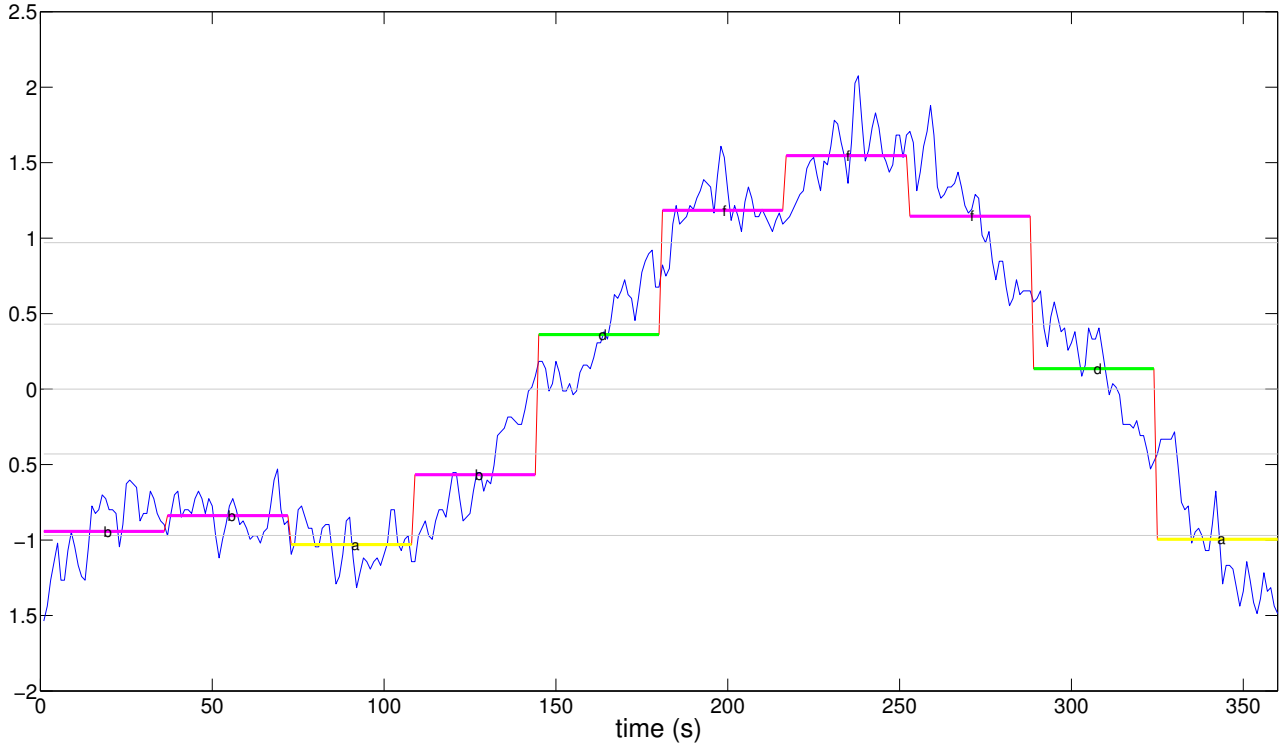


Figure 3: *Example of how SAX is applied to the speed in a large curve and how it is segmented using PAA and then changed into a symbolic representation. It is normalized as in Lin et al. (2002).*

3.3 Latent Dirichlet Allocation

Latent Dirichlet Allocation (commonly referred to as LDA) is a generative probabilistic model for collections of discrete data. The collections of data are usually a text corpus; a large and structured set of texts. The goal of LDA is to "find short descriptions of the members of a collection that enable efficient processing of large collections while preserving the essential statistical relationships" (Blei et al., 2003). This is achieved by explaining a set of observations by unobserved groups; these groups explain similarities in the data. For example, if observations are words in a document, then unobserved groups are a small mixture of topics and the creation of each word is derivable from one of the document's topics. This in turn is useful for classification, novelty detection and similarity and relevance judgments.

To describe LDA and topic modeling, most of the terminology used will be related to the field of natural language processing and actual papers, such as words, documents and corpora. This is in order to guide intuition and facilitate understanding. However, the LDA model can be used with other problems involved with collections of data such as naturalistic driving data. It is important to note that LDA works under bag-of-words assumption, i.e. that words are interchangeable and therefore the order and grammar of the words can be disregarded which is especially useful when working on data.

Formally the terms are defined as follows:
A *word* is a basic unit of discrete data and also an item in a vocabulary, where a *vocabulary* is a collection of all words that exist in a corpus. A *document* contains a list of words with the corresponding number of occurrences of that word in a natural document. Hence, documents can be seen as a collection of word counts, otherwise known as term frequencies. A *corpus* is a collection of documents which become a term-document matrix. Finally, a *topic* is a probability distribution of a vocabulary of words. For each word in the vocabulary there is a probability from 0% to 100% associated with that word. Different topics will contain different probabilities

associated with each word, making it so that each topic has a different set of words and probabilities.

LDA uses latent variables in its model. A latent variable is a hidden variable or a variable that is not known, but assumed to exist and can be inferred from other observed variables. A Dirichlet distribution (often referred to as $Dir \sim \alpha$) is a collection of multinomial distributions. In other words, it is a collection of our prior beliefs with regards to the multinomial distribution chosen from α. In literature the hyperparameters (i.e. α and β values), which are the input parameters to LDA, are usually referred to as single variables rather than distributions (as in Figure 4). This is because if the initial distribution is made of some vector $v = [v_1, ...v_{end}]$, then the final distribution becomes $\alpha * v = \alpha_1, ..., \alpha_{end}$, where α can be chosen as input. Low α values are in order to gain sparsity (i.e. allow zeros) in the data, which allows a document to contain a mixture of very few topics. High α values will make it so that documents may contain a mixture of many documents, i.e. documents will be more similar to each other in regards to which topics they contain. Similarly with high β values, each topic will likely contain a mixture of most of the words while a low β means the topic may contain a mixture of few words.

Usually when working with natural language processing there are three common preprocessing steps required. One is tokenization which is the process of extracting words from a document. The second one is removing stop words such as "and", "or", "but", etc. which are uninteresting words when looking for the meaning in a text. The third is stemming, which is the process of making words that are similar to be identified in the same way, the simpler examples are the removal of pluralization and suffixes.

In LDA, the process works as follows (also shown in Figure 4); For each document in the corpus (M), a topic mixture distribution (θ) is chosen from a per-document topic Dirichlet distribution (α). For each topic (K), a word distribution (ϕ) is chosen from the per-topic word Dirichlet distribution (β). Then for each word in the document (N), a topic (z) is chosen from θ and a word (w) is chosen from (ϕ).

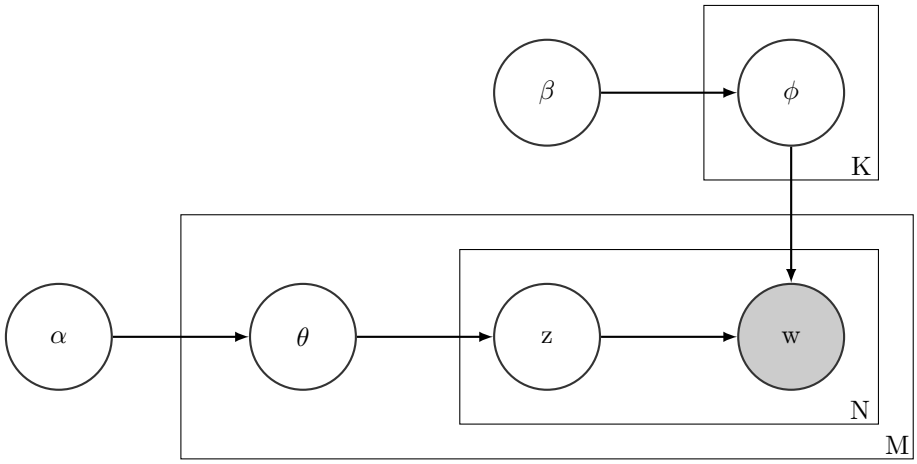


Figure 4: *The LDA model explained using table notation. Within a frame is a loop and within the inner frame is a nested loop. The shaded circles are known variables and the white circles are latent variables.*

This is a generative process where LDA, given a collection of topic-distributions for each document, selects one of the distributions. Then it repeatedly picks both a topic and a word from that topic distribution to generate words in a document. Thus we have a model which generates a corpus. However, as is often the case in practice, the goal is to estimate a model based on a corpus, which essentially is the LDA model in reverse order.

This requires parameter estimation, which is using LDA with a corpus to generate a model, i.e. finding the latent variables; the document topic mixtures (α) and the word distributions (β) that generate the input corpus. The aim is to find a model of the corpus which assigns a high probability to the members of the corpus but also assigns a high probability to other similar documents.

Blei et al. (2003) used variational inference (or variational Bayes) for parameter estimation. A common alternative is Gibbs sampler which is slower than variational inference but simpler to calculate, since the

amount of derivations needed with variational inference grows quickly as the number of dimensions increases. Gibbs sampling is a Markov chain Monte Carlo algorithm (i.e. a probabilistic random algorithm), which is used for obtaining a set of observations from a multivariate probability distribution. The set of observations can then be used to approximate latent variables. As opposed to Gibbs sampling which makes use of random numbers, variational inference is strictly deterministic, causing some bias in the output. This makes it so that Gibbs sampling can produce slightly better results at the cost of being somewhat slower.

LDA is a powerful model partly because it works under the bag-of-words assumption. However, some generalisations are made because the aim is to make sense of a large set of texts. It is only natural to lose some detail. By using LDA there is no need to go through all documents; hence, this loss is acceptable. LDA attempts to build topics unsupervised; however, it is up to humans to give name to these topics and their content.

4 Method

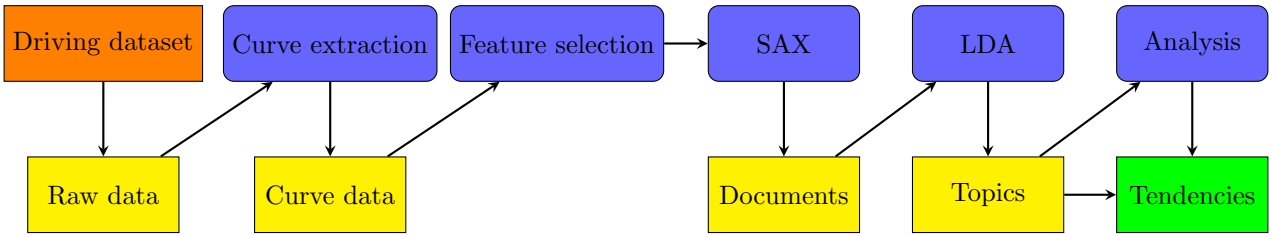


Figure 5: *Flowchart of the work process.*

The method is divided into multiple parts and a brief overview can be seen in Figure 5 above. This chapter is divided into similar sections. The first section is a short description of the dataset used and some difficulties arising from the use of naturalistic driving data. Following this is the curve extraction method in Section 4.2 and an analysis of the selected features in Section 4.3. The subsequent sections describe how the curve data were used, starting with a reduction technique; SAX. After this step, the natural language processing method LDA is used which provides topics for each driver. These topics are then analysed to give an idea of how drivers behave in curves, which is expressed in tendencies.

4.1 Naturalistic dataset

All the data used originated from the first European large-scale Field Operational Test on Active Safety Systems, the EuroFOT project (EuroFOT, 2012). The goal of the project was to record data from day-to-day driving, with driving purpose and destinations decided by the driver. Hence, the driving takes place on roads with ordinary traffic and traffic situations. The vehicles used were the drivers' own, the vehicles the drivers are most familiar with. In this way a more natural driving is ensured. A total of 28 organizations with more than 1000 vehicles participated in this European wide operation. Sweden contributed with 100 Volvo V70 and Volvo XC70 cars and this data are used in this thesis. For most vehicles, there is more than one driver, although generally one of them drives a lot more than the rest. A total of over 200 drivers were recorded and over 100.000 trips are available for analysis.

Each vehicle was equipped with sensors and logging devices. Parameters including GPS, speed, longitudinal and lateral acceleration as well as sensor data on windshield wipers, lights and type of road, etc. were recorded and stored. Most trips have over 100 different features (some are derived from others) collected however a few trips are missing a few. Video of driver (face and feet) as well as the exterior of the vehicle were recorded for all trips. All recording of parameters, including video, was done unobtrusively as to make the drivers forget they are being observed. That the data collection took place under a long period of time, about a year, also contributed to make the driver behave naturally and not change behavior on account of being monitored. When simulators or instrumented vehicles are used a possible change of behavior needs to be taken into account. By using naturalistic driving data this factor can safely be ignored. The downside is that there are no operational controls and as a result the data is hard to analyze and compare. The amount of generated data is large which makes all processing and analyzing tasks more difficult.

There are some negative aspects to working with naturalistic driving data. The most prominent is linked to analysing the data. The routes where the vehicles drive cannot be influenced therefore comparisons and groupings of events are more difficult. When using simulators or instrumented vehicles, where the routes are predetermined, patterns in the data can easily be found. Another weakness of naturalistic driving data compared to the alternatives is the cost and time. Each vehicle in such a study needs to be fitted with equipment under the whole duration of the test, which should be a couple of months to ensure unbiased data. Furthermore, under this long period of time all collected data need to be taken care of somehow. As the amount of data is large it needs to periodically be saved to another location. This will preferably occur automatically without

requiring the driver to do anything.

In summary it can be said that using and collecting naturalistic driving data is bothersome, but the data is unbiased and represents real traffic situations. The need of observing real situations exceeds the difficulty involved; hence, using naturalistic driving data is preferred when possible.

4.2 Locating curves in naturalistic driving data

The definition of a curve is not well defined yet and exploring a formal definition of a curve is outside of the scope of the project. However, there exists a need to make an informal definition in order to extract the curves from the dataset. It is easy to find individual curves by examining the videos, but we use a large dataset and thus individually selecting curves is impractical. Hence, we try to define a curve, such that we from the complete set of all data can extract only curve sections and nothing else. At the same time we want to get different kinds of curves and preferably not misclassify other occurrences as curves. Turns are outside the scope of this study and therefore additional caution is taken to not classify these incorrectly.

To separate the sections with curves we used a systematic approach where we started by discarding all turns and then proceeding to take away sections with straight driving and lane changing. To confirm the validity of the utilized method and parameters some trips were randomly selected and the signals were viewed together with the video. In addition, special care was given to the largest and smallest curves found and selected when attempting the technique on the whole dataset. Both the smallest and the largest curves selected can be seen in Section 5.1.1.

A trip has been chosen to help illustrate the process of extracting start and end points of the curves. Graphs of this trip are shown after each major step taken to give better understanding of situations that can happen. We started this thesis by studying driving data from both urban and rural areas. However, other road users and vehicles in urban environment will greatly affect the driver's behaviour. E.g. a pedestrian can suddenly walk out on the road in front of the vehicle which requires the driver to suddenly brake. If this happens in a curve then during data analysis there is no way (except manually looking through all videos) to know what was the cause of the sudden braking and therefore it could potentially be attributed to aggressive driving. Therefore, we chose to take away these parts and focus on analysing data where these types of events are less common, i.e. rural areas. For the dataset used, a variable containing information if the current vehicle position is in an urban or rural area has already been calculated earlier. Hence, we simply used the supplied variable. The variable value is it is based on the vehicle's GPS position.

To find turns to be discarded the steering angle parameter was used. Turns were at first detected by finding any steering angles exceeding 130 degrees. The beginning and end of turns were determined by back- and forward-tracking in time until the angle of the steering wheel dropped below one degree. The heading parameter was tried, to locate turns but had comparatively more noise, and therefore the current method was determined to be superior for our needs.

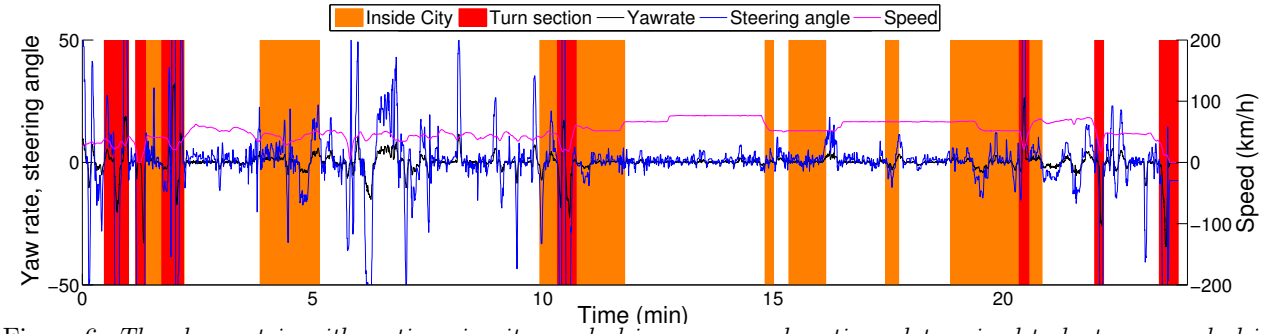


Figure 6: *The chosen trip with sections in city marked in orange and sections determined to be turns marked in red.*

The downside in the current approach is that some very sharp curves (especially in lower speeds) may be classified incorrectly as turns and discarded. There has been an earlier study (Sullivan et al., 2014) where one

of the steps was to locate turns in a naturalistic driving environment. They confirmed if it was an actual turn by using the state of Michigan roadway geometric data where they could get geographical coordinates for all turns. The feasibility of such an approach was investigated using data from Trafikverket in Sweden but later discarded. These kinds of databases only include larger intersections but in our case all types of turns should be removed. Hence, using this approach was determined to be unsuitable.

After the turns were isolated and discarded the next step was to separate curves from straight driving and lane changes. Curves where the speed was less than 30 km/h largely occur in traffic queues or inside cities. These curve sections are not analysed and are taken away in this stage. We do not take into account if the curve are simple or compound, see Figure 1 in Section 2.2.1 for an illustration of different types of curves. Reverse curves are regarded as two separate curves. If spiral transitions are present or not are similarly not accounted for. The initial research presented in Section 2.2.3 showed that all parameters that can potentially be used to find curves are uncertain in some degree. After experimentation the steering angle was chosen to be the main parameter used to find curves. Other possible parameters to find the initial curves include yaw rate, lateral acceleration and heading. A negative aspect of them all, including steering angle, is that start and end points of the curves can be different for different trips and drivers. They depend on where the driver started turning in the curve.

Curves were found using the same method as turns but here steering angle values above five degrees were considered as potential curves.

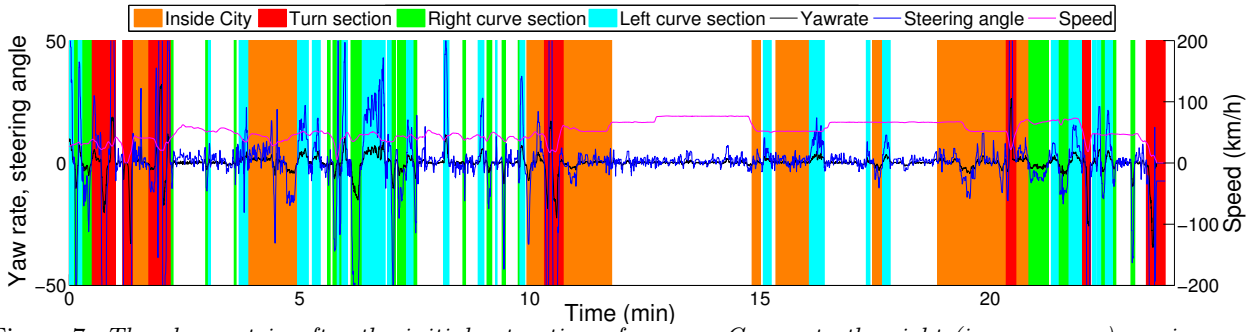


Figure 7: *The chosen trip after the initial extraction of curves. Curves to the right (inner curves) are in green and left (outer) curves are colored cyan.*

In mild curves the driving style is approximately the same as when driving straight and these are uninteresting for our purpose. Sharper steering angles are used at lower speed due to curve designs. Hence, different thresholds were applied to the steering angle for different speeds, see Table 1. To use road speed limit would be preferred but is not always available. The thresholds were mainly chosen by manual inspection of the videos. Here extra care was given to curves in the lower regions of each group with respect to both speed and steering angle. These are the mildest curves in each group.

To ensure the extraction obtain about as many inner (right) as outer (left) curves we had to adjust the required steering angle slightly depending on direction. As can be seen in Table 1 there is one degree difference between inner and outer curves.

The radius of a curve is related to steering angle, more specifically related to the necessary steering angle required to traverse the curve. Godthelp (1986) uses the following equation for approximating the required steering angle from the radius. However we use it to give an indication for the radius at different speeds and steering angles. The radii calculated here are not used in later analysis as other methods are investigated for this purpose.

$$\delta_s = GI\Big(1 + K\big(\frac{u}{3.6}\big)^2\Big) \cdot \frac{1}{R} \cdot \frac{180}{\pi}$$

where the variables used are

$\delta_s =$ steering-wheel angle	deg
$R =$ radius of the curve	m
$G =$ steering ratio	
$I =$ wheel base	m
$K =$ stability factor	s^2/m^2
$u =$ speed	km/h

Note that these values are approximate. The vehicles used during the data collection were Volvo V70 and Volvo XC70 cars and the values from these were used in the approximation. If a driver uses a higher than required steering angle to traverse a curve then the radius of his traveled path will be less than the value in the table. The path traveled can be thought of as a circle, when a higher angle is used the radius of the circle will be smaller.

Speed (km/h)	**Steering angle inner curves (degree)**	**Steering angle outer curves (degree)**	**Curve radius (m)**
30 - 40	24°	25°	120-130
40 - 50	20°	21°	150-165
50 - 60	16°	17°	205-225
60 - 70	12°	13°	295-330
70 - 80	10°	11°	390-430
80 - 90	9°	10°	475-530
90 - 100	8°	9°	590-655
100 - 110	7°	8°	745-825
110 - 120	6°	7°	955-1060
120 >	5°	6°	1250 >

Table 1: The steering angle at least required for curves to be accepted at different speeds and direction. The curve radius R is a rough calculation using the above formula. It describes how big the curve radius will approximately be for different speeds and steering wheel angles. These values are only here to illustrate what type of curves are used and will not be used in later analysis.

Yaw rate was initially successfully used to extract curve start and end points but were later replaced as it contained more noise compared to steering angle while still displaying the same characteristics. While the yaw rate is not used for extracting curves from the data the maximum yaw rate in a curve must be at least one degree/s. In addition to the extra condition on yaw rate, the heading must also change by at least five degrees from start to end of a curve.

The above method will, in addition to curves that we want, find some lane changes and some rare cases of road obstructions. To remove this unwanted data a few methods were used, at first, all identified curve sections with a length of under two seconds were removed.

For lane changes, a simple method was applied which checked if the direction indicator was used during the first second of an identified potential curve, such a curve section would be removed. Lane changes that are located in an actual curve are allowed and are not removed. This method will not remove all lane changes as the direction indicator is not always used (Boyce and Geller, 2002). Hence, an additional heading change test was included. If two adjacent potential curve sections (one right and one left) had a total heading change of less than five degrees, it could be a lane change and are removed.

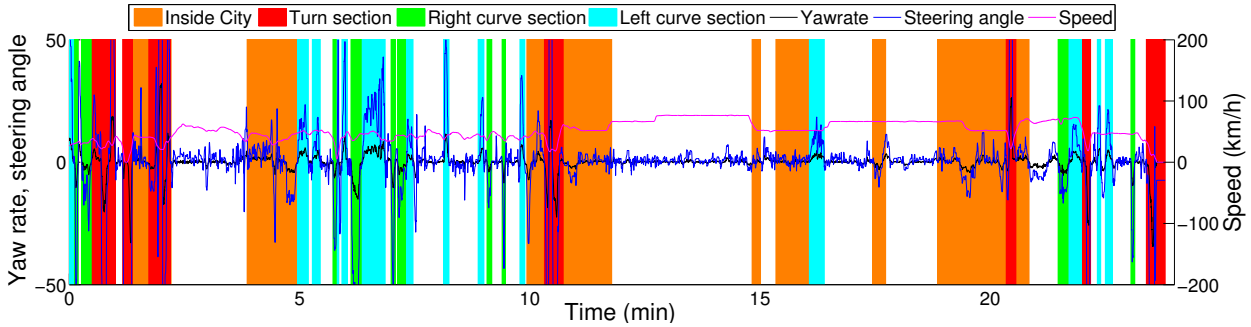


Figure 8: *The chosen trip after removal of curve sections with too low maximum yaw rate, too low heading change or if the direction indicator was used the first second of a section (indicating a lane change, not a curve).*

All remaining sections were determined to be curves and had 100 meters added to the beginning and to the end following the findings in Section 2.2.2 that previous curves affect the behaviour. The findings indicate that the point of deceleration could be as far as 100 meters from the beginning of the curve and therefore this distance was added to the data analysed. The same distance was chosen for when departing the curve. A subset was chosen for manual inspection to confirm the selection. Curves occurring within the first or last 100 meters in a trip are therefore discarded. The curves were split into three different categories; outer curves, inner curves and consecutive curves. Here consecutive curves are defined as the union of curves within 100 meters of each other, including the sections in between.

After all this some final steps are taken. The first on is to ensure that these new sections do not intersect with sections marked as urban or turn. If this were the case it could influence the later analysis. Curve sections where the vehicle was immobile within 100 meters before the curve were discarded as well. The reason being, that these are hard to compare with curves where the driver is up to speed from the beginning. Sections that have areas marked as roundabout are also removed. Not all roundabouts have been marked such in the data, therefore some roundabouts may be included in the final set of curves. All curves that have a radius shorter than 20 meters or longer than 1200 meters are likewise removed.

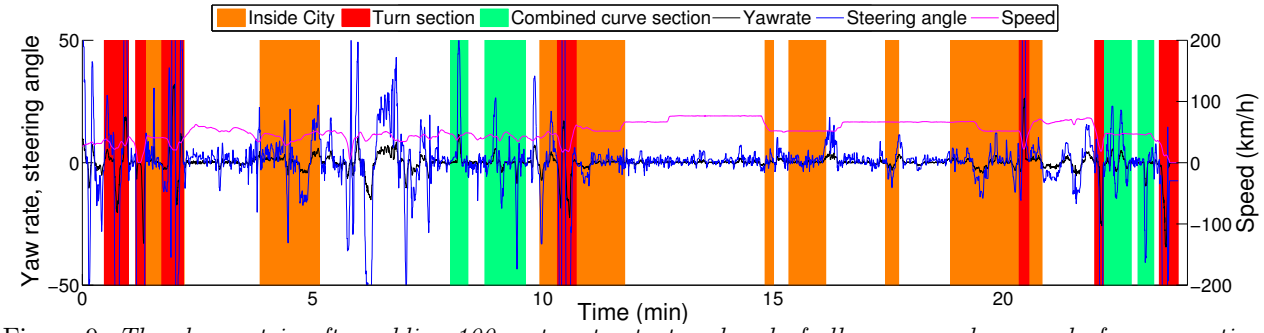


Figure 9: *The chosen trip after adding 100 meters to start and end of all curves and removal of curve sections not meeting the requirements above. This is the final collection of curves.*

4.3 Features selected for analysis

It is difficult to know the best set of features to use for the task of analysing driving styles. Selecting which variables are of interest is a deliberation between benefit and complexity. Constantinescu et al. (2010) identified a weakness with their study was the low number of features used on the other hand with more selected features the later analysis increase in complexity. Here a balance needed to be found. There are cases where some variables contain the same type of information, in practice only one of these needs to be picked. Below in Table 2 are the final set of features used. The selection was made by heuristics and reviewing previous studies done within the area of driving styles along with direct video observations. This section will explain the selection process as well as the modifications done to some of the features.

From the curve extraction above both inner and outer curves as well as consecutive curves were extracted. Only inner and outer curves are used in later analysis steps. This is to make sure the curves are comparable,

when curves are positioned very close to each other the driving style is not necessarily the same as for simple curves. For a simple curve drivers usually decelerate during the approach and accelerate afterwards, this is not necessarily true when two curves are next to each other. Hence, the decision to only use simple (and compound) curves was made.

Each separate curve will be divided into 12 separate sections to be able to make comparisons between different parts of the curves. One segment each for the first respectively last 100 meters and 10 for the actual curve. These 10 segments are all of equal length. The idea to segment the curve have been used extensively in literature, e.g. 10 segments were used in (Imberg and Palmberg, 2015).

Feature	**Unit**	**Modification**
Curve radius	m	-
Segment number	-	-
Direction	-	-
Speed	km/h	-
Over speed limit	km/h	-
Braking	bar	-
Longitudinal acceleration	m/s^2	Abs
Longitudinal jerk	m/s^3	Abs
Lateral acceleration	m/s^2	Abs
Lateral jerk	m/s^3	Abs
Steering angle rate	deg/s	Abs
Night	-	-
Rain	-	-

Table 2: The selected features used for analysis. An description of the features themselves as well as explanation of the modification are below.

Histograms of the features can be seen in Appendix A. Graphs of both the maximum value in each separate curve as well as the mean value from each segment are available. The definition of some of these features is not obvious or ambiguous and will be explained.

Curve radius
This feature have been used in nearly all previous studies to group curves into separate groups for comparisons. It can be seen as a combination of curve length and curve angle. The curve radius is not directly accessible from the data and needs to be calculated. There exists different techniques to estimate curve radius and three of these have been tested.

To determine which method is most accurate we used data from earlier work done by Imberg and Palmberg (2015), they used seven curves of different radii (110m, 120m, 290m, 300m, 385m, 410m and 500m) from the EuroFOT dataset. Their estimation was done with maps and should therefore be accurate. As the same dataset were used we arbitrarily picked 14 trips that passed by these curves (one in each direction and with different drivers). The data from the curves were used to estimate their radius. The method with the smallest total root-mean-square error when compared to the radii estimated by Imberg and Palmberg (2015) were used.

The first estimation method examined has been used by Othman et al. (2012). They used mean values over multiple trips to approximate the curve radius while we only have curves with data from one trip. More noise in the estimations is expected compared to using mean values. Othman et al. (2012) used the heading change and curve length for the estimation. The points used were not the estimated start and end of the curve but points a bit further in on the curve, only about 1/3 of each curve were used for the curve radius estimation. This is done to make sure the points actually are on the curve as the start and end are approximated. The formula used was

$$R = \frac{180^\circ \cdot L}{\pi \cdot \Delta\text{Heading}}$$

where R = radius of the curve, ΔHeading = change in heading between two points in the curve, L = distance the vehicle traveled between the same two points.

The second method used the same formula, but instead of selecting two points to use for estimation multiple points were used. In this way multiple estimations were done for each curve, the final curve radius was the mean of all these estimations. 11 points were used to partition the curve into equal sections and 10 of these were used for the estimation. A figure of an example curve can be seen in Figure 10. The points that were used together for the curve radius estimation are marked as A, B, C, D and E.

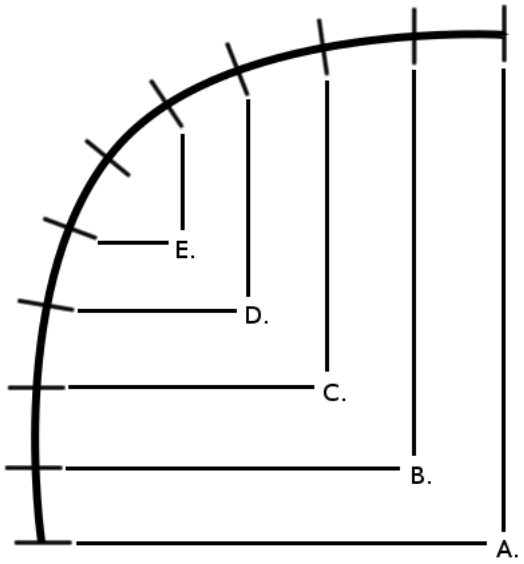


Figure 10: *The points used for curve radius estimation using the second method.*

The last technique tested was to approximate the curve as a simple circle and then calculate the radius from that. For the circle approximation the GPS coordinates of the curve were used. One possible variation that was not tried is to use a clothoid approximation instead.

In truth all these methods estimates the path radius, the estimated radius of the vehicle path - not the actual curve radius. There is a difference as drivers shift the vehicle laterally to flatten the curve slightly for a more comfortable drive, see Figure 11. In a later paper Othman et al. (2014) takes this into account following the values presented by (Bonneson et al., 2007) as seen in Table 3.

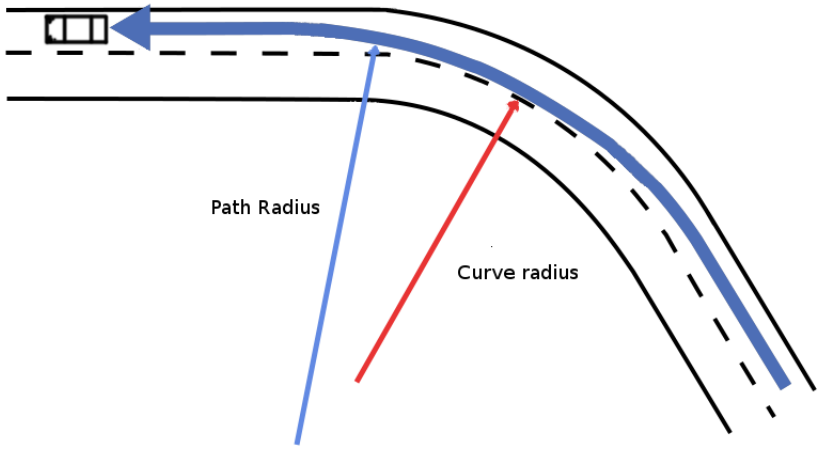


Figure 11: *Illustration of the difference between path radius and curve radius.*

ΔHeading (degree)	5°	10°	15°	20°	25°	30°	35°	40°	45°	50°	55°	60°
Increase in radius (m)	961	240	107	60	39	27	20	15	12	10	8	7

Table 3: Increase in curve radius due to a lateral shift in lane position (Bonneson et al., 2007; Othman et al., 2014).

Road geometry alike to Figure 11 have been used to calculate the increase in radius due to lateral shift within the lane. The curve radius (red arrow) is estimated by taking the path radius (blue arrow) and subtracting values as seen in Table 3 depending on the heading change, i.e. the deflection angle of the curve. This correction to the calculated curve radius could increase estimation accuracy.

Both which points in the curve to use for the approximation and if the correction due to lateral shift improves the reliability of the result were investigated for each method. Besides using the whole curve for the curve radii estimations, utilizing points 1/3, 1/4, 1/5 inside the curve from both sides were also tried.

Direction
As seen in the literature study, this feature is necessary due to the different behaviour expected when driving on the inner lane of a curve (turning right) as opposed to the outer lane of a curve (turning left). By using this feature we can take the absolute value of many features that have negative or positive values depending on direction of the vehicle, e.g. yaw rate and lateral acceleration. This can be done as the information about direction is still accounted for with the use of this feature.

Over speed limit
This feature is calculated from current speed and the road speed limit. Regrettably, the speed limit is not always available and can only be used in approximately 60% of the extracted curves. When the speed limit is unavailable, it is assumed that the driver is not speeding.

Longitudinal acceleration
This is the only feature that have very different meaning depending on the sign. For positive values the vehicle is accelerating and negative values decelerating. Here the direction feature is of no use therefore we cannot use the absolute value of this specific feature. An issue with deceleration is that it can be braking, simply letting friction slow the vehicle down or an uphill slope. However, the braking feature is used to determine if a deceleration is initiated by the driver or not. The same is true for acceleration; it can be both the environment that affects the vehicle or the driver.

Steering angle rate
All mentions of steering angle refer to the angle of the steering wheel. Steering angle rate is the angular acceleration and is based on steering angle and time. In addition to steering angle rate both steering angle and steering angle jerk were considered as features. Steering angle jerk is very noisy even during straight driving and was therefore not selected. Steering angle was already used in the curve selection and have a close resemblance to the lateral acceleration feature, hence, the steering angle rate was used.

Rain
This feature is set to 1 if the wipers are active and 0 if they are not. No information of the actual road conditions is available; hence, the road could be wet or slippery even if the wipers are off.

Yaw rate
Yaw rate is the angular velocity of the vehicle's rotation around its vertical axis. In other words, how fast the vehicle is turning to the left or right. Yaw rate is directly related to lateral acceleration when a vehicle is turning at a constant speed around a constant radius, lateral acceleration = tangential speed · yaw rate. In later results the lateral acceleration and yaw rate were extremely similar. The relation mentioned and these results lead to the removal of yaw rate as a feature.

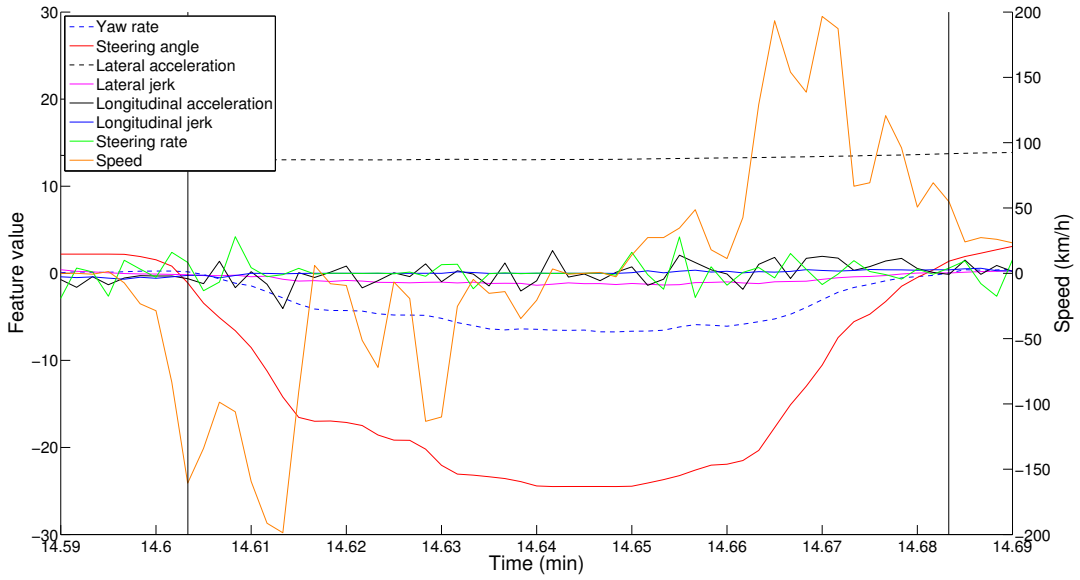


Figure 12: *Example of how the different features can look in a curve.*

4.4 Reduction of data dimensions

Symbolic Aggregate approXimation (SAX) was selected for data reduction and is described in Section 3.2. SAX is one of the better methods for data reduction because it reduces dimension while the characteristics of the data remains unchanged. It has been successfully applied to naturalistic driving data before as mentioned in Section 2.3.2.

Lin et al. (2002) made MATLAB code available for SAX which initially showed some promise. However, their normalization makes it so a change in speed from 40km/h to 50 km/h could show the same results (letters) as a change in speed from 100 km/h to 110 km/h. In other words, 40 km/h and 100 km/h could both be represented by an a.

Thus we implemented SAX ourselves in a way specifically suited to our needs. To implement the piecewise aggregate approximation we divided a curve of length L meters into 12 segments. Then the mean of each value would be calculated giving us 12 segments with one value for each feature in each segment. The first and last segments are the tangents (the 100 meters added to the beginning and end of the curve). The ten middle segments are of equal length; $(L - 200)/10$.

The complete list of features together with their breakpoints are shown in Table 4. The breakpoints were determined by looking at histograms, which can be found in Appendix A. A similar idea was used by McLaurin et al. (2014) where they let the different speed limits be the breakpoints rather than using a normal distribution to calculate the breakpoints. With values sorted from low to high, 40% of the values was decided to be the first of our breakpoints, hence, values within the first 40% were considered close to normal driving with the symbolic representation of a. The other breakpoints where after 60%, 80% and 90% of the data. All of the breakpoints would then be between each percentage (0-40, 40-60, 60-80, 80-90, 90-100) with a symbolic representation of a, b, c, d and e respectively.

The validity of these breakpoints were manually tested to ensure that changes in data are expressed. Near maximum values in any of the features should be more indicative of driving which divergent behaviour. Hence, the breakpoints are closer to each other for higher values. There were some exceptions however, all variables with only two symbols can only have a and b, such as direction, daylight and rain. The thresholds for curve radius were selected such that the groups should signify different kinds of curves, see Section 2.2.2.

Normally the words are symbols expressing different values for the same feature in time. However, we use a slightly different method, where a word consists of several features collected over the same time period. Hence, each separate word consists of 13 letters where each letter represents one specific feature. In Figure 13 an example of a word is shown. The features are always in the same order for each word and for each driver.

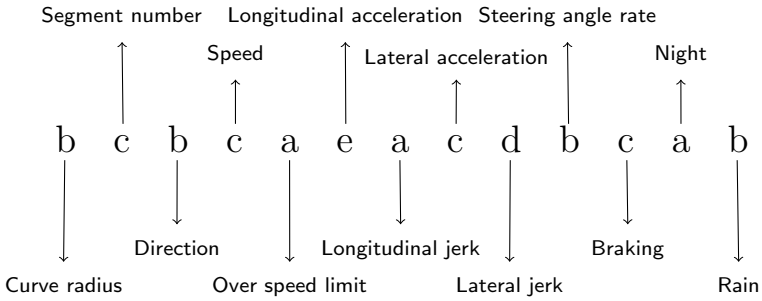


Figure 13: *An example word after performing SAX on the data with explanations of what each letter represents.*

The longitudinal acceleration is a bit different to other features, since it has both positive and negative values. An attempt was made to use more breakpoints, however, the result was significantly harder to interpret. Hence, the same amount of breakpoints are used for longitudinal acceleration. Starting from 0 (the middle) breakpoints were selected to encompass 40% and 80% of all data. In this way the longitudinal acceleration uses 5 letters for the symbolic representation.

The breakpoints for speed are selected to reflect different speed limits, except for the last breakpoint which is 125 km/h. Values over 125 km/h could be indicative for speeding for those curves where no data for speed limit are available. There are only a small amount of segments where braking was observed in our dataset (3.4% of the segments). Therefore, only information on whether braking was performed or not was used instead of measuring how hard braking was done. Below in Table 4 are the breakpoints used for all features.

Curve radius

Value range	Symbol
20 - 200	a
200 - 400	b
400 - 600	c
600 - 800	d
800 - 1200	e

Direction

Value range	Symbol
Inner	a
Outer	b

Section in curve

Value range	Curve location	Symbol
1	Entry tangent	a
2 - 4	Beginning of curve	b
5 - 8	Middle of curve	c
9 - 11	End of curve	d
12	Exit tangent	e

Steering angle rate

Value range	Symbol
0 - 2.2	a
2.2 - 3.4	b
3.4 - 5.5	c
5.5 - 8	d
> 8	e

Vehicle speed

Value range	Symbol
0 - 60	a
60 - 80	b
80 - 100	c
100 - 125	d
> 125	e

Over speed limit

Value range	Symbol
< 5	a
5 - 10	b
> 10	c

Longitudinal acceleration

Value range	Symbol
< -0.26	a
-0.26 - -0.05	b
-0.05 - 0.05	c
0.05 - 0.26	d
> 0.26	e

Lateral acceleration

Value range	Symbol
0 - 0.43	a
0.43 - 0.69	b
0.69 - 1.0	c
1.05 - 1.4	d
> 1.4	e

Longitudinal jerk

Value range	Symbol
0 - 0.71	a
0.71 - 1.02	b
1.02 - 1.405	c
1.405 - 1.75	d
> 1.75	e

Lateral jerk

Value range	Symbol
0 - 0.62	a
0.62 - 0.77	b
0.77 - 1.03	c
1.03 - 1.31	d
> 1.31	e

Braking

Value range	Symbol
False	a
True	b

Wiper activity

Value range	Symbol
Off	a
On	b

Daylight

Value range	Symbol
Day	a
Night	b

Table 4: Symbolic representations of all variables used during analysis. The value ranges for Curve radius, Steering angle rate, Vehicle speed, Longitudinal and Lateral acceleration, Longitudinal and Lateral jerk are left inclusive and right exclusive.

4.5 Topic creation

This section contains how we used Latent Dirichlet Allocation (LDA) on the output from SAX and how we made a visual representation of the data in order to get a good overview. Following this is how the topics were used to determine driving styles of drivers. For the theoretical aspects of LDA and an explanation of the terminology used it is recommended to read Section 3.3.

LDA was chosen partly because McLaurin et al. (2014) used it together with SAX on naturalistic driving data and we were interested in seeing if it could be used successfully on our data. Overall, LDA is a well-known and robust model which offers us a means to classify drivers in an unsupervised manner.

4.5.1 Implementation

the toolbox created by Griffiths and Steyvers (2004) was used. The first step to implement the toolbox consisted of adopting the output of SAX into something that could be used as input into LDA. Each driver's words from SAX are interpreted as separate documents, where each word is a curve segment as described in the previous section. Before using the toolbox a vocabulary had to be built; this was done by going through all of the documents and gathering all unique words. In addition to the vocabulary two other inputs were required. A list of all words and a list of which documents consist of which word(s). Finally, the toolbox would generate topics; for each topic there is a list of words (curve segments) and their respective probability of appearing within that topic. With each topic also came the knowledge of how many percent of the corpus that was expressed by that specific topic.

Blei et al. (2003) used a 100 topic LDA as an example in their paper. However, they mention that the number of topics is dependent on the size of the corpus. Since we could not with certainty know how many topics we needed, we experimented with topics ranging from 10 to 1000. In order for LDA to find a stable state, LDA

was run for 5000 iterations. Ten topics, were too few to learn anything useful. On the other hand, 1000 topics were too many. After testing different amounts of topics and manually inspecting the topic mixture we chose 200 topics. This number was selected to allow one driver to be expressed by multiple topics and so that enough topics existed to express a wide array of different behaviours. More topics would make later analysis insipid and the topics would not represent the data in an intuitive way. Fewer topics would be possible, however, more uncommon behaviour might not be found because only the most probable words will exist in a topic.

Only drivers with more than 25 separate curves were included in the analysis. There were 20 drivers with very few trips, who therefore had few curves. Since nearly no data were available for these drivers they were removed. Hence, the analysis was performed on a total of 141 separate drivers.

The hyperparameters α and β were chosen to be 0.05 and 0.01. They were chosen after testing different values for the hyperparameters and comparing the output. Similarly to how we chose the amount of topics, a higher β gives a higher number of words per topic, making it more difficult to understand the meaning of each topic. A low α makes it so that a driver's individual styles gained more importance by allowing fewer topics to describe each driver.

4.5.2 Making sense of the topics

200 topics with their respective word distributions are difficult to analyze. Hence, we set out to make a visual representation of the topics, in order to get a better overview and be able to see patterns more easily. This is done by taking a weighted average (after word frequency in a topic) of the top words above the threshold and then assigning it a colour from 0-255 where a would mean 0 (blue) and the last letter would mean 255 (dark red) in each feature. This weighted average gives us some idea of what type of curve and driver behaviour each topic could potentially describe. In general, blue means very low values and red means high values, except for longitudinal acceleration. Here, blue means the vehicle decelerates, red accelerates and green represents constant speed.

Each topic contains many words and each word has a weight, which expresses the probability of that word appearing within that topic. This probability varies from around 2% to close to 0. We chose to only include words which have a higher probability than 0.01% because we wanted a topic to accurately express the most prominent words (curve segments) in that topic. 0.01% is enough to allow the most important words within each topic to be considered; however, a few topics only have unlikely words. Hence, we decided that even topics with unlikely words should contain at least 8 words.

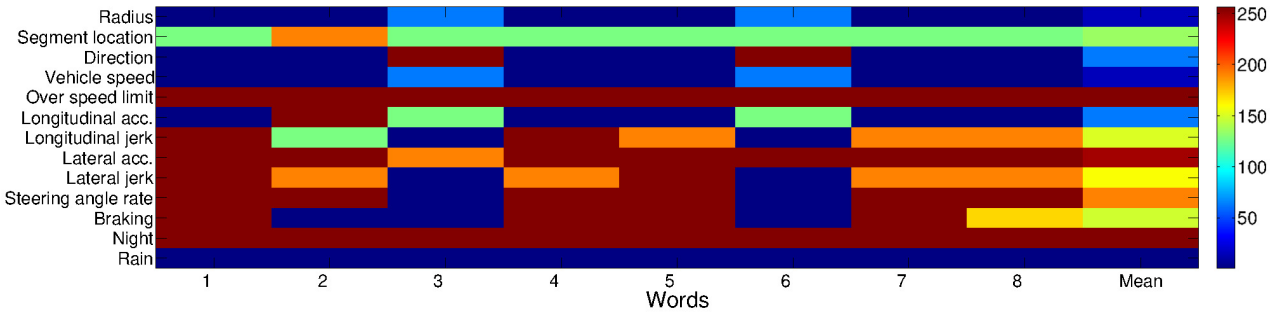


Figure 14: *An example of how eight words get averaged into a mean. The words are sorted in order of probability from left to right, with the mean furthest to the right.*

A consequence of this averaging is that the data are approximated a second time, and in contrast to SAX there is no guarantee that the values obtained will accurately reflect the old values. An example of this is the over speed limit feature which has three letters, a,b and c. If these are assigned the values 1, 2 and 3, respectively, then the sequence *aacc* and *bbbb* will have the same average value. Hence, we introduced a weighted standard deviation measure (weighted with regards to the word probability within a topic) which is shown in the figure as white dots if the standard deviation is small and they get gradually darker until black if the standard deviation is high, indicating that within a topic the values vary considerably.

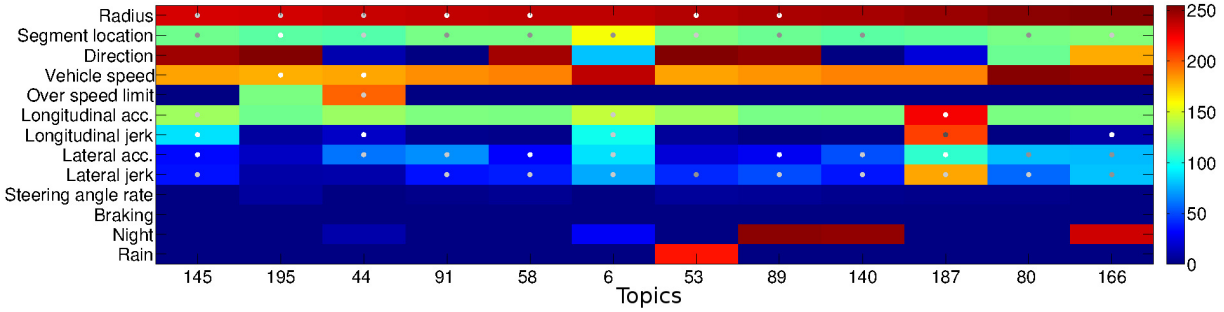


Figure 15: *Visual representation of 12 topics with large radii. Colour represents the average of the top words for each topic and the dot represents the standard deviation. No dot means no variation, a white dot represents little variation and darker dots represents higher variations.*

4.5.3 Determining tendencies

The next step was to determine what the different topics express. LDA only looks through the documents (drivers) to find similarities inside and between the documents to create the topics. Hence, a thorough analysis of the results is required. We decided to divide the analysis into two main parts. One part focuses on environmental and traffic related features such as curve radius, weather and time of day, whilst the other part focuses on driver related features such as speed, braking and lateral acceleration. We call the first part "tendencies related to driving context", and the second part "tendencies related to driver characteristics".

Different behaviours, in both parts, were analysed by sorting each feature in increasing order. The extreme values are those which are the most indicative for the tendencies related to driver characteristics; hence, these values are in focus.

Tendencies related to driving context
Curve radius
Approach and departure
Curve direction
Night
Rain

Table 5: The tendencies related to driving context.

Tendencies related to driver characteristics	**Features used**
Speeding	Over speed limit
Braking	Braking
Jerky curve handling	Longitudinal jerk
	Lateral jerk
	Steering angle rate
Rough curve handling	Lateral acceleration

Table 6: The tendencies related to driver characteristics and the features they contain.

The tendencies related to driver characteristics are highly related to some of the specific driving styles (under the longitudinal control and lateral control subcategories) suggested by Sagberg et al. (2015). The tendencies related to driving context do not directly indicate a driver's behaviour in curves, and were only used to see broad trends. For example, by analysing the topics mostly related to inner and outer curves we could see differences which indicate different behaviour. The curve radius analysis was divided into groups depending on the radius length. The same groups that were used for breakpoints were also used here. Direction was divided depending on curve radius since different behaviour can be seen for different radii. The other features were not strictly dependent on curve radius and were analysed separately.

Each of the tendencies related to driver characteristics were built up by topics that describe that particular tendency well. Values of the features within topics differ, and the topics have to contribute to the tendencies in an objective way. For each feature, the topics were divided into four levels depending on the values of that feature within the topics.

The four levels were high, moderate, low and no effect. The threshold used for deciding which topics belong to which group were derived from the average value of the features within a topic. For example, if a, b, c, d, e are represented by 1 to 5 then average values above 4.5 would be in the high group, above 3.5 in the moderate group and above 2.5 in the low group. The higher values indicate that there exist features with values above the 60th, 80th and 90th percentile of the data for c,d and e respectively.

The first of the four tendencies investigated was the tendency for Speeding, which was based on the amount of speeding (i.e. driving speed above the speed limit). The feature used is over speed limit, which is divided into no speeding, above 5 km/h and above 10 km/h. How common speeding is within a certain topic and how much as indicated by the over speed limit feature, are combined in this tendency.

The second tendency is Braking which tells us that the driver is driving too fast when entering a curve and needs to slow down. The small amount of braking in the dataset also indicates that normally, drivers do not brake in curves.

The third tendency is Jerky curve handling which contain the features; lateral jerk, longitudinal jerk and steering angle rate. These are all features which indicate unusually sharp or aggressive curve handling which could be from turning late in the curve or not slowing down appropriately. Since some changes in these features are expected in all curves, the thresholds for the different levels are focused at the higher end of the range for each feature. This was in order to capture more aggressive curve handling. Three separate features are used for this tendency. Hence, topics that contain high values in more than one of these features were counted multiple times, with appropriate weights depending on levels.

The fourth tendency is Rough curve handling which is based on lateral acceleration. The tendency indicates how strongly drivers are driving into and through the curve, whereas Jerky curve handling is more about sudden changes (jerky driving) within the curve.

The topics were converted into the four tendencies for each driver to create a driving style profile. Since a low level topic is merely indicative of driving slightly faster than average in the Speeding tendency whereas the high level is indicative of speeding a majority of the time. We wanted to be able to differ between the different

levels. Hence, the different levels were assigned a weight of 1, 2 and 3 for low, moderate and high respectively. The remaining topics not in any of these groups have a weight of 0.

A document (driver) contains a list of how many appearances there are of each topic, as seen in Figure 16. Thus, we could easily calculate a topic distribution for each document. The topic probabilities were multiplied with their corresponding weight. Thus, for each document (driver) d we received the following:

$$T_i = \sum_{t=1}^{200} r(d,t) \cdot w_{ti}$$

Where T_i is a tendency and $i = 1, .., 4$. $r(d,t)$ is the relation between a topic t and the driver d. w_{ti} contains the weight for topic t as seen from tendency i as described above.

After the tendencies have been calculated for each driver, we normalized over each tendency so that they were all in a range from $0 - 1$. This allowed for comparison of the different drivers and different tendencies. The tendency score each driver received were only based on data from this pool of drivers. For example, the driver with score 1 in Speeding has the highest tendency to speed among the analysed drivers. The score in Speeding for other drivers is based on how much they speed compared to the driver with score 1. A driver with Speeding score of 0.5 therefore means this driver tend to speed half as much as the driver with score 1.

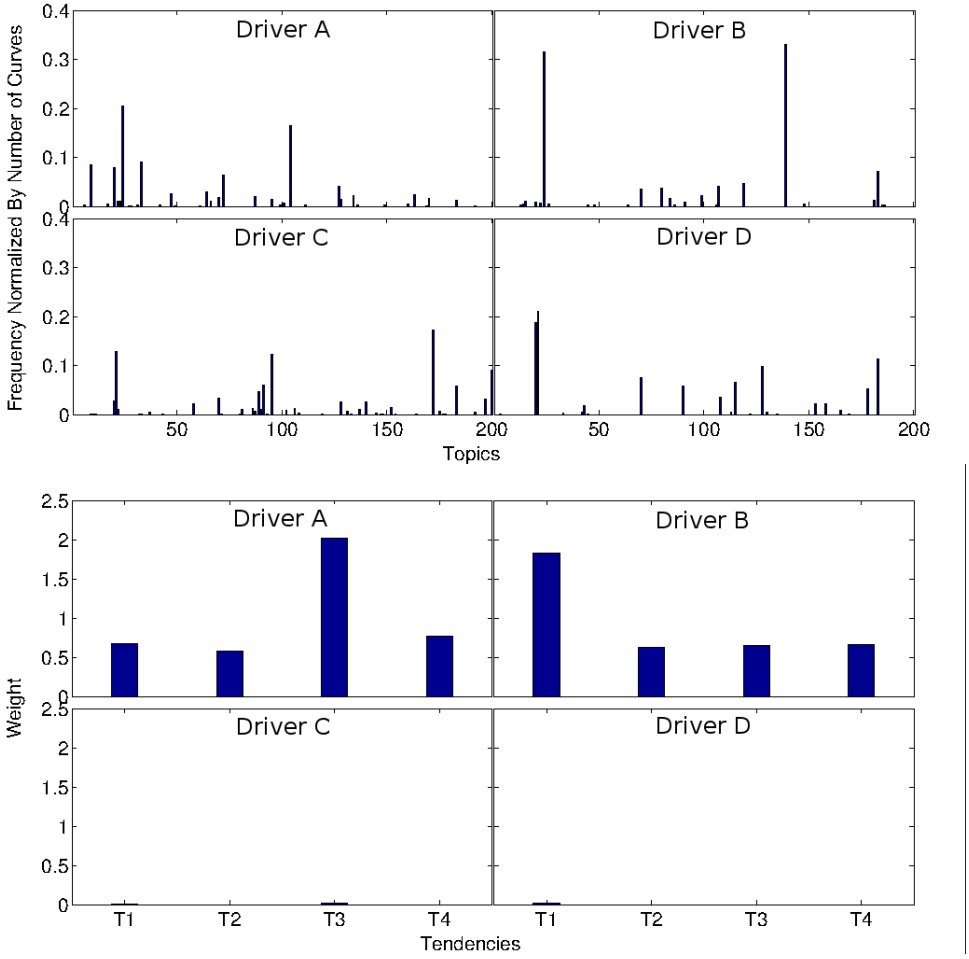


Figure 16: *Example of four different drivers with very different tendencies. At the top are their topic distributions over the 200 topics. The calculated tendencies are shown in the lower figure. The values shown are before the normalization.*

5 Results

In this chapter, the results from the analysis are presented in two parts. The first part describes the number and type of curves that were extracted from the dataset. Results from the investigation of curve radius estimation methods are also presented.

The second part describes the results of using LDA. First the results from analysing the tendencies related to driving context are presented. Following this are the results from the analysis of the tendencies related to driver characteristics, which are presented together with drivers with a high level of tendencies and drivers with low levels.

5.1 Curves

5.1.1 Curve extraction

Following the method in Section 4.2 a total of over 100.000 curves were extracted from the dataset. To further understand what type of curves that were detected by the method, the curves were grouped according to their radius as shown in Table 7 below. Curves classed as consecutive curves were also extracted, see Table 8, but were not used. The curves used for analytical purposes were outer and inner curves with radius of less or equal to 1200 and above 20 meters, i.e. 74020 curves.

Curve radius (m)	**Inner curves**	**Outer curves**	**Total**
20 - 200	12500	8302	20802
200 - 400	12462	11663	24125
400 - 600	6188	8619	14807
600 - 800	2698	3550	6248
800 - 1200	3477	4561	8038
1200 >	2428	2749	5177
Total extracted	39753	39444	79197
Total used	37325	36695	74020

Table 7: Total number of curves extracted from the dataset and number used for further analysis.

Number of bends	**Consecutive curves**
2	17875
3	6240
4	1940
5	841
5 >	900
Total	27796

Table 8: Total number of consecutive curves extracted from the dataset.

The number of inner and outer curves which can be seen in Table 7 are about the same. However, there are more outer small radius curves than inner small radius curves; consequently there are more inner large radius curves. In Figure 17, images of the smallest as well as the largest curves can be viewed together with a graph depicting how some selected features change along the curve. As can be observed from the graphs and images these are undoubtedly two relevant curves that belong to the dataset. Together with the arbitrary sampling of other curves we can say with reasonable certainty that our dataset only contains valid curves.

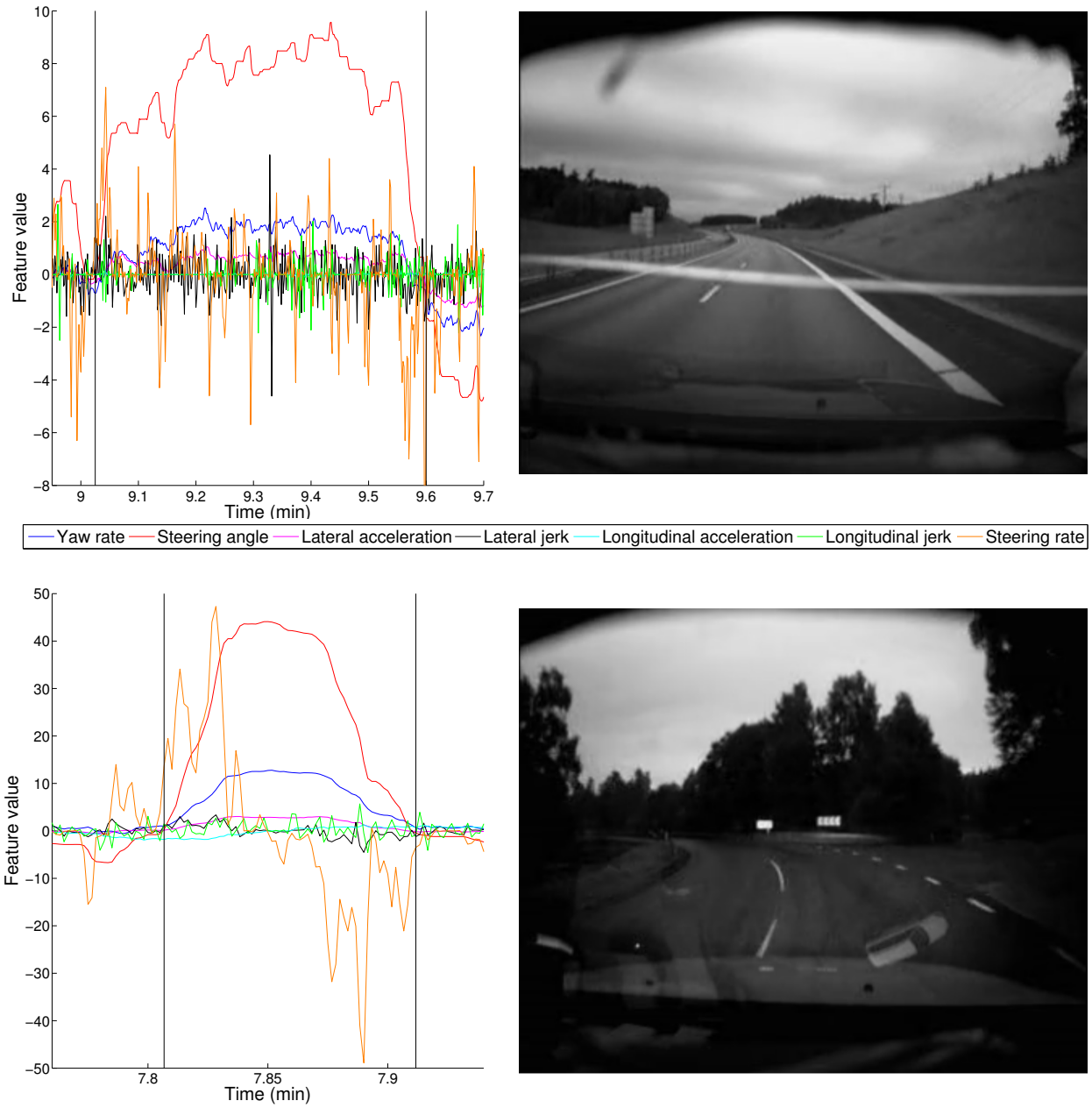


Figure 17: *The largest and smallest curves extracted. At the top row is an image of the largest curve in the set with a curve radius of 1200 meters together with a graph of how different features changes in the curve. The bottom row depicts the smallest curve with a curve radius of 20 meters along with a similar graph. The black lines indicate where the curves start and end.*

5.1.2 Curve radius estimation

As mentioned in Section 4.3, curve radius is commonly utilized in the literature as the feature used for grouping curves into different categories. This feature is not directly accessible from the data and needs to be estimated. The method from Othman et al. (2012) using heading change combined with curve length was selected for further investigation as well as two others. The second method used the same equation but segmented the curve into multiple parts and used a mean value and the last technique tried was to estimate the curve as a circle. In Table 9, the final Root-Mean-Square Error (RMSE) of each tried method can be seen.

Parameters	ΔHeading	Segmented ΔHeading	Circle estimation
Whole curve	197	277	252
Whole curve with corr.	175	**266**	**243**
1/3 inside	255	1536	252
1/3 inside with corr.	222	1480	337
1/4 inside	170	805	268
1/4 inside with corr.	**128**	786	255
1/5 inside	225	647	264
1/5 inside with corr.	131	621	259

Table 9: Root mean square error (RMSE) in meters for the three different techniques tried to estimate curve radius. Utilizing the whole curve as well as points at relative distances inside the curves (depending on the curve length) are shown. The RMSE is the estimated curve radii compared to the estimation done by Imberg and Palmberg (2015). The correction is due to lateral shift as discussed in Section 4.3. When the correction is used, a value is subtracted from the estimation depending on the heading change; see Table 3. The best value for each method is shown in bold.

Table 9 shows the most suitable technique. The lowest RMSE is 128 and is achieved when using the first method, where the heading change and curve length are used to calculate the curve radii. The best results were obtained when only a part of the curve was used. In our case, the lowest error was achieved when an approximation was done when choosing a starting point 1/4 of the way into the curve and the ending point at 3/4 of the curve; thereby only using half of the curve to approximate the curve radius. The correction for lateral shift seems to have reduced the error in nearly all cases and were therefore used as well.

Figure 18 shows the relation between curve lengths and their estimated radii using the selected method. Only curves with a curve radius of less than 1200 meters are shown. As can be observed in the plot, when the estimated curve radius is low then the curve length is usually low as well. For higher curve radii values there is a higher spread of the lengths of the curves. This is to be expected since the heading change is used in combination with the length.

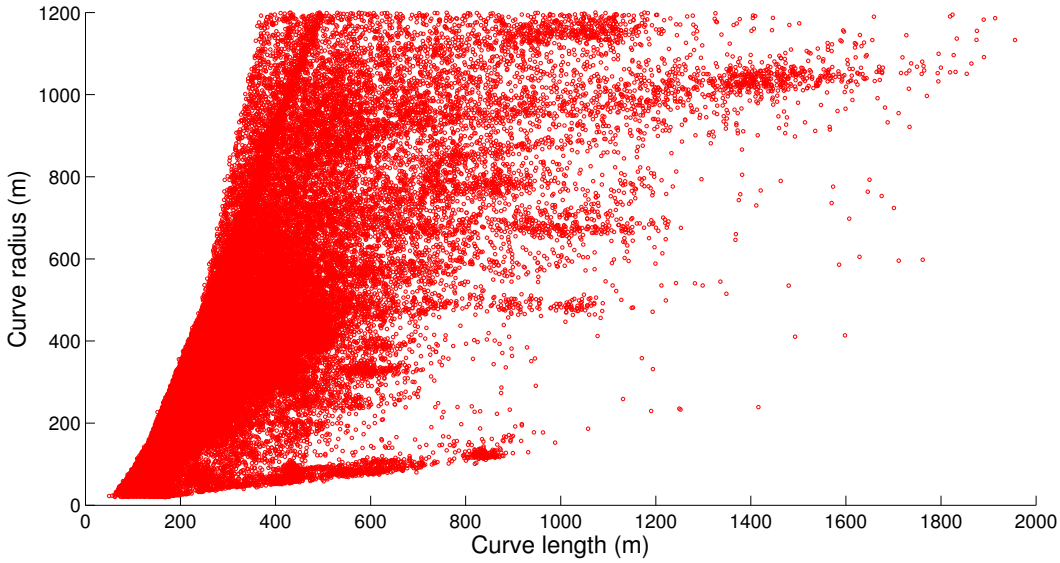


Figure 18: *Relation between curve lengths and curve radii.*

5.2 Results from LDA

After removing the drivers with too few curves, 141 drivers remained. Hence, 141 documents were used for LDA to create 200 topics. The vocabulary contained a total of 242877 unique words. The three topics with highest probability are shown in Table 10. The topics had a probability of 0.04757, 0.04450 and 0.04367, respectively.

Topic 70	**p(w|t)**	**Topic 20**	**p(w|t)**	**Topic 183**	**p(w|t)**
BCABACACAAABA	0.01788	BBBBACAAAAABA	0.01481	CCACACACAAABA	0.00957
BCBBACACAAABA	0.01057	BABBACAAAAABA	0.01313	CBBCACAAAAABA	0.00918
BAABACAAAAABA	0.00825	BEBBACAAAAABA	0.01130	CCBCACABAAABA	0.00885
BCABACABAAABA	0.00780	BDBBACAAAAABA	0.01120	CCBCACACAAABA	0.00796
BBABACAAAAABA	0.00656	CBBBACAAAAABA	0.01061	CDBCACAAAAABA	0.00782
BBABACACAAABA	0.00541	BCBBACABAAABA	0.01033	CACCACAAAAABA	0.00759
BCABACACABABA	0.00538	BBBBACAAACABA	0.00814	CABCACAAAAABA	0.00718
BBABACAAACABA	0.00504	BBBBACAAABABA	0.00759	CEBCACAAAAABA	0.00699
BDABACACAAABA	0.00473	CCBBACABAAABA	0.00752	CCACACADAAABA	0.00568
BBABACABAAABA	0.00461	BCBBACACAAABA	0.00710	CDACACACAAABA	0.00499

Table 10: The three most frequent topics in the corpus together with the top ten words and their respective probability of appearing in that topic, p(w|t).

In all figures below blue colour represents low values and red colour high values. The only exception is longitudinal acceleration where blue means deceleration, red means acceleration and green represents constant speed. For features that do not have any high or low values, e.g. rain, an explanation is provided under the figure.

5.2.1 Tendencies related to driving context

Curve radius

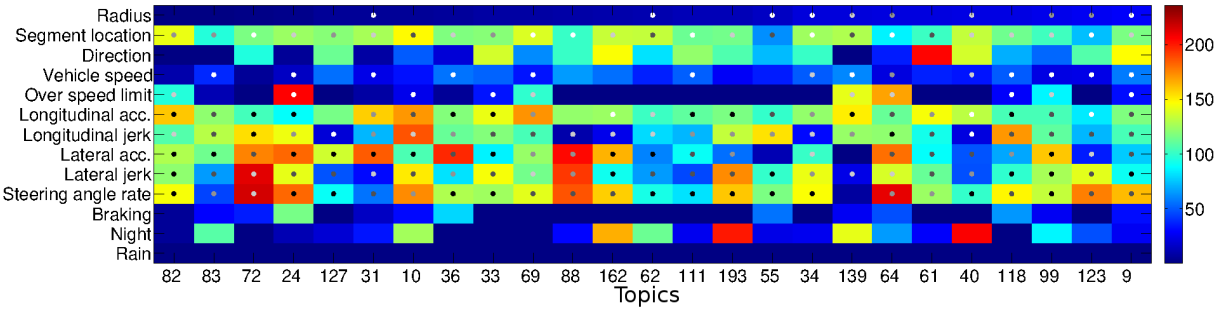


Figure 19: *The topics containing curves with small curve radius, between 20 and 200 meters.*

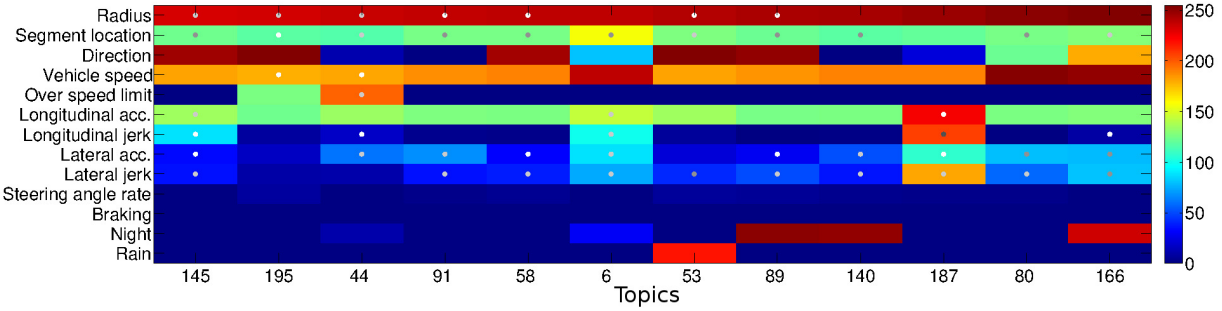


Figure 20: *The topics containing curves with curve radius between 800 and 1200 meters.*

The topics were divided into five groups based on the radius of the curves represented in each topic. The curve radii in each group are 20-200, 200-400, 400-600, 600-800 and 800-1200 meters. The groups with the smallest and largest curve radii can be seen above in Figures 19 and 20, the rest are in Appendix B. The same trends seen when comparing the smallest and largest curves are mirrored in the other groups.

As expected, the speed for small radius curves is low while the speed for large radius curves is high. This is a consequence of the curve design; high speed roads have curves with larger curve radius compared to low speed roads. In Figure 19, there are more inner curves (blue colour) than outer curves (red colour) due to more small radius curves being extracted for inner curves, as explained in Section 5.1.1.

In general, for large radius curves the speed is more constant, which is easily seen in longitudinal acceleration and jerk features. An exception here is topic 187, which has higher acceleration and jerk. This topic represents large curves with high speed, which suggests the curves are placed on the highway. Additionally, there are mostly inner curves with high longitudinal acceleration and jerk values, an indication that this could be ramps leading onto the highway. Steering angle rates are low and at the same time a rapid longitudinal acceleration is possible, indicating that the curves are mild. The longitudinal acceleration is positive, suggesting that only curves that are entering the highway are accounted for in this topic, not exiting curves. As the vehicle is rapidly accelerating the lateral acceleration and jerk are somewhat higher.

For smaller curves, there is much more variation in the longitudinal features (acceleration and jerk), where both acceleration and deceleration are present together with higher levels of longitudinal jerk. In addition to the longitudinal features, both lateral acceleration and jerk are present to a great extent for small radius curves. There is also a higher degree of braking taking place when the curve radius is small. All these features are higher for small radius curves due to the difference in geometry. When the curve is sharp, the lateral acceleration and jerk will be higher than for a mild curve, even if the speed is identical. When approaching a curve with smaller radius, there is a higher need to decelerate and to accelerate up to speed afterwards; therefore more longitudinal acceleration variations and braking are present for small radius curves. For larger curves there is enough room to allow the vehicle to slow down smoother and braking is not required as much.

Of the four topics that contain the most speeding (high level of speeding), three of them have large curve radii. However, the topics with a moderate amount of speeding are nearly exclusively in curves with small or very small curve radii. This suggests that some drivers drive at a speed of more than 10 km/h above the speed limit when the curves are large, i.e. on highways and larger roads. When the roads have lower speed limits (and therefore have curves with smaller curve radius), there are still drivers who drive too fast, but to a lesser degree. Note that speeding on lower speed roads is riskier than on highways as can be seen in e.g. the review done by (Aarts and Van Schagen, 2006).

Approach and departure

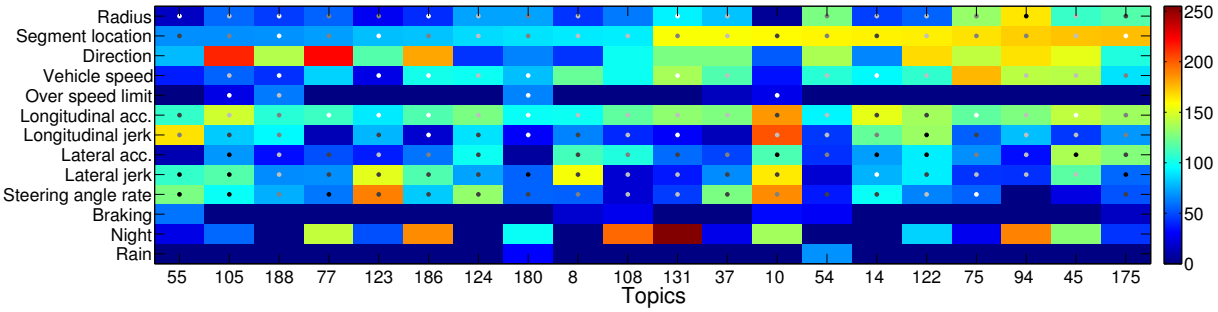


Figure 21: *Topics related to the approach and exit segments are shown here. To the right are topics with high relation to early curve segments in blue colour (i.e. approach segments) and to the left topics related to late segments in red colour (i.e. departure segments).*

There are no topics representing exclusively the approach or departure tangents (the topics do not have strong colours), but some trends can be observed for the approach/beginning and departure/exit segments. For the topics related to the approach and beginning of curves, there are only curves with smaller radius. This suggests more variation in the driver specific features when approaching a small curve compared to approaching a large curve. In a large curve the feature values during the approach and inside of the curve are more similar and do not change much.

For the topics with exit segments there are still mostly curves with smaller radius but here more variation can be seen. This trend suggests that the behaviour of exiting small and large curves is more similar compared to approaching small and large curves. As for the topics related to approach, mainly small radius curves are present. This suggests that driver specific feature values in large curves are similar when comparing inside the curve and exit segments.

For the topics with approaching segments, there are more negative (blue) values for longitudinal acceleration. This suggests that drivers lower their speed to a speed more suitable for the curve during the approach. For topics with exit segments, the longitudinal acceleration is instead mostly positive (red and yellow) or constant (green). For the curves with small radius, the longitudinal acceleration is higher which means that the need to accelerate to an appropriate speed again is especially true for smaller curves. When the curve radius is larger, the speed is kept more constant.

Topic 54 is an exception with negative longitudinal acceleration in an exit segment. A more careful analysis of the topic words suggests that some of the curves contained in the topic could be in uphill slopes. There are also separate words containing rain or braking. As can be seen below, when it is raining drivers tend to drive more carefully. The combination of these three kinds of occurrences could explain why the topic looks like it does, with low driver characteristic values and deceleration in later parts of curves. That these separate occurrences are grouped together suggests some commonalities between them, for topic 54 this appear to be careful driving, however, due to different reasons. In addition, the drivers with high relation with this topic obtained low values for most tendencies, indicating that they are more careful than many other drivers analysed.

Direction

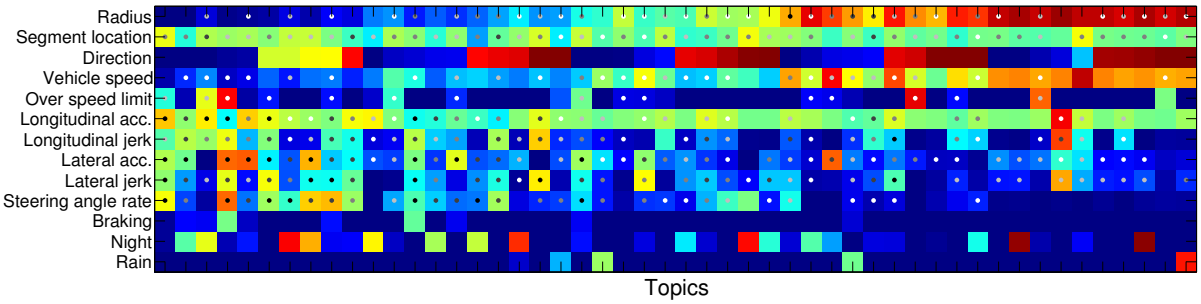


Figure 22: *Topics highly related to direction are shown here. Five inner and five outer curves from each radii group. Here blue is inner curves and red are outer curves.*

The first thing to observe is that nearly no topic contains a high concentration of outer curves when the radius is small (the ten topics to the left in Figure 22). This is very likely due to the different amount of curves extracted for outer and inner curves. As more inner curves are available for topics with small curve radius, the number of topics containing inner curves increases for this group. The same trend can be seen for the topics with large curve radius, one of the topics for inner curves does not have as large concentration as the others, which indicates that more outer curves exist for that group of curves.

Some trends can be observed in the different radii groups. In general, a higher speed is more common in topics with inner curves (blue in Figure 22) compared with outer curves. This behaviour can be observed for all radii groups. Lateral acceleration is also somewhat higher in inner curves.

Night

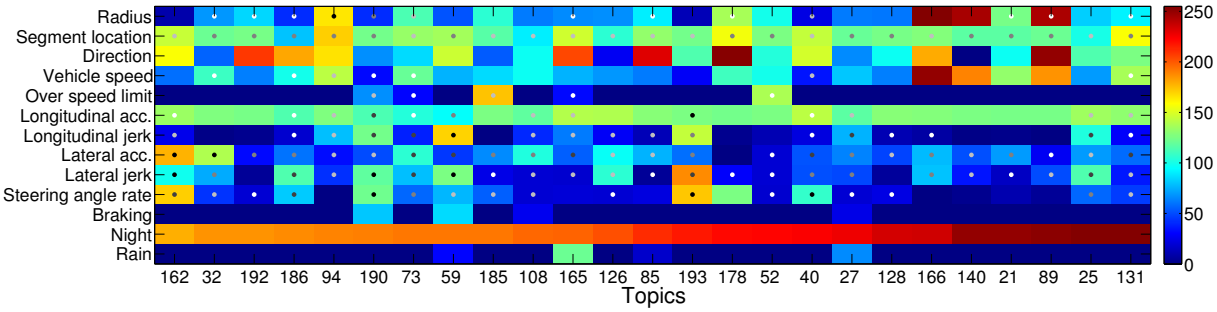


Figure 23: *Top 25 topics when sorting after night frequency. Dark red is only night while dark blue is only day. Anything in between is a combination of the two.*

Nighttime driving only occurs in 20.2% of the analysed trips. Most of the topics with the highest relevance for nighttime driving have a tendency for more careful driving, with low values on the features related to driver characteristics. In other words, careful accelerations, little jerk and nearly no speeding.

There are topics with slightly more aggressive handling; two topics include a moderate amount of speeding (topics 185 and 52) and some topics have higher values for lateral and longitudinal jerk as well as steering angle rate. There are no significantly high or low values for longitudinal acceleration, indicating that keeping constant speed and careful accelerations are common in these conditions.

The three topics with higher longitudinal jerk, steering angle rate, lateral acceleration and jerk (topics 162, 59 and 193 in Figure 23) are all curves with small radius. This follows what has been observed before, that these features have higher values for smaller curves. This is still true for nighttime driving; however, here these values are not very high. This strengthens the indication of more careful driving behaviour during nighttime.

Rain

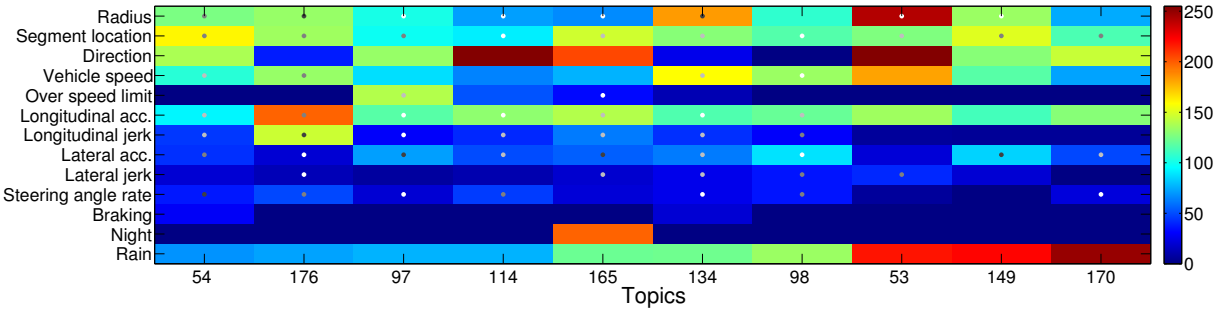


Figure 24: *Top ten topics when sorting after rain frequency. Higher on the scale have more rain (red) and low values represents no rain (blue).*

There are not many trips with rainy weather, only 3.3% of the trips were made in such circumstances. Hence, there are not many topics containing a high percentage of this kind of trips. As can be seen in Figure 24, only three topics contain a high amount of rainy weather. There are an additional three topics with approximately equal amount of curves with rain as no rain.

In these few topics there are no significantly high values for any of the features related to driver characteristics. Topic 176 has higher values in longitudinal acceleration and comparatively more in longitudinal jerk, but on the other hand not very much rain. The topics with rain indicate a more cautious driving behaviour in rainy weather. However, as there is so little data with rain it is difficult to draw any definite conclusions.

5.2.2 Tendencies related to driver characteristics

As mentioned in Section 4.5.3, each of the features related to driver characteristics, is divided into four groups; large, moderate, low and none. In Table 11 below are the different topics which are connected to each tendency. As can be seen, there are few instances where drivers brake within a curve, which resulted in few topics containing any braking and no topic that contains a high degree of braking. Jerky curve handling on the other hand, has many topics included since it is made out of three separate features. Some of the topics have high values in more than one of these three features and are therefore counted more than once.

Tendency	**Level**	**Topics**
Speeding	High	15, 24, 44, 189
	Moderate	19, 26, 35, 52, 64, 97, 139, 150, 156, 185
	Low	57, 69, 71, 74, 78, 82, 99, 107, 116, 132, 158, 168, 180, 181, 190, 195
Braking	High	
	Moderate	24, 112
	Low	7, 36, 59, 104, 118, 190
Jerky curve handling	High	5, 10, 64, 72, 88, 104, 187
	Moderate	16, 24, 118, 123, 133, 141, 174, 193
	Low	1, 9, 30, 41, 55, 59, 84, 92, 104, 143, 144, 157, 162, 173
Rough curve handling	High	31, 36, 88, 191
	Moderate	24, 64, 72, 174
	Low	99, 104, 162, 198

Table 11: Which topics fall under which level for each tendency sorted after topic number. Topics not present under a tendency have no relation to that specific tendency. For the Jerky curve handling tendency some topics are counted multiple times, in this case the topic is placed in the highest group it belongs to (in the table).

Speeding and Braking had a correlation of 0.6, which indicates that drivers who tend to speed also tend to break more; for example there is a driver who has the highest value in both tendencies. When speeding in large radius curves, breaking is not necessary, making the correlation somewhat lower. Jerky and rough curve handling also seem to be more correlated compared to the other tendencies; for example, there exists a driver with the highest value in both tendencies. However, individual drivers can exhibit one tendency substantially more than others, even though drivers who exhibit much of one tendency tend to exhibit more in other tendencies too. For example, two drivers appear several times in the top 3 of the different tendencies, they are also shown in Figure 16 in Section 4.5.3.

In some tendencies there exist multiple drivers with a value of 0. Just as those with high tendencies in one area tend to exhibit more tendencies in the other areas as well, the reverse is also true. Careful drivers tend to have a lower level in most tendencies, with zero values or close to it on most tendencies even though variations exist. Some careful drivers have very low values in Speeding and Braking while Jerky curve handling and Rough curve handling are high. Hence, their average tendencies are comparatively high.

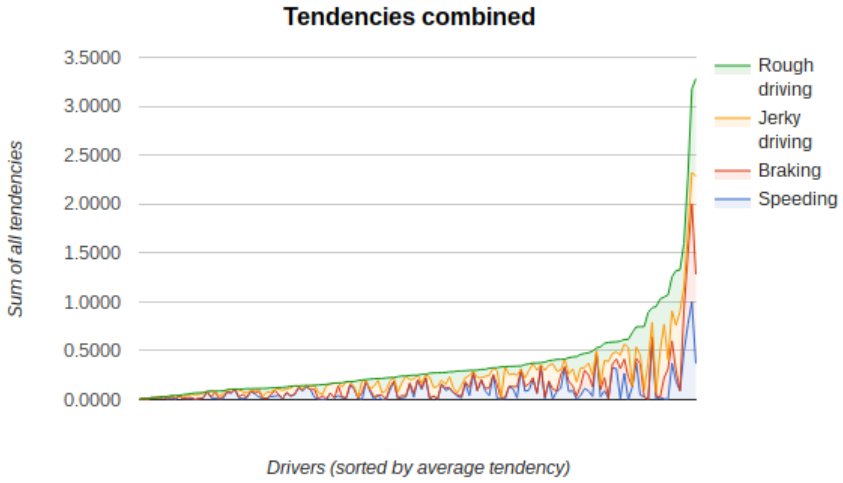


Figure 25: *Tendencies for the 141 drivers sorted by the average amount of tendencies. The tendency values for each driver are summed together here. How much each tendency add to the sum for each driver can be seen as well.*

As can be seen in Figure 25, there are a large amount of drivers with relatively low total amount of tendencies. A few drivers have a higher amount of tendencies and could be considered more aggressive drivers, prone to risky behaviour. The drivers above the 85th percentile have an average tendency level above normal behaviour (average + standard deviation). As can be seen in Table 12, the median value is substantially lower than the average, hence, a few drivers with high level increases the average. We call the drivers above the 85th percentile aggressive, and the drivers below the 15th percentile careful drivers. The rest of the drivers are considered normal. In Figure 26, the drivers have been divided into these groups, and the average tendency for each group can be seen. Values for each specific tendency separated by group can be found in Appendix C. As can be seen the average values for the aggressive group are significantly higher than the other two.

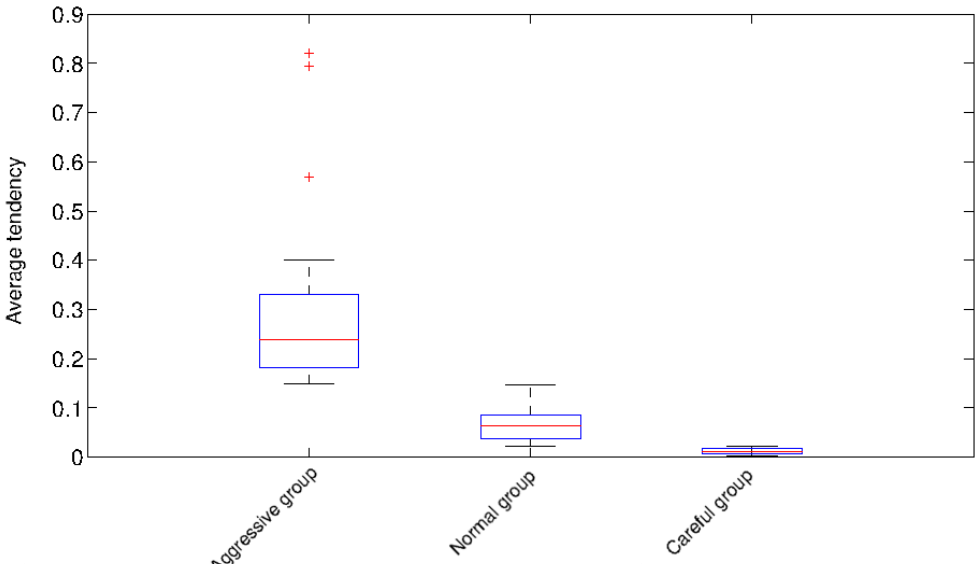


Figure 26: *Average tendency values for the three driver groups. As it is average tendency the highest possible value is 1.*

In Table 12 the Braking tendency has lower values than the other three tendencies. Both the median and average are significantly lower, especially the median. This is an indication that many drivers do not brake much at all inside curves and that a few drivers brake substantially more than the others. The other three tendencies are more common than Braking and have similar values. The standard deviation σ is similar for all tendencies inside each driver group. However, σ is significantly higher in the aggressive group than in the other groups.

Measure	**Speeding**	**Braking**	**Jerky driving**	**Rough driving**	**Sum**	**Mean**
Median of all-drivers-group	0.0400	0.0092	0.0644	0.0441	0.2501	0.0625
Median of aggressive-group	0.1294	0.1456	0.2191	0.4531	0.2377	0.9509
Median of careful-group	0.0118	0.0011	0.0095	0.0037	0.0108	0.0431
Average of all-drivers-group	0.0967	0.0579	0.1047	0.1159	0.3752	0.0938
Average of aggressive-group	0.2416	0.2469	0.2836	0.4427	1.2148	0.3037
Average of careful-group	0.0158	0.0066	0.0183	0.0077	0.0484	0.0121
σ of all-drivers-group	0.1456	0.1417	0.1397	0.1840	0.4780	0.1195
σ of aggressive-group	0.2818	0.2960	0.2573	0.2742	0.7805	0.1951
σ of careful-group	0.0173	0.0137	0.0186	0.0119	0.0067	0.0265

Table 12: The average, median and standard deviation in the tendencies for all drivers are shown here together with values for the 15th percentile drivers with the lowest level of tendencies (which we call careful-group) as well as the above 85th percentile drivers with the highest level of tendencies (which we call aggressive-group). σ is the standard deviation.

The age, gender and estimated amount of driving per year (estimated by the drivers themselves)are shown in Table 13. The careful driver group has more women than men (even considering there are more men in the dataset), they are usually older than average and they drive more per year than average. There are a few outliers with a few drivers who only drive around 1000-3500 km per year as opposed to the mean of 20000 km, as they drive less they could be a bit unsure and therefore drive more carefully.

A deeper analysis of the aggressive group suggests two subgroups. One group is entirely made out of men who drive more than 30000 km per year and are between the ages of 37-62. The other group consists of both men

and women who drive less than 5000 km per year. These could be considered to have less driving skill and not very used to driving, which could lead to the current classification.

Data partition	**#Male**	**#Female**	**Mean age**	**Mean distance driven annually**
All drivers	85	56	47.4	20830 km
All drivers in careful group	10	11	48.8	26890 km
All drivers in aggressive group	14	7	44	20997 km

Table 13: Driver data for different partitions of the dataset.

In Table 14, the gender, average age and distance traveled for the aggressive drivers within each tendency can be seen. Those deemed to be aggressive within each tendency are those who have values which are above the mean plus one standard deviation.

Interestingly, the drivers who tend to speed more often are older than average and drive less per year. Drivers who drive less also seem to brake more than the majority of the drivers. Those drivers who exhibit more of Jerky curve handling are in the standard range for those who we consider are aggressive drivers. When looking at Rough curve handling we find that there are more men than women among the drivers with high tendency values.

Aggressive tendency	**#Male**	**#Female**	**Mean age**	**Mean distance driven annually**
Speeding	9	7	51.6	12820 km
Braking	6	4	43.875	15392 km
Jerky curve handling	9	4	44.8	19600 km
Rough curve handling	12	3	45.5	25730 km

Table 14: The statistics for the drivers which are considered aggressive within each tendency.

6 Discussion

The following discussion is divided into three parts. The first part concerns the curve extraction, features used and the data reduction method. A discussion about the advantages and disadvantages of using LDA follows in the next part. Finally, the results from the LDA are discussed and compared to previous studies.

6.1 Curves and features

6.1.1 Curve extraction

The method for extracting the curves seems to have worked well for finding start and end points in the curves. The correctness was confirmed with video observations and therefore it is up to human interpretation. Some thoughts and potential improvements are discussed below.

Consecutive curves are not used in this thesis as they are not comparable to single inner and outer curves. However, these curves were also extracted, as seen in Table 8 in Section 5.1.1. To incorporate these kinds of curves into the analysis would have been interesting as different behaviour is expected when entering and exiting them. Consecutive curves can be regarded as either a single curve or as multiple separate curves. If they are regarded as multiple curves it will not be possible to include the tangents 100 meters before and after the curve, hence, missing possible deceleration and acceleration before and after the curves. To look at a consecutive curve as a single, big curve has problems as well. It would not be possible to divide the curve using equidistant segments. A possibility would be to mark these kind of curves in some way or analyse them separately and then combine the results with the analysis of inner and outer curves.

The total number of inner and outer curves used in the analysis are about the same, 36695 and 37325 respectively, but the estimated radii distributions for inner and outer curves are different. There are two possible explanations for the skewness in radii between the inner and outer curve groups. The first possibility is that the imbalance is introduced during the curve extraction. The steering angles used for the identification of curves always differ by exactly one degree. It could be that the difference in angle should be adjusted depending on the curve radius, speed or some other factors. The second possibility is that the radius estimation is the reason for the imbalance. When the radius is approximated it is the path radius that is estimated, which depends on when the driver actually starts turning and not when the road turns. We have taken this into account by adjusting the radius slightly but have not considered whether it is an inner or outer curve. Hence, it is possible that the estimated radii for inner and outer curves are different even if the same geometrical curve is used.

The use of the steering angle for curve extraction introduced some problems but, as an independent variable, is reliable. To use another variable, such as the heading change, could be preferred but then other errors need to be taken into account, e.g. GPS inaccuracy. Yaw rate and lateral acceleration are two other possible candidates to use. Yaw rate was used at first but later changed to steering angle. These two variables are basically equivalent from a curve extraction viewpoint but steering angle was chosen because it is much more intuitive. Yaw rate and lateral acceleration would also be more affected than steering angle by superelevation, although this is only a small issue for high speed roads.

The values on the required steering angle thresholds were decided by direct video observations. These values can be adjusted to find more but less sharp curves, or less but sharper curves. Increasing and decreasing the thresholds have both benefits and drawbacks. More data is always an advantage but the data for sharper curves are more interesting from a driving style perspective. Mild curves, where not much change in speed and other variables can be observed, are less interesting since drivers do not drive in a different manner compared to straight road sections. The values in Table 1 in Section 4.2 gave both a good amount of curves and curves that are sharp. The large curves could be a bit sharper, but as we did not have very many of them we decided not to change these thresholds. For lower speeds, the required steering angle was high but many small radius curves were still obtained.

As can be seen in Table 7 in Section 5.1.1, there were comparatively more small radius curves than large radius curves. Generally when the speed is higher, the curve radii are larger, hence, there are two possible explanations. Either the thresholds for high speeds should be lowered somewhat, or trips contain more small radius curves than large radius curves. As stated above, the thresholds for high speeds should not be lowered as they already are a bit too mild, hence, we can conclude that the trips contain more curves with small radius even when only data from rural areas are used. As the dataset used is normal day-to-day driving this is not very surprising as smaller radius curves are more common to come across when driving normally.

6.1.2 Curve radius estimation

The curve radius was estimated for all curves but whether the estimations are truly accurate is difficult to assess. When selecting radius estimation method a brief evaluation of different methods was performed; however, more trips could have been used. The biggest curve used in the evaluation was 500 meters which was the biggest curve Imberg and Palmberg (2015) analysed while our data included curves with a radius up to 1200 meters. In other words, it is unknown how the method performs for curves with larger radius than 500 meters. However, Othman et al. (2012) used the same technique before on naturalistic driving data with multiple trips averaged. We expected larger errors but as we did not require very exact estimations, we consider this method to be adequate for our purpose.

To obtain better estimations for the curve radii it would have been possible to first find all curve sections and connect them together using GPS positions; in this way the geographical curves could be obtained. Data from multiple trips could then be used to estimate the curve radius. Othman et al. (2012) wrote that a mean of multiple trips is necessary for curve radius estimation, and by performing this extra step this would have been possible. Regrettably, time constraints made this improvement impossible to accomplish.

6.1.3 Features used

Segment location
Each curve was divided into 10 segments in accordance with Imberg and Palmberg (2015), with an additional 2 for the tangents. However, Imberg and Palmberg (2015) did not select their segments with uniform length; they had more segments in the middle of the curve and less further away. The optimal number of segments was not tested, but is highly interesting as each segment added will also add one more word for each curve. Hence, more segments will give more words to use with LDA. 12 segments are enough for shorter curves, since they do not contain much data. Additionally, the speed for a typical large curve is illustrated in Figure 3 in Section 3.2 and the use of 12 segments appear to be satisfactory when approximating longer curves.

Only significant changes in the feature values can be detected when using SAX. As a result, values close to each other will be reduced to the same letter. Increasing the number of words could potentially give more duplicate words, which would increase computing time. A possibility could be to use a different amount of segments for small and large curves. If more segments are used for large curves, the number of words containing large curves would increase. Hence, a bias towards large curves would have been introduced because the number of words with different curve radii would not reflect the actual number of curves. To use fewer segments would make later approximations with SAX unreliable for larger radius curves, which is undesirable.

Two of the segments represent the 100 meters in the beginning and the end of a curve. These segments are mostly interesting for smaller radius curves, since for larger curves not much change during approach and departure can be observed. To consider the tangents before and after the curves is important, because interesting patterns can be observed in these road sections. Pérez-Zuriaga et al. (2013) stated that the point of deceleration is between 50 and 100 meters before a curve. Therefore, 100 meters both before and after curves are taken into account. However, as we use SAX on these segments we only obtain mean values. If the driver starts slowing down 50 meters before a curve, then we have an additional 50 meters with normal straight driving in this segment. Montella et al. (2015) found deceleration lengths between 230 and 50 meters depending on the curve radius. If the driver starts decelerating 230 meters before the curve, then at the point 100 meters before, the vehicle could already have decelerated to a speed acceptable to traverse the curve. In such a case no deceleration will be detected on the approaching tangent. Therefore, the approach and departure segments can

have many values closely related to straight driving in our case, as a mean is used and not much change in the features is expected. The exception is smaller curves where a larger change can be observed. Hence, finding the point of deceleration for each curve and use that point as the start of the first segment could be of interest.

The curves that are found are treated as simple curves, but they may also be compound curves. As a result, the sharpest section in a curve does not have to be in the middle segments after SAX. For LDA, the middle four segments are labeled the same (as the letter c) to try to catch the sharpest part of a curve, but there could be exceptions. Even if the segment location is in the middle (the letter c) it does not guarantee that this is the middle of the geometrical curve.

Over speed limit
The speed limit variable was not always available, making the feature representing speeding incomplete. The simplification to treat an unknown speed limit as non-speeding made the analysis easier, but the use of a separate symbol for unknown data was another action considered. If a new symbol is used when data are unknown, we would still not be able to determine whether the drivers are speeding or not, hence, these topics would be treated the same as non-speeding.

Some drivers who speed will be missed or have a lower tendency due to the incomplete data for speed limits. This will be true regardless of analysis method and cannot be changed. To predict the speed limit when unavailable should be possible and is something that could be incorporated to improve this feature.

Longitudinal acceleration
The longitudinal acceleration feature is slightly different compared to the other features as it is not possible to make use of the absolute values due to the different meaning of positive and negative acceleration. Attempts were made to split this feature into two separate features to allow the scale to be the same as for all other features. This introduced some problems, for example, the situation where in a segment there was one positive value and the rest negative. Then the value for the positive acceleration feature would be this single value while the negative acceleration feature would have the mean of the other values. This segment would then have both positive and negative longitudinal acceleration, but the knowledge of how much of each is lost. Another method was to split the feature into the absolute value and percentage of time with positive acceleration, the rest of the time there would be deceleration. This did not work out since after LDA this feature did not give us any sensible information. Hence, a single combined feature for acceleration and deceleration was used.

Additional features considered
There are features that could have been useful but cannot be directly derived from the available data. A more in depth investigation of each separate curve would be needed, but the features are mentioned here because they definitely affect driving styles in curves. An interesting feature would be the degree of superelevation, i.e. the difference in elevation between the two edges of the road. The outer edge of a curve is often raised to allow vehicles to use a higher speed through the curve. If the curve is in an uphill or downhill slope would also be good to know since it affects e.g. the speed and the longitudinal acceleration.

The time of year has not been considered. During the winter months the presence of snow can change the driving behaviour radically. Other studies have shown that driving speed and other driving behaviours tend to be more careful in the presence of snow or perceived icy roads, as described in Section 2.1. Since we could not determine whether there was snow or not automatically from signal data, we chose to not take this into account. It should be mentioned that the winter during the data collection was the coldest in Sweden since temperatures started being recorded. Hence, to compare differences in driving styles during the winter and summer seasons for this dataset would be interesting future work.

6.1.4 Data reduction

Other methods than SAX were considered, such as Discrete Fourier Transform (DFT) and PCA. Since SAX was proven by Lin et al. (2003) to not be less accurate than DFT, there were no qualms about using SAX in that regard. PCA has been used before by Constantinescu et al. (2010), Chandrasiri et al. (2012) and Towfic (2014). However, PCA merely reduces the number of features and therefore the amount of data for each component created by PCA will be the same as for each feature; thus, requiring the use of some other data reduction technique. However, PCA could have combined features into components, thereby increasing the total number of features used. It is not certain that utilizing PCA would be superior to selecting variables via heuristics since the output from LDA needs to be interpreted by a human and PCA components might not be as easy to understand.

Lin et al. (2003) suggested an alphabet size of five to eight symbols stating that there are diminishing returns for larger alphabet size since less reduction of data takes place. We also found that more letters obfuscate what we can learn from LDA. A lower amount of letters obviously reduces too much data. Most features had an alphabet size of five. The few features that had fewer letters either did not have enough variation in the data to warrant more breakpoints, or they were binary in nature in the first place, such as rain and night.

One could argue that the breakpoints we selected could be changed. For example; we added a breakpoint at 125 km/h instead of the highest breakpoint being 100 km/h for speed which we thought could help find cases where a high speed was indicative of speeding, especially where the legal speed was not available. This did not change the results much, except that the radius to speed correlation was improved.

6.2 LDA method discussion

As explained in Section 4.5 there are many parameters which can be adjusted such as number of topics, number of iterations, and as α and β values. We chose a high number of iterations so that the solution would not change the output much between iterations, which is the case with a lower number of iterations. How to interpret the output is something highly subjective which is often the case with LDA. Hence, heuristics were used to make judgments of which parameters would give an interpretable result. The parameters could presumably be optimized further.

There are many features that decide what a specific topic contains, and as such, it is difficult to draw any definite conclusions on the contents of the topic. Only the trends in the topics can be investigated. The different amount of curves extracted for the different curve radii can skew some of the results. The output of LDA is sometimes somewhat skewed in the number of topics that express a driver; for some drivers this number is relatively small. In other words, topics representing one driver rather than a certain driving style can appear. We have tried to counter this by adjusting the input parameters α and β, but this behaviour is still visible in a few topics. The number of topics was also increased to prevent this kind of behaviour. In this way, the number of topics each driver is expressed by, is increased. Thus, even if some topics have one driver overrepresented, other drivers are also described by that same topic.

LDA is very good at finding similarities on a high level, which is somewhat detrimental since we are interested in finding uncommon as well as common behaviours. Increasing the amount of topics (200) past the amount of drivers (141) increased the chance of finding outliers. We also chose to disregard if there is lower probability of a certain topic to appear. Topics with lower probabilities have a higher chance of containing more uncommon behaviours and, as these are of interest, we do not differentiate between the topics depending on their respective probability. Likewise we chose to focus on the top words as representative of an entire topic. As a consequence of only using the top words for the analysis, some topics that are classified as e.g. high level of Speeding could, if more words were used, become a moderate level of Speeding instead. If the unused words, which are less likely to appear, were included in a topic, then it would in some cases be harder to see clear trends within that topic.

To understand the topics, a visual representation was made which performs a second approximation on the features. The second approximation has flaws compared to SAX since curve segments vary within a topic and therefore large variations in words can be seen. To try to illustrate the variations we used the standard

deviation to mark which topics were affected by large variations. However, the standard deviation was not used extensively during the analysis and may require further adjustments. The term tendency was introduced to indicate certain behaviour because we cannot guarantee that a topic expresses only the behaviour we are looking at. We merely observed that within a topic, drivers tend to, for example, drive quicker than the majority of the drivers. Thus, we claim that drivers that have a relation to that topic might also have a tendency to speed.

To decide which drivers expressed a large amount of the tendencies related to driver characteristics, we chose to look at the topics that had high values, i.e. values close to the higher range of the breakpoints. We also made the assumption that a majority of our drivers have a normal driving style. How the breakpoints are selected is based on this assumption. The most aggressive driving and other behavioural outliers are expected to be when the values of the features are in the upper ends of values in the dataset (inside curves), i.e. above 80% or 90%. Not all features can use this method to select breakpoints, e.g. braking and over speed limit used other thresholds. Feature values of e.g. lateral acceleration will change depending on the position in a curve and values will likely be higher in the middle of a curve. Hence, even if a value is above 90% it does not necessarily have to be a strong indication of aggressiveness. However, it may give an indication nonetheless.

The topics used are broad in scope and some approximations have been done. Therefore, we do not provide absolute values for aggressive driving styles but tendencies. Since all steps have been mathematical in nature, we are confident that the tendencies presented are reasonable from what we can infer by looking at the available data. However, an expert in analysing driving styles might not agree. Our results are only based on the behaviour characteristics in some features derived from signal data, while an expert would probably use the whole dataset, including videos. We believe that tendencies are still indicative of driving styles and a useful way of categorizing what we learn from our topics.

Similarly to Constantinescu et al. (2010) we tried to find topics which exhibit some kind of aggressiveness or abnormal behaviour. Determining whether the output is correct is difficult when using unsupervised learning and assessing a relative risk with the help of groupings, as Constantinescu et al. (2010) did, seems more appropriate. Towfic (2014) assigned levels of risk depending on the value of certain components similar to our topics. The high risk components contained speeding and high positive accelerations as well as frequent braking and high mean speed; these are similar to our tendencies.

The tendencies related to driving context are a bit peculiar because it is only possible to compare general trends within each tendency. For example, there are many more topics with a high relationship to day compared to night. Therefore, there will be more variations in the topics related to day and more behavioural outliers can be identified. Hence, it is difficult to say whether there is more speeding (or some other behaviour) during day or night. However, it is still possible to look at all topics with night and identify some trends in these topics. It is likewise difficult to combine multiple features during the analysis, as only 200 topics are generated. For curve radius it works, however, for e.g. night and segment location it is more difficult. There are not enough relevant topics for these features, making any fair comparisons or reaching any motivated conclusions difficult. Therefore, all tendencies related to driving context are analysed separately except for when the curve radius is used.

141 drivers and over 74000 curves comprise a large amount of data. However, it might not be representative for a larger population to such a degree that the same breakpoints can be used on other datasets. The breakpoints could, however, be indicative of good breakpoints, and the method for selection could be used to calculate new ones. Larger populations, e.g. the Swedish population, could exhibit more or less aggressive behaviour. To determine the driving style of a driver we compare it to the other drivers in the dataset, hence, more available drivers will make these classifications more precise and provide more accurate results. To improve the classifications even more some labeled data would need to be used. If an expert could label a couple of drivers or trips as having a specific driving style, then it would be possible to use semi supervised machine learning methods on the data. The negative aspect of such an approach would be that it is built upon a subjective input. Furthermore, it is hard for an observer to only study data and video to determine driving style without being physically present inside the vehicle. However, if an observer is present the driver is likely to drive in a different way than normal.

The selected number of segments and breakpoints gave us a high number of unique words. Slightly less than 1/3 of the words were unique. Since there are many unique words, this means that LDA has trouble grouping

the drivers together in a good way since it is based on commonalities between the drivers. To reduce this number, less features or segments could have been used. Another alternative would be to increase the data size; however, this is usually difficult to do as the data size is one of the constraints when analysing these types of problems. Data from multiple sources could, however, be used if they contain the selected features.

6.3 LDA results discussion

6.3.1 Tendencies related to driving context

Five different tendencies related to driving context were analysed and some trends could be observed in all of them. Curve radius have according to previous literature a high influence when driving in curves. Particularly when speed is examined, the radius has been identified as the factor with the highest impact (Quaium, 2010; Othman et al., 2014; Imberg and Palmberg, 2015). We found that the speed was higher for curves with large radius which conforms with the reviewed literature.

A trend that drivers keep a more constant speed when the curve radius is large was identified by studying changes in the longitudinal acceleration. In contrast, small radius curves had higher levels of speed differentials. Similar behaviour was noticed by Imberg and Palmberg (2015), who found that the speed does not decrease as much during the approach tangent if the curve radius is large; consequently, drivers accelerate less after such a curve. These results also conform with studies by Hu and Donell (2010) and Altamira et al. (2014). Both of them found high deceleration rates during approach and high acceleration rates during departure in cases where the curve radius was small. An explanation could be that as the radius is larger it is possible to keep a higher, more constant speed and therefore drivers do not need to decelerate as much compared to curves with small radius. Since a high deceleration is unnecessary, only a low acceleration is needed for drivers to reach their desired speed after a curve.

Higher values for both lateral acceleration and lateral jerk were present in curves with smaller radius when compared to large radius curves. Othman et al. (2014) and Imberg and Palmberg (2015) observed similar behaviour and therefore support these findings.

Speeding seems to take place independent of curve radius. However, a higher level of speeding was seen when the curve radius was larger and for higher speeds (>100 km/h), whereas a lower level of speeding was observed for curves with smaller radius. Nie (2006) found that operating speeds on freeways exceeded the speed limit to a significant degree (>15 km/h). On lower speed roads we expect the speed over the posted limit should be lower compared to freeways and highways due to the curve geometry (i.e. smaller radius curves). This expectation and results from Nie (2006) match our results.

The segment feature did not give any significantly high or low values. However, a trend to decelerate before a curve and accelerate afterwards can be identified. This observation is supported by previous studies (Hu and Donell, 2010; Pérez-Zuriaga et al., 2013; Altamira et al., 2014). There also seems to be more deceleration present before a small curve compared to a large curve, which Imberg and Palmberg (2015) agrees with. Othman et al. (2014) found the highest lateral acceleration to be at the entrance of a curve, and that the lateral acceleration decreased as the vehicle approached the exit. Looking at the topics with the highest relevance to lateral acceleration, i.e. high and medium levels, the values for the segment location feature are only in the beginning or middle of the curve. Hence, our results seem to be conform with what Othman et al. (2014) found. Earlier studies identified different behaviours depending on the length of the tangents (Hu and Donell, 2010; Imberg and Palmberg, 2015). However, the only information available in our case is that the tangents are 100 meters or more, as only curves with 100 or above meter tangents are extracted. Hence, any behaviour differences depending on the tangent length could not be seen. The presence of spiral transitions is also a factor known to affect driving behaviour but this feature is also unavailable.

The direction feature has a general trend towards higher speed in topics with inner curves. These curves also have a slightly higher lateral acceleration. The higher lateral acceleration is likely due to the higher speed and sharper angle required to traverse an inner curve (as compared to the corresponding outer curve). Othman et al. (2014) observed similar behaviour in his study and Syed (2005) found a higher lateral acceleration for inner curves in a study using instrumented vehicles.

Compared to the total amount of curves in the dataset, there are few curves during nighttime. Therefore, only the most common trends can be observed and some outlier behaviours as aggressive driving could have been missed. What can still be seen is a general trend of more careful driving during the night. Topics with different curve radii, segment locations and directions are all present in the top topics for relevance to nighttime driving, suggesting that the behaviour is due to darkness.

Earlier studies on two-way rural highways have shown no significant speed differences for day and nighttime driving (Fitzpatrick et al., 2000; Donnell et al., 2006). Quaium (2010) especially studied curves under these conditions and found no correlation between speed and the time of day. Not only two-way rural highway curves were used in our study and in the results small speed differences cannot be detected as the breakpoints for speed are not detailed enough. Both night and day driving look the same from a speed perspective. Curves on two-way highways tend to have larger radius, the features of tendencies related to driver characteristics tend to be low on these topics even during the day. We can neither confirm nor oppose these earlier results for the speed on two-way rural highways. An analysis of only two-way rural highways focused on the speed feature (by removing some other features and adjusting the breakpoints for speed) would be necessary for further comparisons.

Previous research found that drivers tend to slow down more during nighttime in curves with low visibility or low retroreflectivity compared to during the day, as seen in Section 2.2.2. A low visibility can be present in many of our curves, especially those that have smaller radius. This could explain the more careful behaviour observed in our results. Hence, there can be two explanations for the careful behaviour observed at nighttime: that the drivers are more careful (especially when the curve radius are smaller) or that more topics and data need to be used to find outlier behaviour. A separate analysis on only data from nighttime driving could be attempted; however, not much curve data are available compared to the total amount. To only use this data as input for LDA would still be possible and, in this way, all curves would be during nighttime and different driving behaviours could be identified.

Even less data are available during rain compared to trips at night. A more careful driving behaviour was also observed for this feature but there is too little data available. Therefore, no conclusion can be drawn except that careful driving is more common than aggressive driving. Earlier studies on driving in rain, see Section 2.1, showed a more careful behaviour and as we did not detect any aggressive behaviour we can at least corroborate that drivers do not drive more aggressively during rain than without rain.

6.3.2 Tendencies related to driver characteristics

Jerky drive handling is represented by three features, some of which are somewhat correlated to each other. High values are common to see at the same time, especially for lateral jerk and steering angle rate. To remove topics that are counted multiple times is an option; however, we want to catch the differences between the three variables. Hence, the topics with high values in more than one feature have higher weight than topics which only appear once.

As mentioned earlier, there is a possibility for a situation to occur where only a few topics describe a document (driver), instead of many topics. This might have happened with one of the most aggressive drivers who had a high relation to topic 24, as this topic might have been derived primarily from that driver. Topic 24 had a high level in Speeding and moderate level in Braking; thus, the driver with highest relation to this topic will have high values in both these tendencies. The same trends can be observed when looking at another of the most aggressive drivers and topic 88, which had high levels in both Jerky curve handling and Rough curve handling. This behaviour is still acceptable since these topics exist in many other drivers as well.

Rough curve handling affects the average value of the tendencies the most. Before normalization, the highest value was 0.77, which is the lowest maximum value for all tendencies. Due to the normalization, all values for this tendency will therefore increase and will affect the average more than other tendencies. As each tendency is normalized separately it is not entirely correct to look at the average value of all tendencies. Braking also has lower values but there are also fewer high values overall in braking. Furthermore, a few drivers have very high maximum values for Speeding and Jerky curve handling and when normalizing all drivers', the value will decrease by more than half. Thus, these values are more relevant when comparing drivers within the same

tendency.

When looking at the calculated level of aggressiveness for each driver, it is therefore not appropriate to look at the average value. There are 34 drivers who had values above the norm within at least one tendency, some are not in the aggressive driver group. These drivers could be considered more aggressive than other drivers who, while having higher average, have no single tendency over the norm. However, one might consider weighting the different tendencies towards e.g. which tendencies are more likely to be a source of accidents. For example, if speeding affects more than the other tendencies with regards to crash risk, increasing the weight for this tendency might be appropriate. For this to be feasible, appropriate weights for the different tendencies are necessary, which require a detailed knowledge of what causes accidents in practice. The features selected to analyse should be chosen to reflect this as well.

Sagberg et al. (2015) and MacAdam et al. (1998) found that younger people tend to drive more aggressively. In the dataset there are a lot of older drivers and a few young drivers; hence, the average ages for the different groups are above 40. This means it is difficult to draw any conclusions about new drivers. However, we found a higher average age in the group of careful drivers but drivers with high Speeding tendency have even higher average age.

Solely the gender of the driver does not appear to affect tendencies much. However, the careful group contains more women and the aggressive group contains more men (compared to the average number in the dataset). Rough curve handling have 80% male drivers and they also drive a lot annually. This indicates that they are experienced while still driving quite aggressively, as opposed to driving aggressively because of inexperience. In the review by Sagberg et al. (2015), the authors found that while gender differences in driving are small, men appear to be on average slightly more aggressive.

There are drivers who drive less than 5000 km per year in both the careful and aggressive groups (of different genders and ages). For these careful drivers it is difficult to draw any conclusions. However, the subgroup of aggressive drivers with high mileage are all male, which gives an indication that women who drive much each year are more careful. There are males with high mileage in both the aggressive and careful groups, hence, no such distinction can be observed for men. Previous studies found mixed results with regards to annual mileage and aggressiveness. Our results conform with Lajunen and Parker (2001), who found that women with higher mileage had a more careful driving style while the mileage for men had no relation to aggressiveness.

7 Conclusion

A dataset containing exclusively road curves has been extracted from the EuroFOT naturalistic driving data. Utilizing this dataset, a machine was trained to group common driving behaviours together and connect these to the drivers. First, SAX was successfully used to reduce time series numeric raw data to abstract words. Using these words together with LDA made it possible to create topics which can be used to distinguish between different driving styles, on a high level (tendencies related to driving context) and a lower level (tendencies related to driver characteristics).

Topics created when utilizing LDA express a wide array of driving behaviours in different environments. Using these topics, the drivers were categorized into three separate groups depending on their perceived tendency towards aggressive driving styles compared to the other analysed drivers. The groups of the most aggressive and most careful drivers had trips arbitrarily sampled and a calmer driving style was observed in drivers classified as careful. No significant differences between the groups could be seen when comparing the drivers' age. A more careful driving style was observed in women with a high annual mileage, while no relation between annual mileage and aggressiveness was found in men.

Some general trends were identified among the analysed drivers. For large radii curves, compared to small radii curves, we found higher speeds, less braking, lower speed differentials as well as lower lateral acceleration and jerk. Drivers speeding in large radii curves were found to have higher speeds above the speed limit than drivers speeding in small radii curves. During curve approach drivers tend to slow down and then accelerate up to speed during departure. Inner curves were found to have slightly higher speeds and lateral acceleration compared to outer curves. The results suggest more careful driving during nighttime and in rain, even though there is insufficient data for those features to be confident in the findings. These observed trends are similar to earlier studies in the area.

8 Future Work

Much time was spent on data mining for the extraction of curves. We were only interested in curves but different sections of trips are needed for different projects, such as turns, straight driving, etc. A dataset could be built containing these sections to simplify their use. Proper, formal definitions should be created to help with this. If formal definitions exist, machine learning could be used to separate trips into different sections and label the sections accordingly. Our extracted curves are a good starting point for a set of curves; however, improving the curve radii estimations as mentioned in section 6.1.2 should be considered.

To analyze driving behaviours in more than just curves is a possibility, either for other kinds of road sections or extended to analyzing entire trips. With regards to segmentation with SAX it is possible to combine words which are close and the same, thus reducing the data even further, e.g. allow "abc abc abc abc" to become "abc4". This could be especially useful when reducing data for entire trips where long periods of straight driving can be reduced to just a few words with the right breakpoints.

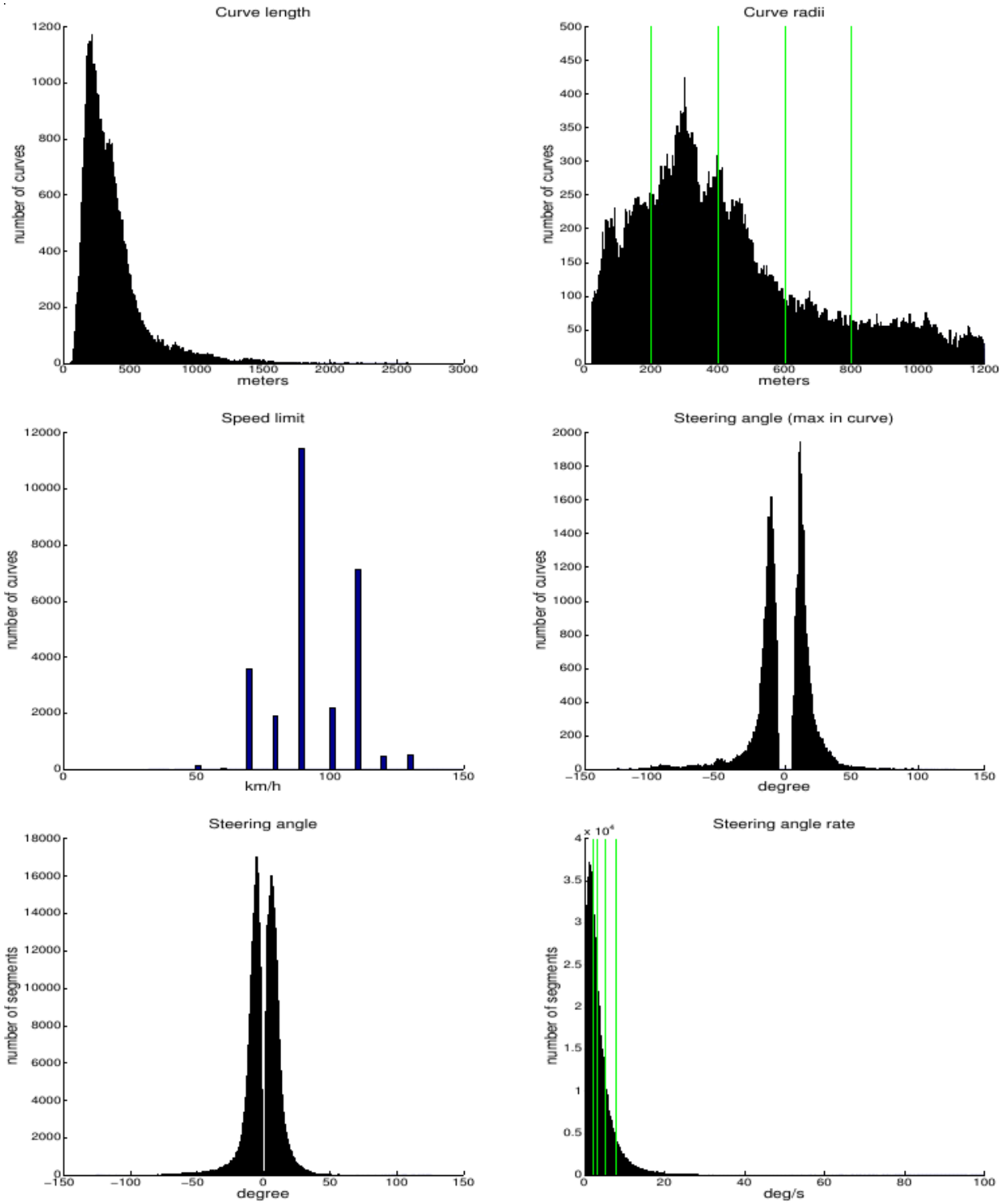
Interpretation of the results from LDA requires more work. Since abstract words are used, the analysis of the outputs is difficult, a fully automatic interpretation would be preferred. The development of such a method could be of high interest in other fields as well, where similar techniques to work with abstract words can be applied.

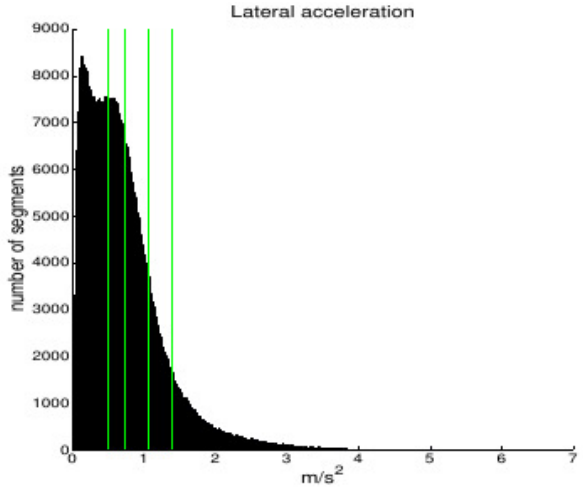
Using the same topics we obtained, it is possible to add new drivers or data in the form of documents. This data could be from other sources or it is possible to try to add the urban driving from the EuroFOT data. The urban data could be used to try match the urban and rural driving styles of each driver together, without checking the driver ids, to see if drivers can be identified simply by looking at topics or tendencies. Another approach could be to compare drivers' urban and rural driving styles.

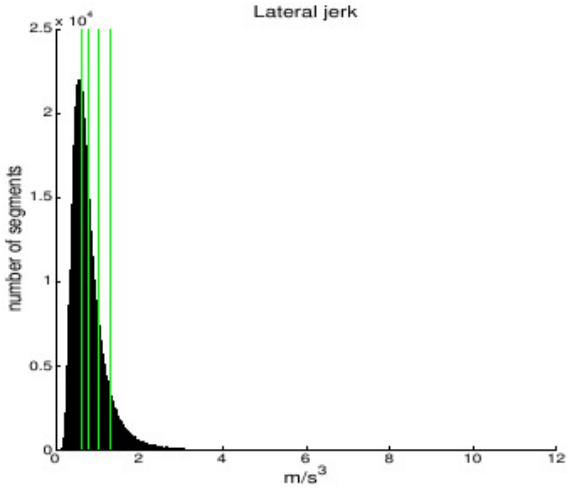
As the drivers now have different values in the tendencies, it is possible to do similar analyses using these values as labels. In such future work, it would therefore be possible to use supervised or semi supervised learning approaches. This does not necessarily need to be for curves; a driver with a more aggressive driving style in curves should have a more aggressive driving style in other road sections.

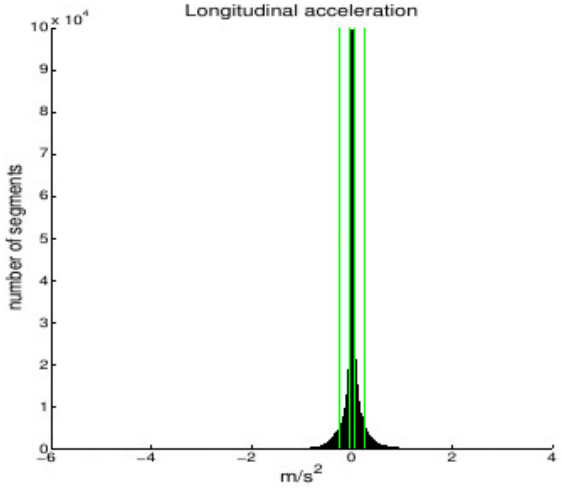
A Histograms

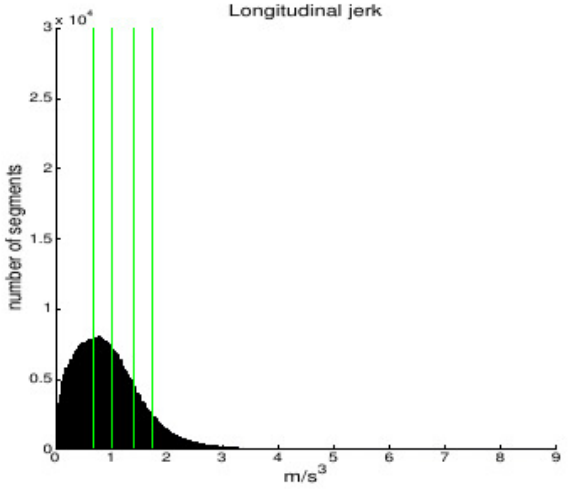
The first four histograms have one value from each curve, all histograms after these count the mean values in each separate segment (the 10 segments for a curve, not including the two tangent segments). The green lines present in some of the histograms show the 40%, 60%, 80% and 90% groups used as breakpoints for SAX. The histogram with speed limits only takes curves where we have that data available into account.

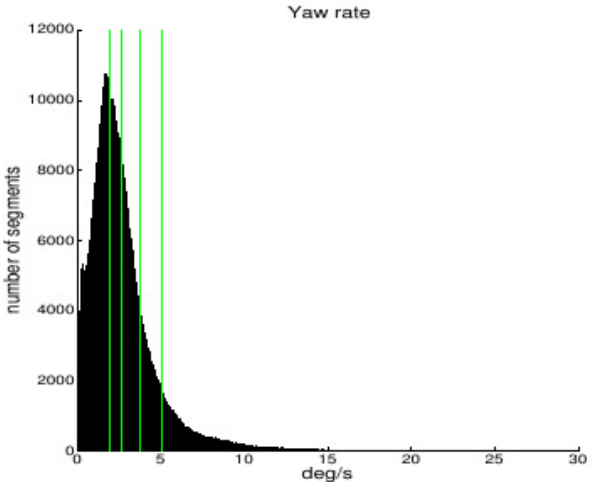



Lateral acceleration
number of segments
m/s2

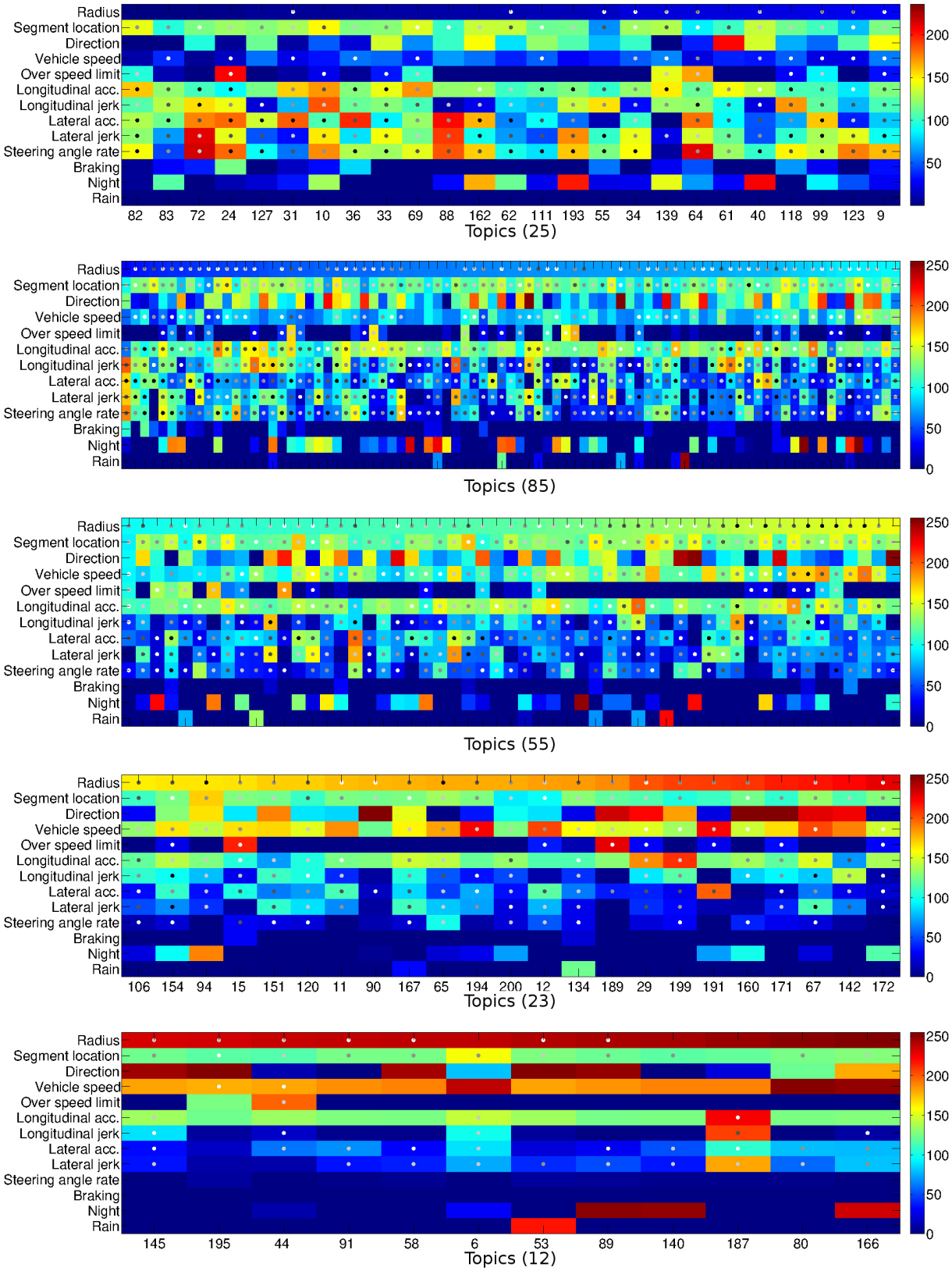

Lateral jerk
x 10^4
number of segments
m/s3


Longitudinal acceleration
x 10^4
number of segments
m/s2

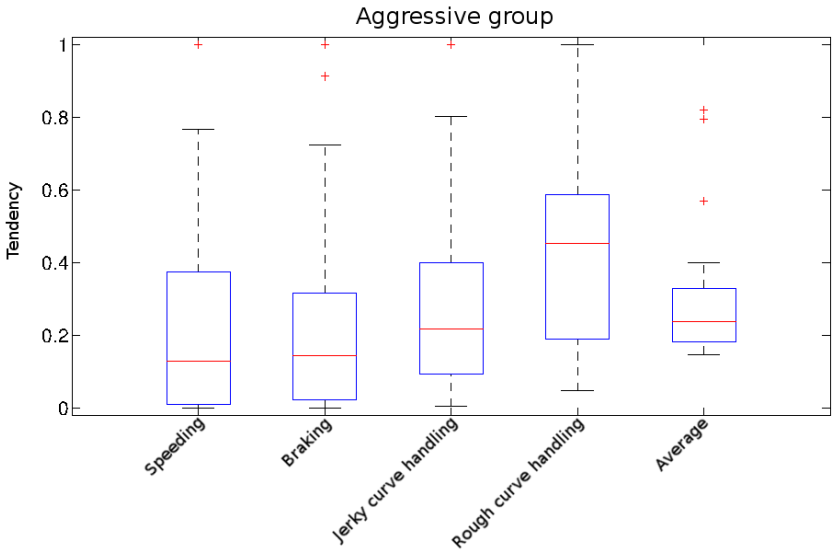

Longitudinal jerk
x 10^4
number of segments
m/s3

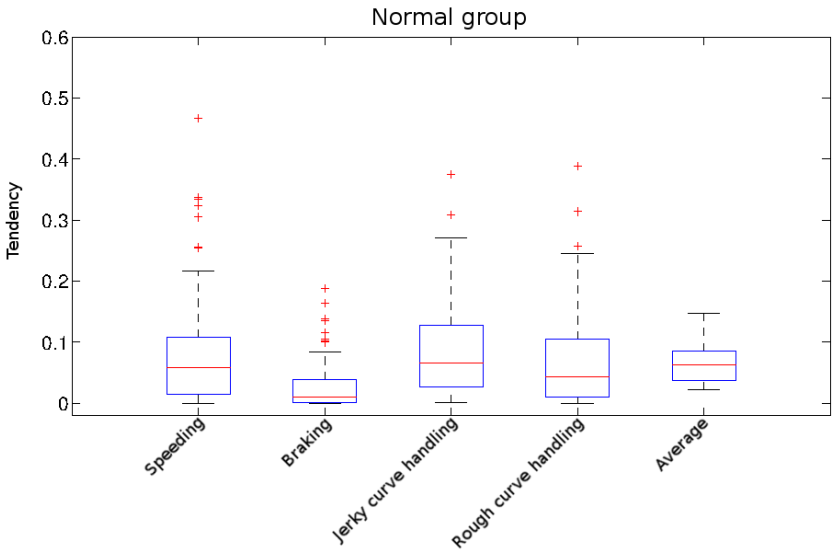

Yaw rate
number of segments
deg/s

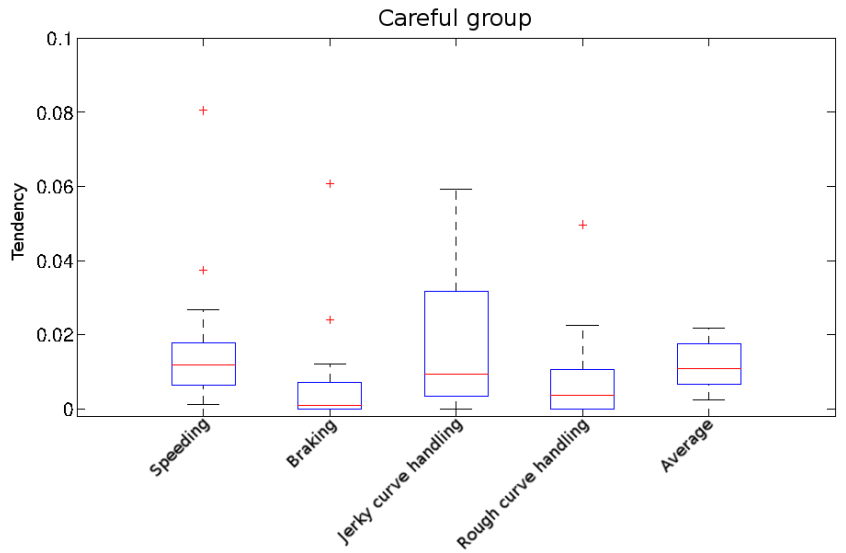
B All topics sorted after radii (from small to big)



C All driver tendencies







References

AAA Foundation for Traffic Safety (2009). *Aggressive Driving: Research Update*. Available at: `https://www.aaafoundation.org` (2016-01-12).

Aarts, L. and Van Schagen, I. (2006). Driving speed and the risk of road crashes: A review. *Accident Analysis & Prevention*, 38(2):215–224.

Altamira, A., García, Y., Echaveguren, T., and Marcet, J. (2014). Acceleration and deceleration patterns on horizontal curves and their tangents on two-lane rural roads. In *Transportation Research Board 93rd Annual Meeting*, number 14-2627.

Amado, S., Arıkan, E., Kaça, G., Koyuncu, M., and Turkan, B. N. (2014). How accurately do drivers evaluate their own driving behavior? an on-road observational study. *Accident Analysis & Prevention*, 63:65 – 73.

Banks, J. H. (2002). *Introduction to transportation engineering*. McGraw-Hill, London, 2 edition.

Bella, F., Calvi, A., and D'Amico, F. (2014). Analysis of driver speeds under night driving conditions using a driving simulator. *Journal of Safety Research*, 49:45 – 52.

Bender, A., Agamennoni, G., Ward, J. R., Worrall, S., and Nebot, E. M. (2015). An unsupervised approach for inferring driver behavior from naturalistic driving data. *IEEE Transations on Intelligent Transportation Systems*, 16(6):3325–3336.

Blei, D. M., Ng, A. Y., and Jordan, M. I. (2003). Latent dirichlet allocation. *J. Mach. Learn. Res.*, 3:993–1022.

Boer, E. R. (1996). Tangent point oriented curve negotiation. In *Intelligent Vehicles Symposium, 1996., Proceedings of the 1996 IEEE*, pages 7–12.

Bonneson, J. A., Lord, D., Fitzpatrick, K., and Pratt, M. (2006). *Development of tools for evaluating the safety implications of highway design decisions*. Texas, USA: Texas Transportation Institute.

Bonneson, J. A., Pratt, M., Miles, J., and Carlson, P. (2007). Development of guidelines for establishing effective curve advisory speeds. Technical report, Texas Transportation Institute, Texas A & M University System.

Boyce, T. E. and Geller, E. S. (2002). An instrumented vehicle assessment of problem behavior and driving style: Do younger males really take more risks? *Accident Analysis & Prevention*, 34(1):51–64.

Boyle, L. N. and Mannering, F. (2004). Impact of traveler advisory systems on driving speed: some new evidence. *Transportation Research Part C: Emerging Technologies*, 12(1):57 – 72.

Carlson, P., Burris, M., Black, K., and Rose, E. (2005). Comparison of radius-estimating techniques for horizontal curves. *Transportation Research Record: Journal of the Transportation Research Board*, 1918:76–83.

Chandrasiri, N. P. et al. (2012). Driving skill analysis using machine learning, the full curve and curve segmented cases. In *ITS Telecommunications (ITST), 2012 12th International Conference on*, pages 542–547.

Constantinescu, Z., Marinoiu, C., and Vladoiu, M. (2010). Driving style analysis using data mining techniques. *International Journal of Computers, Communications & Control (IJCC)*, 5(5):654–663.

Deffenbacher, J. L., Lynch, R. S., Oetting, E. R., and Swaim, R. C. (2002). The driving anger expression inventory: A measure of how people express their anger on the road. *Behaivour Research and Therapy*, 40(6):717–737.

Donnell, E. T., Gemar, M. D., and Cruzado, I. (2006). *Operational effects of wide edge lines applied to horizontal curves on two-lane rural highways*. Pennsylvania, USA: Pennsylvania Transportation Institute.

EuroFOT (2012). *EuroFOT - Bringing intelligent vehicles to the road*. Available at: `www.eurofot-ip.eu` (2016-01-12).

Fitzpatrick, K., Elefteriadou, L., Harwood, D. W., Collins, J. M., McFadden, J., Anderson, I. B., Krammes, R. A., Irizarry, N., Parma, K. D., Bauer, K. M., , and Passetti, K. (2000). Speed prediction for two-lane rural highways. Federal Highway Administration.

Garber, N. J. and Hoel, L. A. (2009). *Traffic and Highway Engineering - SI Verson.* CL Engineering, 4 edition.

Godthelp, H. (1986). Vehicle control during curve driving. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, 28(2):211–221.

Griffiths, T. and Steyvers, M. (2004). Finding scientific topics. *Proceedings of the National Academy of Sciences, 101 (suppl. 1)*, pages 5228–5235.

Helmers, G. and Törnros, J. (2006). *Effekt av övergångskurvor på förares säkerhetsmarginal samt inverkan av träning : ett försök i körsimulator (Effect of transition curves on drivers' safety margin and the impact of training : an attempt in a driving simulator).* Technical report, Swedish National Road and Transport Research Institute. VTI.

Hong, J., Margines, B., and Dey, A. K. (2014). A smartphone-based sensing platform to model aggressive driving behaviors. *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 4047–4056. New York, NY:ACM.

Hu, W. and Donell, E. T. (2010). Models of acceleration and deceleration rates on a complex two-lane rural highway: Results from a nighttime driving experiment. *Transportation Research Part F: Traffic Psychology and Behaviour*, 13(6):397–408.

Imberg, J. and Palmberg, A. (2015). How curve geometry influences driver behavior in horizontal curves. Master's thesis, Chalmers University of Technology, Sweden.

Ishibashi, M., Okuwa, M., Doi, S., and Akamatsu, M. (2007). Indices for characterizing driving style and their relevence to car following behavior. In *SICE Annual Conference 2007*, pages 1132–1137. Psicataway, NJ: IEEE.

Kong, J., Zhang, K., and Chen, X. (2013). *Digital Human Modeling and Applications in Health, Safety, Ergonomics, and Risk Management. Healthcare and Safety of the Environment and Transport*, chapter Personality and Attitudes as Predictors of Risky Driving Behavior: Evidence from Beijing Drivers, pages 38–44. Springer Berlin Heidelberg, Berlin, Heidelberg.

Krahe, B. (2005). Predictors of women's aggressive driving behavior. *Aggressive behavior*, 31(6):537–546.

Krahé, B. and Fenske, I. (2002). Predicting aggressive driving behavior: The role of macho personality, age, and power of car. *Aggressive Behavior*, 28(1):21–29.

Lajunen, T. and Parker, D. (2001). Are aggressive people aggressive drivers? a study of the relationship between self-reported general aggressiveness, driver anger and aggressive driving. *Accident Analysis & Prevention*, 33(2):243–255.

Lannér, G., Wengelin, A., and Berntman, M. (2000). Kurskompendium väg- och gatuformning (coursecompendium road- and street design). CTH, KTH, LTH.

Li, X., Yan, X., and Wong, S. (2015). Effects of fog, driver experience and gender on driving behavior on s-curved road segments. *Accident Analysis & Prevention*, 77:91–104.

Lin, J., Keogh, E., Lonardi, S., and Chiu, B. (2003). A symbolic representation of time series, with implications for streaming algorithms. In *Proceedings of the 8th ACM SIGMOD Workshop on Research Issues in Data Mining and Knowledge Discovery*, DMKD '03, pages 2–11, New York, NY, USA. ACM.

Lin, J., Keogh, E., Pranav, P., and Lonardi, S. (2002). Finding motifs in time series. In *Proceedings of the 2nd Workshop on Temporal Data Mining, at the 8th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, Edmonton, Alberta, Canada.

MacAdam, C., Bareket, Z., Fancher, P., and Ervin, R. (1998). Using neural networks to identify driving style and headway control behaviour of drivers. *Vehicle System Dynamics*, 29(1):143–160.

Machado-León, J. L., de Oña, J., de Oña, R., Eboli, L., and Mazzulla, G. (2016). Socio-economic and driving experience factors affecting drivers' perceptions of traffic crash risk. *Transportation Research Part F: Traffic Psychology and Behaviour*, 37:41–51.

McDonald, A., Lee, J., Aksan, N., Dawson, J., Tippin, J., and Rizzo, M. (2013). The language of driving: Advantages and applications of symbolic data reduction for analysis of naturalistic driving data. *Transportation Research Record: Journal of the Transportation Research Board*, 2392:22–30.

McLaurin, E., McDonald, A. D., Lee, J. D., Aksan, N., Dawson, J., Tippin, J., and Rizzo, M. (2014). Variations on a theme: Topic modeling of naturalistic driving data. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, 58(1):2107–2111.

Meseguer, J. E., Calafate, C. T., Cano, J. C., and Manzoni, P. (2013). Drivingstyles: a smartphone application to assess driver behavior. *Computers and Communications (ISCC), 2013 IEEE Symposium on*, pages 535–540.

Mitchell, T. M. (1997). *Machine Learning*. McGraw-Hill, New York, USA. ISBN 0-07-042807-7, p.2.

Montella, A., Galante, F., Mauriello, F., and Aria, M. (2015). *Continuous Speed Profiles to Investigate Drivers' Behaviour on Two-Lane Rural Highways*. Transportation Research Record: Journal of Transportation Research Board.

Nie, B. (2006). Effect of horizontal alignment on driver speed behaviour on different road classifications. Master's thesis, Carleton University, Ottawa, Ontario, Canada.

Othman, S., Lannér, G., and Thomson, R. (2009). Identifying critical road geometry parameters affecting crash rate and crash type. In *Association for the Advancement of Automotive Medicine (AAAM), 53rd Annual Meeting - October 4-7, 2009 - Baltimore*.

Othman, S., Thomson, R., and Lannér, G. (2012). Using naturalistic field operational test data to identify horizontal curves. *Journal of Transportation Engineering*, 138(9):1151–1160.

Othman, S., Thomson, R., and Lannér, G. (2014). Safety analysis of horizontal curves using real traffic data. *Journal of Transportation Engineering*, 140(4).

Palmberg, A., Imberg, J., Selpi, and Thomson, R. (2015). The effect of curve geometry on driver behaviour in curves by using naturalistic driving data. In *Proceedings of the 3rd International Symposium on Future Active Safety Technology Towards Zero Traffic Accidents (FAST-zero 2015)*.

Parker, D., West, R., Stradling, S., and Manstead, A. (1995). Behavioural characteristics and involvement in different types of traffic accident. *Accident Analysis & Prevention*, 27(4):571–581.

Passetti, K. A. and Fambro, D. B. (1999). *Comparison of passenger car speeds at curves with spiral transitions and circular curves*. Transportation Research Board.

Pérez-Zuriaga, A. M., Camacho-Torregrossa, F. J., and García, A. (2013). Tangent-to-curve transition on two-lane rural roads based on continuous speed profiles. *Journal of Transportation Engineering*, 139(11):1048–1057.

Quaium, R. B. A. (2010). A comparison of vehicle speed at day and night at rural horizontal curves. Master's thesis, Texas A&M University, Texas, USA.

Radimsky, M., Matuszkova, R., and Budik, O. (2016). Relationship between horizontal curves design and accident rate. *Jurnal Teknologi*, 78(5-2).

Sagberg, F., Selpi, Piccinini, G. F. B., and Engström, J. (2015). A review of research on driving styles and road safety. *Human Factors*, 57(7):1248–1275.

Sullivan, J. M., Bao, S., Goudy, R., and Konet, H. (2014). Characteristics of turn signal use at intersections in baseline naturalistic driving. *Accident Analysis & Prevention*, 74:1–7.

Syed, L. (2005). Experimental investigation of vehicle's lateral acceleration on highway horizontal curves. Master's thesis, Carleton University, Ottawa, Ontario, Canada.

Towfic, I. Z. (2014). A method for classifying driver performance. Master's thesis, University of Windsor, Canada.

Wallman, C.-G. (1998). Driver behavious on winter roads: A driving simulator study. In *Xth PIARC International Winter Road Congress*, volume 3.

Zhang, Y., Lin, W. C., and Chin, Y. S. (2010). A pattern-recognition approach for driving skill characterization. *IEEE Transactions on Intelligent Transportation Systems*, 11(4):905–916.