CHALMERS | UNIVERSITY OF GOTHENBURG

Pattern extraction to define normal driving from driver-rated data using data mining techniques

Master's Thesis in Engineering Mathematics and Computational Science

Tobias Karlsson

Department of Mathematical Sciences
Chalmers University of Technology
Gothenburg, Sweden 2013

Master's Thesis 2013:1

Abstract

In this thesis driver rated data is studied using data mining techniques. The rated data consists of roughly 72 hours of data from seven drivers. The ultimate goal is to be able to identify patterns of high rating and match them towards a reference database consisting naturalistic driving data. Two segmentations of the drives are used, equilength subsegments and steering operations. An alternative morphed standardised rating scaled is proposed.

Two data mining approaches are applied. The first method is based on using an ensemble classifier on features derived from the CAN-data to predict the rating of each segment of the data. The second method uses an outlier detection algorithm and a hierarchical clustering approach on a distance metric based on the angles between the principal variance components of the observations.

Using the ensemble classifier and general variables a large proportion of rating variance can be explained when including driver and route factors. Large rating values can be identified well. For the standardised rating the prediction of high values is worse with many false positives.

The matching of signals using the covariance structure works well. Using hierarchical clustering clusters with standardised rating high above average can be obtained. Outliers with high standardised rating are extracted and matched towards a larger database. The matches are few but similar to the original situations owing to the fact the matching is strict.

In conclusion the ensemble classifier works well for predicting rating when driver and route factors are included. The covariance-based method performs well for situation matching and clusters with high rating can be identified. It also has potential to be be used for extracting and matching more sofisticated patterns.

Acknowledgements

First and foremost I would like to thank everyone at Volvo Cars Corporation for the opportunity to work an exciting project and gain some insight into industrial research. Especially I would like to thank my supervisor Magdalena Lindman for always taking time to answer my questions and giving advice. A great thanks also goes to my supervisor and examiner at Chalmers, Olle Nerman, for providing insight and experience during the time of this project. Least but not last a great thanks to my friends who have provided both input and support throughout my time working with this thesis.

Tobias Karlsson, Gothenburg, 2013-05-31

Contents

1

Introduction

Car accidents is one of the leading causes of deaths and injury worldwide. The area of traffic safety is widely studied and the research field is getting ever more diverse. Starting from passive safety and damage mitigation research the field has developed to incorporate collision avoidance and driver assistance systems. The trend is that the safety systems are designed to reduce risk prior to accidents and intervene at ever earlier stages in the course of the accident. In order to design and implement such systems an understanding of driving behaviour is needed. This thesis aims to define boundaries for normal driving using driver rated data. By studying driving situations of severity lesser than incidents. To do this driving data rated on a scale from 1 to 100 is used. The ultimate goal is to be able to identify situations where the driver is experiencing unease. Once such situations are identified systems could be developed to assist the driver to get from that state of unease back into states of comfort.

1.1 Background

Motor vehicle traffic lies at the very heart of our society, people depend upon it for their daily transports as well as logistics on a larger scale. Car ownership statistics show that the number of cars per capita increases worldwide - and the trend is predicted to continue[1]. The immense scale of the worldwide vehicle fleet means that a relatively small risk of injury per kilometer traveled leads to a large absolute number of injurys and deaths. Each year over 1.2 million people lose their lives in traffic accidents - and only a fraction of victims perish. This places car accidents as 11th leading cause of death and 9th leading cause of disability-adjusted life years lost according to the WHO[2]. Apart from the tragedy of lost lives there are large monetary costs coupled with car accidents. The estimate differs by country but in a 2007 report an estimate based on statistics from NTSHA the cost of accidents in the US is estimated to be over 4% of the total GDP[3]. This figure includes destruction of physical values, medical costs and lost worktime due to resulting congestion. Increasing the safety, improving traffic flow and decreasing congestion in traffic is a high priority both for car manufacters and policy makers around the world.

The development of vehicle safety systems is moving backwards in the collision phase. The first safety measures taken where all post-crash and damage mitigation, classically called passive safety. Such systems include seatbelts, padded dashboards and safety cages. Subsequent systems focused on assisting the driver to carry out desired operations in a better way. Examples of such systems include anti-lock breaking systems (ABS). Today state of the art systems are in collision avoidance such as pedestrian detection and automatic breaking. In order to move active safety systems away from the domain of imminent crash scenarios into the domain of normal driving a thorough study of normal driving conditions is necessary. To meet this need Naturalistic driving studies are performed. These studies are performed by equipping everyday cars with measurement equipment and then allowing people to use the cars as they normally would. Such datasets include all types of driving data as there are no constraints on driving conditions in the set-up. Performing the data collection in this way means that the data gathered have increased representability and causes less bias in data gathering.

1.2 The data

For this thesis two datasets will be used, and are called the Volvo Human Monitoring dataset (VHM) and the EuroFOT dataset. Both consist mainly of Controller Area Network (CAN-data) and video data but have some other differences as explained below. The CAN-data is made up of signals that describe the state of the car. These signals are measured in while driving and are recorded in a integrated computing unit. Examples of signals are brake pressure, vehicle speed, steering angle and GPS-data. Three signals from the CAN-data plotted as a function of time can be seen in Figure (1.1). All CAN-data used for this study is sampled at 10 hertz.

1.2.1 Volvo Human Monitoring data

The first and most important dataset that will be used for this study is the Volvo Human Monitoring (VHM) dataset. The data consists of a total of 27 drives by 7 different drivers, each drive lasting approximately three hours. The drives where scheduled with a mapped route to drive. The 10 routes where designed to include various types of driving. The routes are all in the same geographical area around Gothenburg. This means that the data is not naturalistic driving data. This data was gathered with the goal to measure and understand the origin of driver workload, cognitive and physical, during different driving conditions. The distributions of the number of drives per driver and number of drives per route is seen in Figure (1.2). After driving the drivers where shown video data from the drive and rated the experienced difficulty of every situation on a scale from 1 to 100, where 100 represents not being able to talk on the phone while driving. The

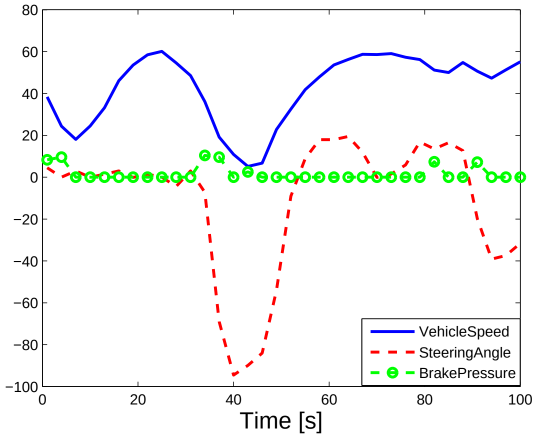


Figure 1.1: Three signals from the CAN-data plotted against time. The y-axis is in different units for the different signals.

length and value of each rated segment was decided by their own choosing. These driver-specified segments are called *Original segments*. In this work this rating data is used as a reference for normality of situations. The original segments are adjacent in time and can be depicted as a time signal with discrete jumps henceforth called the *individual unease rating function* (IURF), also called just rating. There also exists interview data with a driver comment for each original segment.

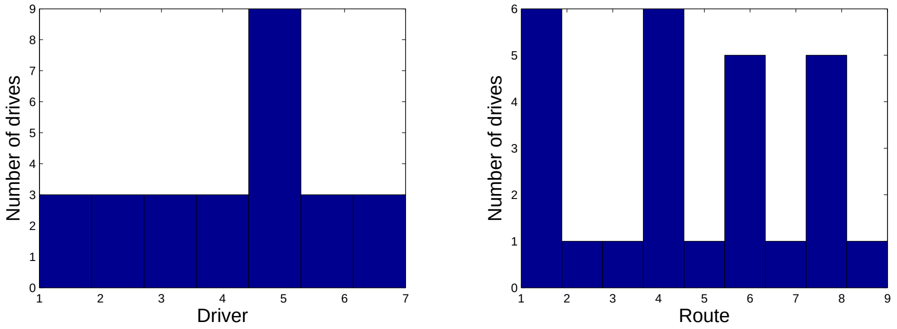


Figure 1.2: The distribution of drives per driver and drives per route.

1.2.2 EuroFOT data

This dataset contains naturalistic driving data. The data was gathered by equipping 100 cars with measurement equipment for CAN- and video data. Data were collected as the cars were used as everyday cars by families in the Gothenburg area for one year and consists of over 40,000 hours of data. We will use the EuroFOT data as a reference for matching patterns that are extracted in the VHM data.

1.3 Objective

The overall aim of the project is to identify situations in which the driver is experiencing unease in the VHM database and investigate if similar situations can be found in the EuroFOT database. The measure used for unease is a high rating scale value. In order to do this the following subgoals are set:

1. Develop methodology to extract common patterns of uneasy situations from the VHM data.
2. Investigate partitionings of the data that can quantitatively be compared using suitable variables.
3. Matching extracted patterns from the VHM data with the EuroFOT dataset to evaluate the occurence and representability of extracted patterns.

1.4 Thesis outline

The work is structured as follows:
In chapter 2 some work on similar problems is presented. In chapter 3 the general methodology for this thesis is presented. One method based on ensemble classifiers is used to investigate the possibility of training a classifier using general variables to identify patterns of high unease over the whole dataset. The second method is an agglomerative clustering method with distances based on comparing the direction of variances in the space spanned by the data. Chapter 4 presents the results of the methods. Chapter 5 offers a concluding discussion on the result, the limitations of the data and the methods.

2

Similar work

This chapter gives a brief description of work in similar areas and presents some previous studies.

2.1 Naturalistic Driving studies

The first large-scale naturalistic driving study was performed by Virginia Tech Transportation Institue. In the study 100 cars where equipped to measure CAN-data and given to everyday drivers. This data is public and avaliable at the institutes website[1]. The study got a lot of attention as it was the first of its kind and introduced the option to study all phases of driving.

The main part of the analysis was annotation-based and focused on driver inattention. Such studies are in the general case based on defining a boundary or subset of all driving conditions and annotating the data into discrete classes. The annotated data is analyzed for differences between the cases present.

One example is the study of relative risk of near-crash and crash under cellphone use compared to a control group of no secondary task engagement. In the study the accident and near-accident frequencies are statistically compared between the two groups.

2.2 Other work

An overview of some studies with similar goals are prestented below. A common factor in these studies is that they are performed on data that is geographically or situationally controlled.

[1] http://forums.vtti.vt.edu/index.php?/forum/13-100-car-data/

In a study[6] Takeda et. al used a composition of braking dynamics and driver voice utterances to detect annotated hazardous situations and extreme brakings. The method is based on isolating the operations and using two-dimensional bins of brake pedal pressure and break pedal pressure rate integrated with the energy of the voice utterances to compare the brakings. The classification rate was high, around 95% with a false positive rate of 5%. To compare it with this study however, that data was geographically confined to a small area with a large set of data and voice utterance data is not avaliable here.

A number of studies have been performed using Hidden Markov Models (HMM) for trajectory recognition. HMMs are commonly used for trajectory tracking, where one example is *Clustering Vehicle Trajectories with Hidden Markov Models Application to Automated Traffic Safety Analysis*[7]. This study is also limited to a geographical area with a larger dataset. These models are very assumption-heavy, making them ill-suited for problems with a large number of unknown states. Especially as the number of parameters of the model grows exponentially with the number of possible states[8].

3

Method

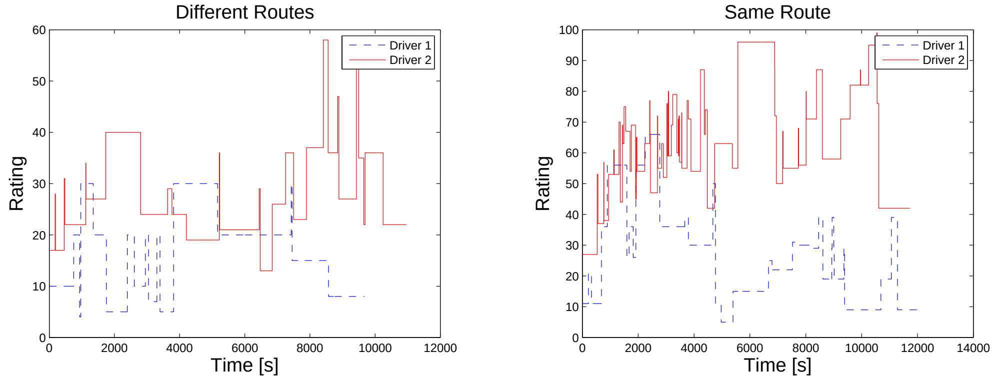
This chapter is focused on describing the implemented process of obtaining patterns of unease from the VHM database which is the main body of this work. Before defining the method a deeper description of the data is given. The signals used are described and a brief overview of the workflow is also given. Two different methods are tested: One of general classification using an ensemble classifier to investigate the possibility of training a classifier to identify patterns of unease. A second approach based on comparing the direction of data variance using principal components is also implemented. The events extracted from this algorithm is then used to extract matching patterns in the EuroFOT database.

3.1 Individual unease rating function and VHM statistics

This section provides a description of the Individual unease rating function (IURF) and shows some basic statistics for it.
Examples of the IURFs for some drives are shown in Figure (3.1). The distribution of rating values per driver as a box-plot is shown in Figure (3.3). There one can see that different drivers use the rating very differently, especially the value range varies a lot, but also the variance. The distribution of length of the original segments over all the drives can be seen in Figure (3.2). A majority of the original segments are shorter than 2 minutes and among those a majority is shorter than 40 seconds.

The distribution of rating changes per time unit as a function of speed can be seen in Figure (3.2a). The fraction of event changes per time unit is larger at lower speed. This can be interpreted as if there is a larger rating variability at lower speeds.

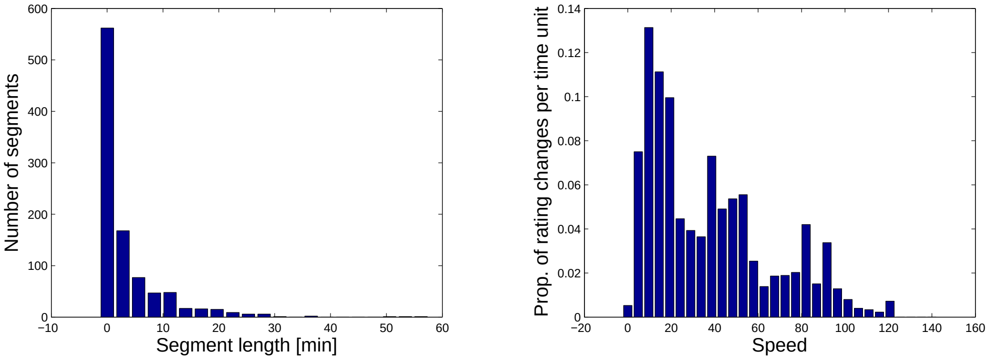
As is seen in Figure (3.3) there is a large between-subject variation in the usage of the rating scale. The maximum values for driver 2 and 3 falls within the lower quartile of driver 1 who uses almost the entire range of the rating scale. Arguably the definition



(a) Different drivers different routes

(b) Different drivers same route

Figure 3.1: The individual unease rating function (a) from two different drives. In (b) the individual unease rating function for two different drivers driving the same route is shown.



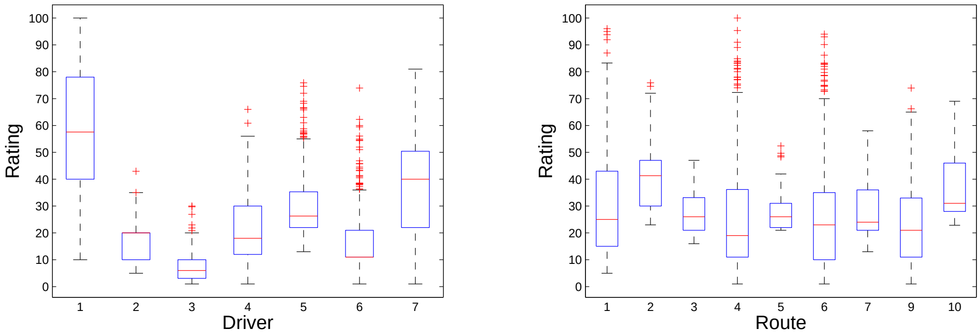
(a) Distribution of original segment lengths

(b) Distribution of rating changes per time unit over speed

Figure 3.2: (a) The distribution of original segment lengths for all drives. Almost 60% of the original segments last for under 2 minutes. (b) The number of rating changes per time unit as a function of speed.

of the rating scale to the drivers as ”value 100 represents not being able to talk on the phone while driving” is heavily subjected to personal interpretation. The sample pool of drivers consists of 7 drivers. Therefore the possibility to draw any generalizable conclusion regarding the usage of the rating scale between drivers is highly limited.

Without a significant effort in studying the behaviour of individuals on similar scales there is no way of knowing if an individual with higher maximum value actually percieves a higher effort or only percieves the scale differently. However a basic assumption must be that within drivers the values are coherent i.e. a rating value above average for



(a) Boxplot of rating per driver

(b) Boxplot of rating per route

Figure 3.3: A boxplot of the rating value distribution per driver and route. The value range used by the drivers differ a lot.

an individual correlates with a higher than average effort for that driver.

To remove the different value ranges between individuals a morphing of the rating scale that retains the within individual relations is proposed: a standardised rating scale.

3.1.1 Within driver normalisation - standardised rating scale

To rescale the value ranges for the individuals a standardisation is done. If we consider the mean value of the individual unease rating function for a driver to be the baseline for that person it is logical to rescale the individual to equal means. To handle the disparity of variances within drivers the rating values are divided by the within driver standard deviations. This yields a standardised rating scale where values above zero represents higher than average rating and the disparity of rating values within drivers being roughly equal. This effectively removes the between driver variance in individual unease rating function. A boxplot over the standardised rating per driver is seen in Figure (3.4).

More formally the standardisation is done by taking the original rating values R_{ij} for driver i and segment j and forming the morphed values $\hat{R}_{ij}$ by:

$$\hat{R}_{ij} = \frac{R_{ij} - \overline{R_{i\cdot}}}{\sqrt{\frac{1}{N_j}\sum_j (R_{ij} - \overline{R_{i\cdot}})^2}} \tag{3.1}$$

where $\overline{R_{i\cdot}}$ denotes the mean rating value of driver i and N_j the number of segments for driver j. The standardised rating is what is commonly known as the z-score or the standard score of the rating.

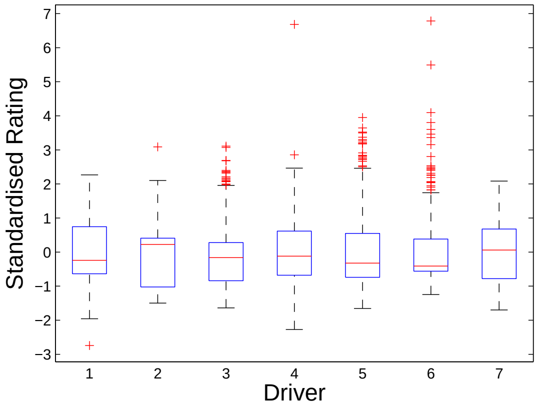


Figure 3.4: A boxplot over the standardised rating per driver. All distributions have equal means and standard deviations.

3.2 CAN-data Signals

A number of signals from the CAN-data are used, henceforth called *arguments*. A brief list with corresponding explanation for the signals is given below. All derivatives used are with respect to time. An overview of a coordinate system fixed at the center of the car describing the signals relation to the car is shown in Figure (3.5).

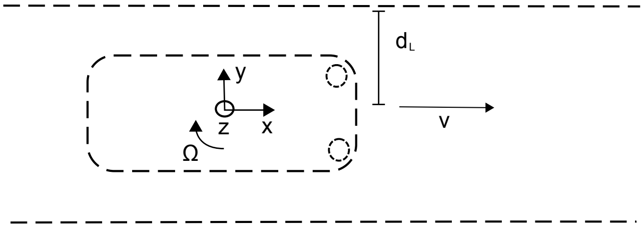


Figure 3.5: An overview of the coordinate system fixed at the center of the car. The rotational angle in the plane of motion, called yaw, is measured clockwise around the z-axis and is denoted Ω. The distance to the left lane marker from the center of the car is denoted d_L.

AbsSteerAng - The absolute value of the SteeringAngle, degrees from the central position.

AbsSteerAngRate - The derivative of AbsSteerAng, observe that this is not equal to the absolute value of SteerRate seen below.

AccPedalPos - The position of the accelerator pedal.

AccPedalRate - The derivative of the AccPedalPos.

BrakePressure - The pressure applied on the brake pedal.

BrakeRate - The derivative of the BrakePressure.

LateralAcc - The lateral acceleration, measured in the axis perpendicular from front to rear of the vehicle. In the coordinate system of the car in Figure (3.5) this is $\ddot{y}$.

LongAcc - The longitudinal acceleration, measured in the axis from front to rear of the vehicle. In the coordinate system of the car in Figure (3.5) this is $\ddot{x}$.

LeftLaneOffset - The distance to the left lane marker, depicted as d_L in figure (3.5).

SteeringAngle - The steeringwheel angle.

SteerRate - The derivative of the SteeringAngle.

VehicleSpeed - The speed of the vehicle.

YawRate - The rotational speed of the vehicle, proportional to the VehicleSpeed times the SteeringAngle. The time derivative $\dot{\Omega}$ of Ω in Figure (3.5)

3.3 Knowledge discovery in databases

In the last three decades the avaliability of data has increased exponentially, following the exponential evolution of data storage- and processing capacity described by Moores Law. At the end of the 1980's a database of size 1 megabyte was considered large, 10 years later databases with size order in the terrabyte domain were not unusual[4]. The estimated increase in avaliable information is estimated to double every 20 months[5].

In many areas today the entire state of a system can accurately be measured over a long period of time. With the growing amount of avaliable data following the cost reduction in data gathering, storage and processing new ways of studying systems have emerged. A few decades ago collecting reliable data was a limiting factor in studies. Consequently the data gathering and data analysis tools where often designed to measure specific phenomena. With this increase in dataset size a new interdisciplinary area known as knowledge discovery in databases has quickly gained popularity [4]. The main focus is to identify unknown patterns in the data with the objective to extract knowledge about the system in order to understand mechanics, predict future behaviour and design strategies to control the system. The steps of the process is depicted as a flowchart in Figure (3.6).

For this work two separate data mining methods are applied. For each method the steps in figure (3.6) are applied.

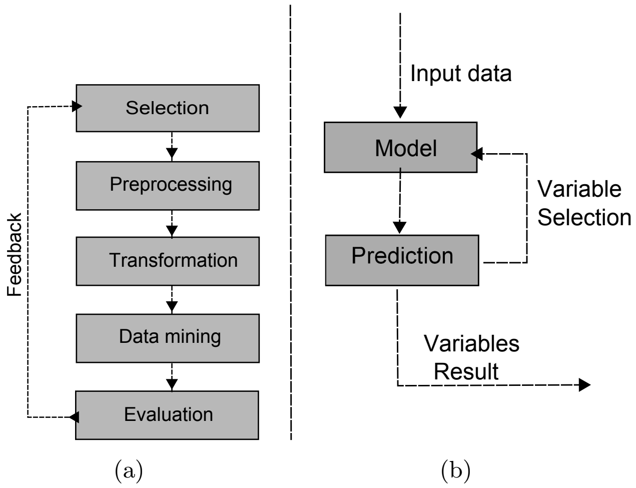


Figure 3.6: (a) A flowchart of the KDD process, it includes the general steps of the knowledge discovery process with an added feedback. The feedback consists of adjusting the model by adding new assumptions according to previous results. (b) Shows the details of the data mining step in (a).

3.4 Approach 1 - Ensemble Classifier

The first part of the thesis is focused on a general model for estimating the rating given the CAN-data. Another separate approach with a separate problem formulation is given in the next section.

The driver can be seen as a system reacting to the environment and providing an output in the form of the rating. A schematic of this is shown in Figure (3.7). The modelling process consists of two steps. First input data x for the model must be derived from the representation of reality, the CAN-data. These are the selection and preprocessing steps. Then the input data x is mapped by a function $f(x)$ to a response $\hat{y}$. The success of the model is measured by the difference in predicted response $\hat{y}$ and the true response y. The model is evaluated both for regular and standardised rating.

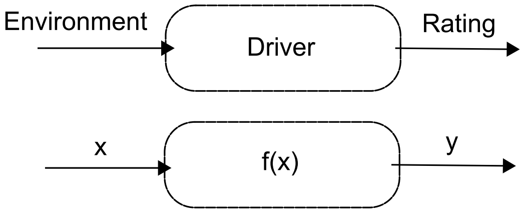


Figure 3.7: A schematic of the real process and the estimation (modelling) of the process.

3.4.1 Data preprocessing

The data preprocessing procedure is divided into two steps. First the drives are divided into subsegments and each subsegment is taken to be one observation. For the subsegments a set of predictor variables are derived, these are taken to represent reality and are used as the input of the model. These predictor variables are used to train and evaluate the model. A description of this process and the segmentation are given below.

Original segments

The first segmentation are the original segments as defined by the drivers. These segments are of varying lengths as is shown in Figure (3.2). One drive with original segments marked with vertical dashed lines is shown in Figure (3.9a). The response for each original segment is the rating value set by the driver.

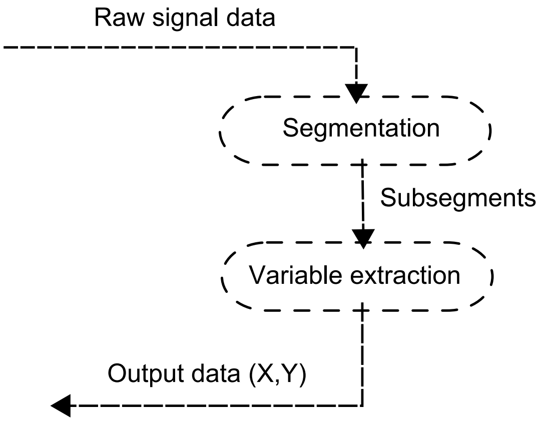


Figure 3.8: A flowchart over the data preprocessing. The two main steps are data segmentation and variable extraction. The input is the raw signals and the output is a feature vector $\vec{x}$ and a response y for every subsegment.

Equilength subsegments

An alternative segmentation to the original segments is the division of the drives into equilength subsegments of chosen length dt seconds. Each such segment is taken as one observation and the rating is formed as the average of the rating values over the segment. The partitioning of one drive into such segments can be seen in Figure (3.9b). This means effectively ignoring the original segmentation defined by the driver and looking for general correlations between driving conditions and rating. For a given drive of length T the segments are given by:

$$\left\{[0,dt[,...,[(i-1)dt,idt[...,[(N-1)dt,Ndt[,[Ndt,T]\right\}$$

$$N = \left\lfloor \frac{T}{dt} \right\rfloor.$$

Where $\lfloor \cdot \rfloor$ is the floor operator.

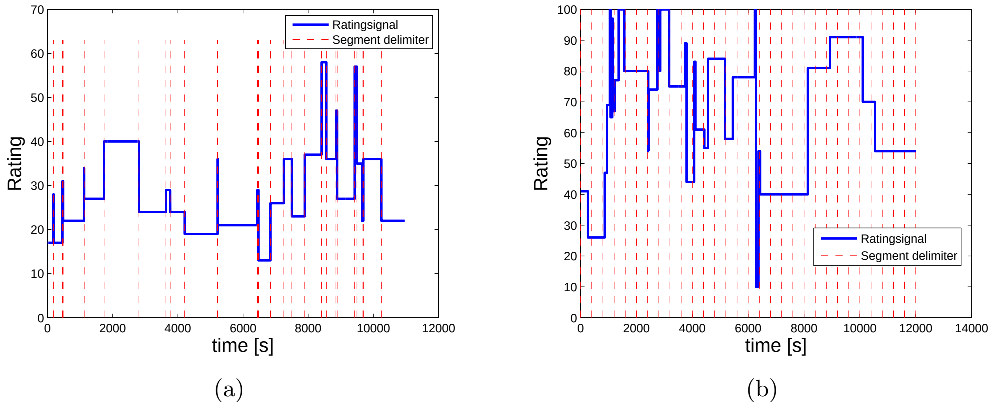


Figure 3.9: (a) The individual unease rating function (blue) and markings of start and end of the original segments (dashed vertical lines). (b) The individual unease rating function for one drive (solid) and the division of it into equilength subsegments of length $dt = 400$ [s].

3.4.2 Predictor variables

The subsegments are now defined by a time interval $[t_1,t_2]$. For each subsegment a set of real-valued predictor variables are derived.

Driver data

As is shown Figure (3.3) there is a large difference between drivers and routes for the value range and usage of the individual unease rating function. That is, a large between-subject variance which must be accounted for. The individual effects are included in the model with the following categorical features:

1. Driver index
2. Gender
3. Route

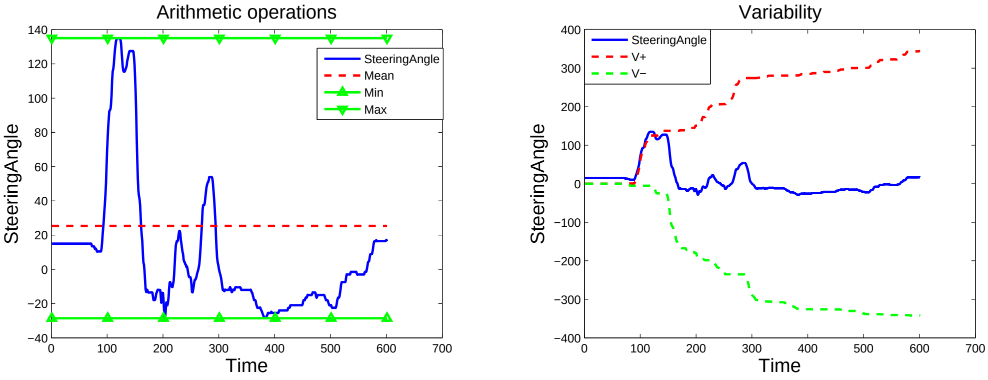
These features are included as predictors, coded by the driver number {1,2,..,7}. The routes are coded by the route number {1,2,...,10}. The routes contain the same proportion of different driving conditions but differ in the layout.

Regulars

The variables for a segment $[t_1,t_2]$ are derived by applying some arithmetic operations on the *arguments* over that segment. This includes the mean, max and min value as:

$$[\text{mean}(argument([t_1,t_2]))\ \text{max}(argument([t_1,t_2]))\ \ \text{min}(argument([t_1,t_2]))\ \ ]$$

The mean value is a description of the activation in the argument over the segment. The max and min values represent the most extreme occurences during the segment. Some signals have a range including both positive and negative values. In those cases the mean of the absolute value of the signals is used. For example using the absolute value of the argument *SteeringAngleRate*, then the mean value represents the average degree of activation of the steering wheel for the segment. An example of these operations applied to signals is shown in Figure (3.10).



(a) (b)

Figure 3.10: Examplification of the variables described in section (3.4.2). The right plot (a) shows the result of applying arithmetic operations and the left (b) shows the results of calculating the variability. The segment length is one minute.

Variability

The variability of a segment $[t_1,t_2]$ is a measure of the total amount of change in the argument. This is calculated as the piecewise differences over the smallest time intervals, corresponding to the sampling frequency, of length Δ:

$$\delta_i = \text{oArg}(t_1 + (i+1)\Delta) - \text{oArg}(t_1 + i\Delta)$$
$$i \in \{1,2,...N-1\};\ N \text{ s.t } t_1 + N\Delta = t_2.$$

Then all positive increases are summed to form the positive variability V^+:

$$V^+(t_1,t_2) = \sum_{\delta_i > 0} \delta_i \tag{3.2}$$

in a similar fashion we can form V^- as:

$$V^-(t_1,t_2) = \sum_{\delta_i < 0} \delta_i. \tag{3.3}$$

The variability, postitive and negative, for one signal is shown in Figure (3.10). The difference $V^+ - V^-$ is the total activation in the argument over the segment.

Arguments	Mean($\|\cdot\|$)	Max($\cdot$)	Min($\cdot$)	Mean($\cdot$)	Variability
VehicleSpeed		x	x	x	x
AbsSteeringAngle		x	x	x	x
AbsSteerRate		x	x	x	x
BrakePressure		x	x	x	x
BrakeRate		x	x	x	x
YawRate	x	x	x		x
LateralAcc	x	x	x		x
LongAcc	x	x	x		x
RightLaneOffset	x	x	x		x
Variable					
Proportion still					
Proportion reversing					
Proportion accelerating					
Proportion decelerating					

Table 3.1: The variables included for the equilength and original segments. MeanAbs denotes taking the absolute value of the signals before performing the mean operation. Fields marked x indicates variable obtained using operation included. The bottom variables are the proportion of the segment that those operations are active. When segments are of unequal length the variability is divided by the segment length to form the variability per time unit.

3.4.3 Decision trees

The model used for analyzing this data is an ensemble classifier of bagged decision trees in the form of classification and regression trees (CART). The specific implementation used is MATLABs *'TreeBagger'* function. A table over the variables used for this analysis is seen in Table (3.1). The classifier is evaluated with regards to both the rating and the standardised rating.

Classification and Regression Trees (CART)

The CART classifier operates by making a number binary of cuts of the n-dimensional space defined by the n predictor variables. These are also called decision splits. In two-dimensional space this can be seen as cutting the whole space into two regions estimates rating by the average rating in each region. Every new observation is thus assigned as having the rating of the majority in the region it is assigned to. For continous response variables the corresponding majority vote is the mean value of the response of the observations in the region. The regions are split serveral times giving a finer partitioning of the predictor variable space. This is applied until no decision split can be made that separates the data such that the error is reduced. The reason for chosing this classifier is that it returns a variable importance measure which increases interpretability. Further with bagging it has small classification bias and works well with variable selection[9].

Ensemble classifier

The ensemble classifier is a scheme that is applied to counter bias from the classifier resulting from over- and underfitting. Instead of training only one classifier a number (N) of classifiers are trained. The output is the weighted average of all the classifiers trained. The result is a less biased classification that is often more accurate than the results of a single classifier.

The ensemble classifier uses bootstrap aggregation. For each individual tree trained a sample X' is bootstrapped uniformly from the original sample X. This introduces a randomness in the construction of every decision tree that reduces over- and underfitting that may occur when using single classifiers on all data. In this application the bootstrapped sample X' is of the same size as the original sample X. An effect is that it is possible to form the out-of-bag training error (OOB err.) while training. The OOB error is formed by evaluating the classifier on the unseen observations that lie in X but not in X'.

The number of variables used for each decision split are a settable parameter for the method. In this application all variables are used at every split. This is motivated by that variable selection will be applied removing variables that are not useful.

As a complement to the out of bag error during training 20% of the data, chosen randomly, is held out for final validation. The final error η is measured as the average quadratic distance between the real rating value R_j of the observation and the estimated rating value $\hat{y_j}$ for the held out data:

$$\eta = \frac{\sum_j (\hat{y}_j - R_j)^2}{N_j} \tag{3.4}$$

where N_j is the number of observations in the validation data. This is compared to the variance of the rating for all observations which is the worst the model could perform.

The model is run and evaluated both for the normal rating and the standardised rating.

Variable selection

To select valuable variables and remove variables containing no information a forward variable selection scheme is applied. Each variable is assessed individually and the best one is chosen. Iteratively every new variable is evaluated with the previous chosen variable(s) and the best variable is added. This is repeated until the error increases or the maximum chosen number of variables are obtained.

3.5 Approach 2 - Situation matching with signal covariate structure

As a complement to the general model above a separate more situation-specific approach, based on indirect matching of signal covariate structure, is applied. Similar methods are generally used for multivariate time series clustering and pattern recognition, for instance in climate models [11]. SVD-based methods for analysis of similar, but unrated, data has also been proposed[12]. An agglomerative clustering approach is applied to detect clusters. Further, as the data is sparse in comparison to the large variation of situations that exists during driving and many situations occur only once, an outlier detection algorithm is applied to detect observations that are significantly different from the rest of the data. There is no guarantee that deviating situations have a high rating. Therefore a rating threshold is applied on the filtered situations. Once these situations are identified they can be matched to situations in the larger EuroFOT database and the matches are manually evaluated.

This method is applied for two different segmentations: The equilength segments described above. The second segmentation is into steering operations.

3.5.1 Steering operations

An alternative segmentation is used namely steering operations. Steering operations are filtered out and used as segments. There is practically always activation to some degree on the steering-wheel. This is an effect of road curvature and the need to correct the vehicles lane position. Therefore, a minimum threshold of 15 degrees of activation is used for filtering steering events. One example of two isolated steerings can be seen in Figure (3.11).

3.5.2 Covariance structure comparison

For these segmentations the covariance structure of signals is studied. The covariance structure can be described as the relation between the signals. For this the following signals are used:

Vehicle Speed

Brake Pressure

Brake Pressure Rate

Steering Angle

Steering Angle Rate

LongAcc

LateralAcc

YawRate

AccPedalPos

AccPedalRate .

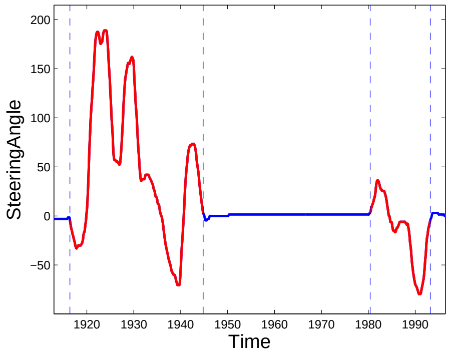


Figure 3.11: Two separate steering events with dashed line as event delimiters.

The signals have very different value ranges, therefore a normalisation is done. As in Section (3.1.1) standard score is used for each signal. In this case the standard score is global, i.e. each value for a signal is standardised with respect to every other value for that signal in the dataset. For each segment a matrix X_i with the respective signals as columns is formed:

$$X_i = \begin{pmatrix} x_{1,1} & x_{1,2} & \cdots & x_{1,n} \\ x_{2,1} & x_{2,2} & \cdots & x_{2,n} \\ \vdots & \vdots & \ddots & \vdots \\ x_{m,1} & x_{m,2} & \cdots & x_{m,n} \end{pmatrix} \tag{3.5}$$

where m is the number of timesteps and n is the number of signals. This matrix X_i can be decomposed into three separate matrices using singular value decomposition (SVD) as:

$$X_i = U_i S_i V_i^T. \tag{3.6}$$

The numerical details of the singular value decomposition are lengthy and are not given in detail. These can be found in virtually any book on linear algebra, where one example is the book by Lay[10]. The matrix U_i contains the left singular values, these are not of interest here and will not be explained further. The matrix V_i contains the right singular values. The columns of V_i are orthonormal eigenvectors forming a basis for the n-dimensional space spanned by the colums of X_i. The eigenvectors are such that the data variance in their direction is maximized. These are the principal components of the data. These can be viewed as the directions of the variance in the data. Matrix S_i is a diagonal matrix where each diagonal element $S_i(k,k)$ is proportional to the square root of the variance in the direction formed by the eigenvector given by the k:th column of V_i.

The measure of similarity for two event matrices X_i and X_j is given by the comparison of the direction of the eigenvectors. Meaning that if the direction of variance for the two matrices are equal they are considered equal. Similar events have similar variance structure in n-dimensional space and thus the corresponding eigenvectors should be in the same direction.

To reduce the dimensionality and only include eigenvectors with a large proportion of variance connected to them the eigenvalues are studied. The distribution proportion of total variance per eigenvalue for the equilength segments can be seen i Figure (3.12). After the fourth eigenvalue little variance remains. Therefore only the first four eigenvectors are used for the comparison.

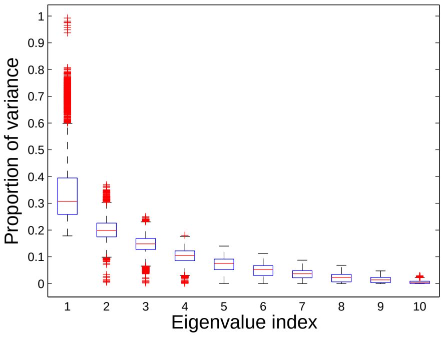


Figure 3.12: The distributions of proportional variance per eigenvalue after the singular value decomposition. Generally after the fourth eigenvalue very little variance remains. Therefore only the four first eigenvectors are used for comparing bases.

The measure of similarity for two matrices is formed by comparing their respective bases given by V_i and V_j. Formally the similarity $S(i,j)$ of the two bases is given by the sum of the squares of the cosines of the smallest angles between all the eigenvectors. This is calculated as:

$$S(i,j) = \frac{1}{4}\text{trace}((V_i^T V_j)(V_j^T V_i)). \tag{3.7}$$

The value $S(i,j)$ lies in the range of [0,1]. It is informally described as total difference between the directions of variances between the two events. From the similarity measure a distance matrix D can be calculated where each value $D(i,j)$ is the distance between event i and j. The elements are given by:

$$D(i,j) = \frac{1}{S(i,j)}. \tag{3.8}$$

This gives the pairwise differences in coviariate structure between the all the events in the data. These pairwise distances are used to form clusters using agglomerative clustering and detect outliers in the data.

3.5.3 Clustering

To detect clusters an agglomerative clustering approach is applied. The hierarchical tree is formed using complete linkage. Meaning that for every observation the distance to every other observation is used when forming the tree. The tree is pruned top down into N subclusters. For each cluster the mean standardised rating and number of observations are calculated and extracted.

3.5.4 Outlier detection

To detect outliers a simple scheme is used. For each observation i the average distance to the 5 closest neighbours $d_5(i)$ is formed:

$$d_5(i) = \frac{1}{5} \sum_{j \in \min_5(i)} D(i,j). \tag{3.9}$$

Where $\min_5(i)$ denotes the set of the 5 closest neighbours to i. An outlier is then defined as a point with a large distance $d_5(i)$.

3.5.5 Situation matching

The extracted situations from the outlier detection algorithm are used for situation matching in the EuroFOT database. The same distance measure as in equation (3.8) and (3.7) is used. The minimum distance between two events X_i and X_j is 1. The following criteria is used for a match:

$$\text{dist}(X_i,X_j) - 1 < \epsilon \tag{3.10}$$

where epsilon is chosen to be 0.005.

3.6 Other models tested

In complement to the methods above some other possible methods and methodologies have been investigated. A small overview of these methods are presented below. However those methods where either such that the basic assumptions proved to not be met or the results where not good enough to be worthy of full presentation.

3.6.1 Situationally limited study and Logistic Regression Models

The initial idea was to limit the study situationally. The situations considered was those revolving around pedestrian crossings. After annotation and filtering it turned out that the number of such situations in the VHM database was 16 and all situations occured at different geographical areas. The possibility of fitting a logistic regression model to those situations was studied and passages over the same areas in the EuroFOT database was filtered out using GPS coordinates. The occurence of such situations in the EuroFOT database was also low. With a limited rated sample and a limited amount of complementary unrated situations this method was deemed a dead end.

3.6.2 Mixed Models

Mixed models are a common approach for studying treatment effects in clinical trials. The model has great flexibility as it includes both individual and common effects between subjects. All unexplainable variance is modeled as an error term that is assumed to be normally distributed. This allows for identifying effects in the chosen variables on serveral levels and the effects that cannot be accounted for by the fixed effects are accounted for in the random effect.

Different segmentations such as the equilength subsegmentation was considered. At implementation the common effect for various variables, such as speed, was small and insignificant. The model showed some significance for individual effects but the error terms included almost all of the total variance and did not meet requirements for normal distribution so no statistically sound conclusions could be drawn.

3.6.3 Extreme values

The third and last of the methods applied was focused on extreme values. The main idea was to study extreme occurences in the different arguments. The assumption was that extreme argument values (for instance upper x%) at different speeds would correlate with deviating situations that would correspond to high rating. For example the minimum longitudinal acceleration as a function of speed was studied. The extreme values corrected for speed showed little or no correlation to high rating values.

4

Results

In this chapter the results for the respective methods are presented.

4.1 Approach 1 - Decision trees

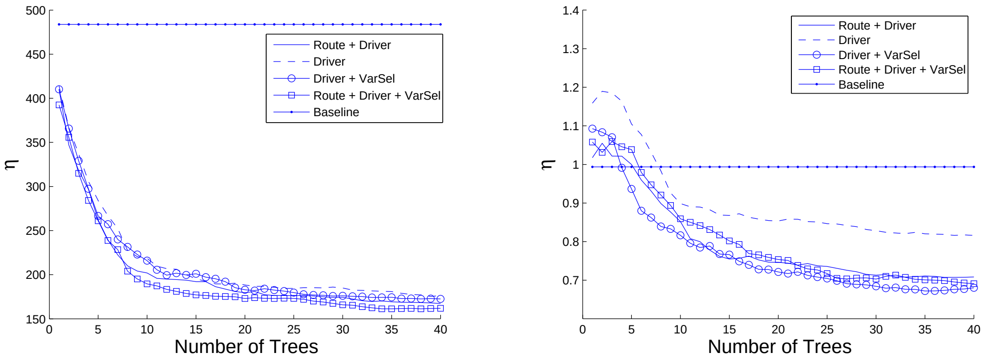
The decision trees algorithm was run on the original segments and the equilength subsegments of length $t = 60$ seconds. The variables included in the process are shown in table (3.1). The plots show the OOB error and any figures given are from the validation data. The OOB error is very close to the validation error in all cases.

4.1.1 Original segments

The results for the original segments are presented here. In total there are 950 such segments. The results from the decision trees on the original segments before and after the variable selection is shown in figure (4.1). The original number of variables are 49, the number of variables after variable selection are 5. With all variables included the final error is 166, which is not reduced notably by variable selection.

4.1.2 Equilength Segments

The equilength segments are of length 60 seconds giving a total of 4681 segments. The results from the treebagging on the equilength segments before and after the variable selection for the rating is shown in Figure (4.2). The same results for the standardised rating can be seen in Figure (4.3). The initial number of variables are 49. The variables obtained from variable selection with variable importance for the respective cases is shown in table (4.1).

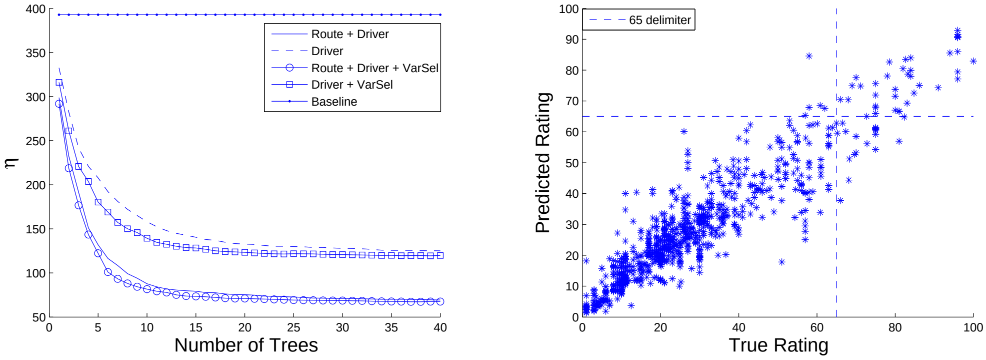


(a) (b)

Figure 4.1: (a) The training error for the classifier model on the original segments on the rating function. Roughly 60% is explained by the model. Including the route does not affect the results. (b) The training error for the classifier on the original segments with standardised rating. Without the route 30% of the variance is explained. With the route 32% of the variance is explained.

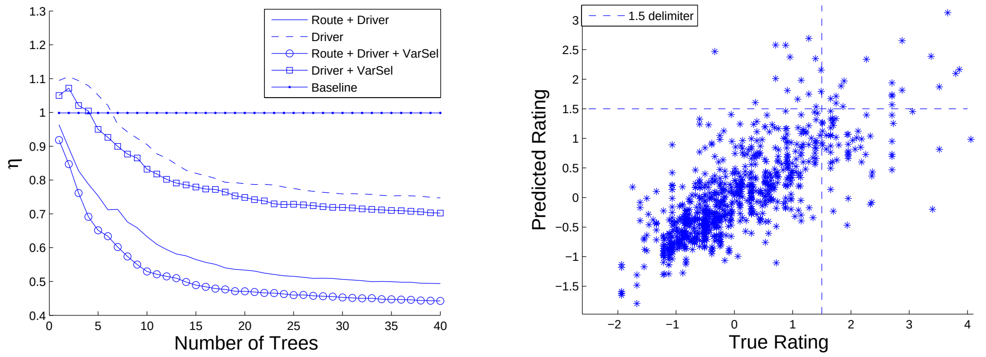
Case:	Rating		Standardised rating	
	Var name	Var. Importance	Var name	Var. Importance
	Driver	23.4	Driver	7.4
	Gender	0.6	Gender	0.6
	Route	9.5	Route	8.3
	min Speed	7.1	min Speed	7.8
	var YawRate	1.9	var SteerAng	1.8
	min AbsSteerAng	2.1	min LongAcc	1.1
	var AbsSteerAng	1.2	mean Brake	0.8
	N/A	N/A	min BrakeRate	0.7

Table 4.1: The variables selected for the model for the rating and standardised rating including the route. The variable importance is a relative number for the importance of the variable for the classification error. Values above 0 represent positive influence.



(a) (b)

Figure 4.2: (a) The training error for the ensemble classifier model on equilength segments on the rating function. More than 65% of the variance can be explained without the route. With the route inclued 82% of the variance is explained. (b) The predicted values plotted against the true values with a horizontal and vertical line at value 65.



(a) (b)

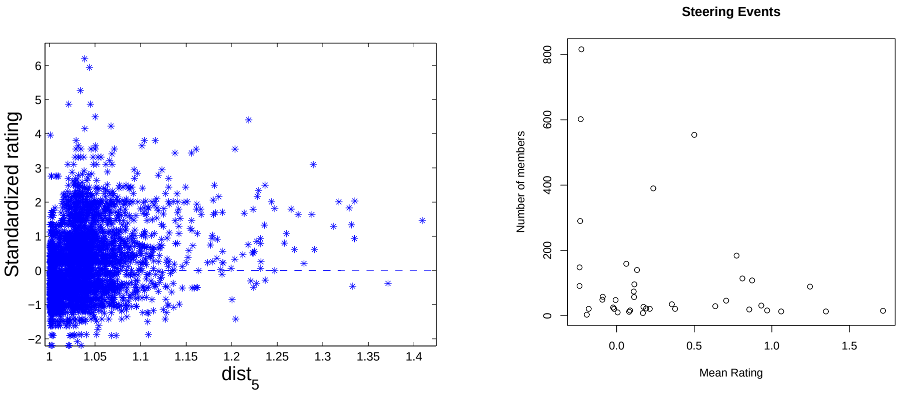
Figure 4.3: (a) The training error for the classifier model on equilength segments on the standardised rating function. Roughly 30% of the variance can be explained without the route. With the route inclued 55% of the variance is explained. (b) The predicted values plotted against the true values with a horizontal and vertical line at value 1.5.

4.2 Approach 2 - Covariate structure

The results from the outlier detection for the steering events and equilength subsegments is presented here. Some examples of matched events and subsegments can be seen in appendix B.2.

4.2.1 Steering Events

The variables used for the steering events are given in section (3.5.1). There is a total of 5112 steering events in the data. A scatterplot of $dist_5$ versus rating for all steering events are shown in Figure (4.4a). In (4.4b) a scatterplot of the mean standardised rating plotted against the number of members for a set of 60 clusters obtained by pruning an agglomerative complete linkage tree top down is shown. Four clusters have an average rating of greater than 1 and one cluster have an average rating of greater than 1.5.



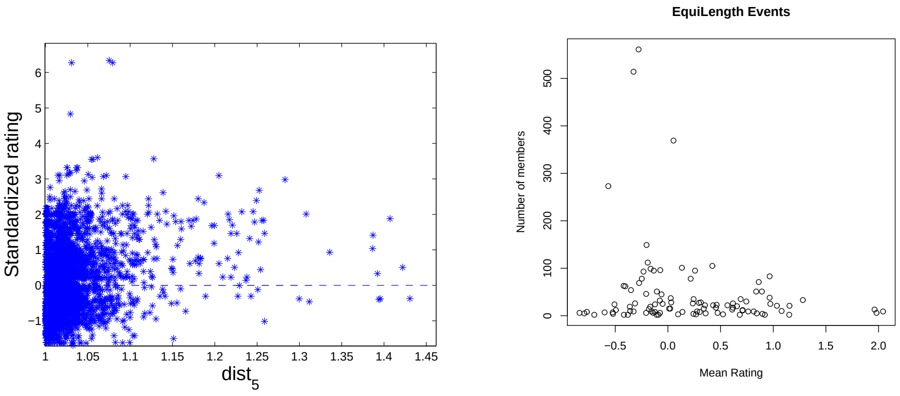
(a)

(b)

Figure 4.4: (a) Shows a scatterplot of the rating versus $dist_5$ for the steering events. (b) Shows the number of members plotted against the mean rating for 60 clusters obtained by pruning an agglomerative complete linkage tree top down.

4.2.2 EquiLength

The total number of equilength segments in the data is 4682. A scatterplot of $dist_5$ versus rating for all equilength segments are shown in Figure (4.5a). In (4.5b) a scatterplot of the mean standardised rating plotted against the number of members for a set of 60 clusters obtained by pruning an agglomerative complete linkage tree top down is shown. Eight clusters have an average rating of greater than 1 and two clusters have an average rating of greater than 1.5.



(a) (b)

Figure 4.5: (a) Shows a scatterplot of the rating versus $dist_5$ for the EquiLength segments. The Red line is the best linear fit to show the trend. (b) Shows the number of members plotted against the mean rating for 60 clusters obtained by pruning an agglomerative complete linkage tree top down.

4.3 Event matching EuroFOT data

Some high rating equilength segments matched against the EuroFOT database to investigate if simular situations can be found. The unease and similarity of the matched situations is estimated by the author. For this four segments are chosen from the equilength segments that have significantly higher standardised rating than average and a large distance to the nearest 5 events. A brief description of the situations are:

Event 1 - Driver is overtaking a lorry on a highway and is experiencing unease due to limited vision.

Event 2 - Driver is driving in a central area and passes a crowded pedestrian crossing.

Event 3 - Driver taking a left turn at a traffic crossing with a tramline through the center.

Event 4 - Driver is driving on a rural road and at a sharp left turn a meeting with an oncoming car driving in the center of the road takes place catching the driver by surprise.

For **event 1** over 300 matches are found. These are all similar to the original event and takes place on the highway with a file change or a slight turn. However a large majority of the events seem casual and cannot be deemed uneasy.

For **event 2** the number of matches are 6, one of these is basicly identical to the original situation and takes place at the same location. Of the six situations four are deemed to be uneasy.

For **event 3** the number of matches are 4 of which all are deemed uneasy.

For **event 4** no matches are found.

5

Conclusion and Discussion

In this chapter the conclusions for the respective approaches are given. This is followed by a discussion on the data, the methods and possibilities for future work.

5.1 Conclusion

5.1.1 Approach 1 - Ensemble classifier

From the results of the ensemble classifier it can be seen that a large proportion of the variance in the individual unease rating function can be explained using a few variables. The variable selection does not affect the proportion of explained variance much but it reduces the number of variables by a large factor.

There is a large difference between individuals as can be seen by the variable importance of the individual driver factor in the model. On the case with the original rating function this is partly explained by the discrepancy in value ranges between individuals. However the effect persists with standardised rating, indicating that there is a significant difference in experienced unease between drivers in different situations. Similarily when a variable for the route is included in the model the results are improved significantly. Speed has by far the largest effect of the driving parameters, followed by variables describing the steering behaviour.

As can be seen in Figure (4.2b) high values can accurately be predicted when including the route in the model. For the standardised rating seen in (4.3) high values are harder to predict. Without the route factor the model performs worse. Therefore the model is not suitable for filtering situations in the EuroFOT database as no routing exists there.

5.1.2 Approach 2 - Covariate structure

Using covariate structure for measuring similarity of situations works well. Some examples of matched events and subsegments can be seen in appendix B.2.

Using an agglomerative clustering approach is possible to detect a few clusters with much higher than average standardised rating both for the equilength events and steering events.

The matching of situations to the euroFOT data indicate that it is possible to detect situations similar to those in the VHM data that in most cases are uneasy and in all cases are very similar to the original situations in signal values.

The main conclusion here is that clusters with high rating and high representability can be achieved. This opens the door for further work with this method.

5.2 Discussion

This section offers a general discussion on the work process and the results.

5.2.1 The data

There is an issue with the rating function being so subjective as the discrepancy between individual value ranges is very large. One approach worth considering is having categorical rating levels with an explanation for each level to reduce subjectivity. This would open up for other methods of analysis between the categories.

Further for situation matching there is an issue with the representability between the EuroFOT and VHM data sets. The routing in the VHM data poses a limitation as it exposes the drivers to driving under conditions they are not used to. The routing also introduces other forms of bias. At times the drivers are unsure about the route which causes stress that leads to a large increase in rating value. These situations are not easily separated from the purely traffical high-rating situations. With that said it is not certain that such situations need to be separated.

The VHM data is also relatively sparse. Considering all situations that can occur while driving, and levels of severity of these, and the large set of environmental factors 72 hours of data over 7 individuals is not much. Many situations occur only once or twice in the data, giving no way to compare such situations between individuals and between levels of severity.

5.2.2 Chosing Models

As can be seen from section (3.6) a lot of different approaches and models was considered. It is not trivial to chose a suitable model for a problem of this type. It would have been desirable to limit the study to some situation, however as has been noted the data is sparse in comparison to the wide range of situations occuring. This makes it difficult to limit the study in such a manner that other factors can be kept constant. An alternative would have been to focus on a single event in the VHM data and focus on filtering out similar events in the EuroFOT database, using annotation based techniques.

5.2.3 Approach 1 - Ensemble classifier

The adaptable structure of ensemble classifiers makes them highly suitable for multivariate problems with high complexity. The downside is that the complexity of the classifier makes it difficult to extract any directly applicable knowledge of the partitioning of the predictor space.

The proportion of variance explained by the model is large. The single most important variable in the model is the driver factor. This is explained partly by the different scales of the rating function. However when using the standardised rating the driver factor still has the largest effect indicating a difference between individuals apart from the usage of the rating scale.

The results are signficantly improved when including the route factor in the model. Therefore it can be said that similar situations as measured by the variables in the model are not (perceived) equal over different days by the drivers. Without further studies it is impossible to say what the cause of this effect is.

From the variables the minimum vehicle speed has the largest effect. This indicates that the current speed is the largest factor effecting the rating. Intuitionally this is logical since the speed limits are heavily correlated to the complexity and risk level at the road. The second largest effects are regarding the steering activation. This could be due to the fact that the low speed driving is in central areas where complexity of task is higher.

5.2.4 Approach 2 - Covariate structure

The covariate structure proved good for matching situations. However it also holds some limitations. Mainly that situations must be very similar to be matched. However with a larger amount of data more and better subclusters could be detected to which the matching could be done. Matching situations to a cluster means a less strict matching and possibly classification of previously unseen situations by assignment to clusters instead of matching to individual situations.

5.2.5 Validation

A stricter validation of the results could have been done by showing the test subjects video footage from the extracted situations and asking them to rate it. This rating value could be compared to the predicted rating. However the subjects where unavaliable. Using other subjects than those included in the study is not desirable due to the difference in the usage of the rating scale. Then a baseline would have to be established for each subject to be used for evaluation.

As the results from the classifier show the route has a large effect on the results. The route data is not avaliable in the EuroFOT database and the settings are significantly different from naturalistic studies. From the comments of the drivers it is clear that the ability to navigate the route is a large factor in the experienced unease while driving coupled with driving in central areas of Gothenburg that they are not used to. Further the driving in central areas might be very similar to driving in familiar low-speed areas such as the subjects local neighbourhood. Such situations are not accounted for in the VHM data but numerous in the EuroFOT data.

5.3 Future work

If any future work is to be done the following is proposed:

The issues regarding the difference in rating scale values between individuals should be adressed. A more standardised scale could be introduced or a study regarding the usage of the rating scale could be performed. A more standardised scale could be categorical with fewer values and a stricter explanation for each value.

Controlling for setting and increasing representability of the study data with naturalistic driving data could be done. A novel way to do this is to extract a number of situations of interest from the naturalistic data. The video from this data could be shown to a number of subjects and rated. This would mean a loss of the first hand experience from the driver but would also greatly reduce the cost of data gathering and increase control of situations.

Using the covariate model and a method for averaging of the subspaces of principal components could be constructed. Using such a method a set of reference clusters could be extracted. This would open up for the possibility of mapping up driving into a set of categories. Further clustering could then be performed within these clusters, or a study of the time spent in these clusters by different drivers could be used to understand the driving habits of individuals. That is a study of intra- and inter-individual effects.

Bibliography

[1] J. Dargay, D. Gately et al., *Vehicle Ownership and Income Growth, Worldwide: 1960-2030.* Institute for Transport Studies, University of Leeds, 2007

[2] M. Peden, R. Scurfield et al., *World report on road traffic injury prevention.* World health organization, 2004; pp. 34

[3] Parry et al, *Automobile Externalities and Policies.* Winston Harrington, 2007; pp. 9

[4] Gregory Piatetsky-Shapiro, *Knowledge discovery in databases: 10 years after.* Knowledge Stream Patners, 1999

[5] U. Fayyad, G. Piatetsky-Shapiro et al., *Knowledge discovery in databases: An overview.* AI Magazine, 1992; 13(3)

[6] C. Mijayima, K. Takeda et al., *A study of driver behaviour under potential threat.* University of British Columbia, 2009; 10(2):201-210

[7] N. Saunier, T. Sayed, *Clustering Vehicle Trajectories with Hidden Markov Models Application to Automated Traffic Safety Analysis.* Transactions on intelligent transportation systems, 2010

[8] Z. Ghahramani, *Introduction to Hidden Markov Models and Bayesian Networks.* University College London, 2001

[9] C. Sutton *Classification and Regression Trees, Bagging and Boosting.* Handbook of Statistics Vol. 24, 2005.

[10] D. Lay *Linear Algebra and its applications 3rd Edition* Pearson Education Inc., 2006.

[11] N. Kumar, et. al, *A New Singular Value Decomposition Based Robust Graphical Clustering Technique and Its Application in Climatic Data* Journal of Geography and Geology Vol. 3, 2011.

[12] S. Spiegel, J. Gaebler, et. al, *Pattern Recognition and Classification for Multivariate Time Series* Technische Universitaet Berlin, 2011.

A

Implementation appendix

This chapter provides an overview of matlabs *TreeBagger*, the implementation used. As the function is versatile and have a wide variety of settings the used settings are given and explained. The input data is given as a matrix X with each row corresponding to one observation and each column to one variable. The response vector Y is of the same length as the columns of X and contains the mean rating for that segment. The specific call used is:

```
B = TreeBag-
ger(nTrees,X,Y,'Method','regression','NVarToSample','all','OOBVarImp','on');
```

Which returns a trained tree structure B. The first parameter *nTrees* is the number of trees to train. This value is set with respect to the convergence of the training error. The *Method* is either *regression* for numeric classes or *classification* for categorical classes. The *NVarToSample* option is the number of variables to select randomly at each split and *all* is the default setting. *OOBVarImp* is either *on* or *off* and indicates wether to form the out of bag training error at every tree trained.

B

Covariance matching principal components and matchings

In this appendix the principal components for the clusters with highest rating are given. Some matched signals using the covariance based method are also shown.

B.1 Principal components

B.1.1 Equilength Segments

The principal components for one observation from each of three clusters with average standardised rating above 2 are given in Tables (B.1), (B.2) and (B.3).

B.2 Matching examples

Some examples of matched sequences from the matching algorithm are given here.

B.2.1 Steering event

The two closest matches obtained using the covariance matching for four steering events are presented below.

B.2.2 Equilength event

The two closest matches obtained using the covariance matching for four equlength events are presented below.

Var	Comp 1	Comp 2	Comp3
VehicleSpeed	0.26	-0.30	0.29
BrakePressure	-0.28	0.15	0.07
BrakeRate	0.13	0.01	-0.76
SteeringAngle	0.13	-0.23	0.27
SteerRate	-0.41	-0.84	-0.26
LongAcc	0.73	-0.14	-0.26
LateralAcc	0.08	-0.20	0.20
YawRate	0.14	-0.18	0.26
AccPedalPos	0.29	-0.17	0.15
AccPedRate	-0.05	-0.10	0.03

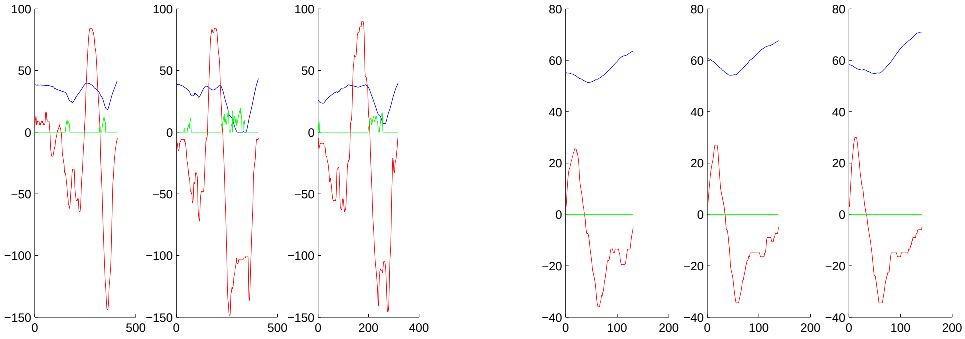
Table B.1: The principal components of one observation in a cluster with standardised rating larger than 2 for the equilength segments.

Var	Comp 1	Comp 2	Comp3
VehicleSpeed	0.17	-0.01	0.32
BrakePressure	-0.02	0.10	-0.25
BrakeRate	0.09	-0.08	-0.88
SteeringAngle	-0.90	0.08	-0.05
SteerRate	0.01	0.96	0
LongAcc	-0.11	-0.18	0.13
LateralAcc	-0.08	0.01	0.10
YawRate	-0.36	-0.17	0.03
AccPedalPos	0.05	-0.05	0.19
AccPedRate	-0.02	0.02	0

Table B.2: The principal components of one observation in a cluster with standardised rating larger than 2 for the equilength segments.

Var	Comp 1	Comp 2	Comp3
VehicleSpeed	0.64	-0.05	0.34
BrakePressure	-0.45	0.41	-0.22
BrakeRate	-0.02	0.03	0.14
SteeringAngle	0.49	0.64	-0.49
SteerRate	-0.02	-0.10	-0.14
LongAcc	0.10	-0.57	-0.74
LateralAcc	-0.07	0.06	-0.06
YawRate	0.13	0.21	-0.07
AccPedalPos	0.34	-0.18	0.01
AccPedRate	-0.01	-0.01	0.02

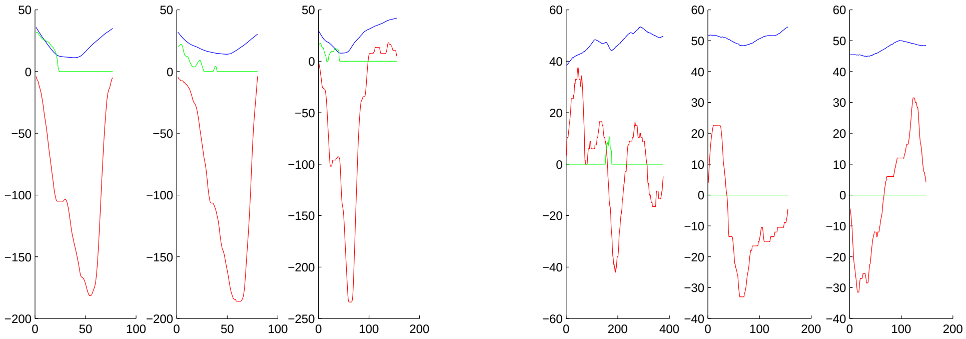
Table B.3: The principal components of one observation in a cluster with standardised rating larger than 2 for the equilength segments.



(a)

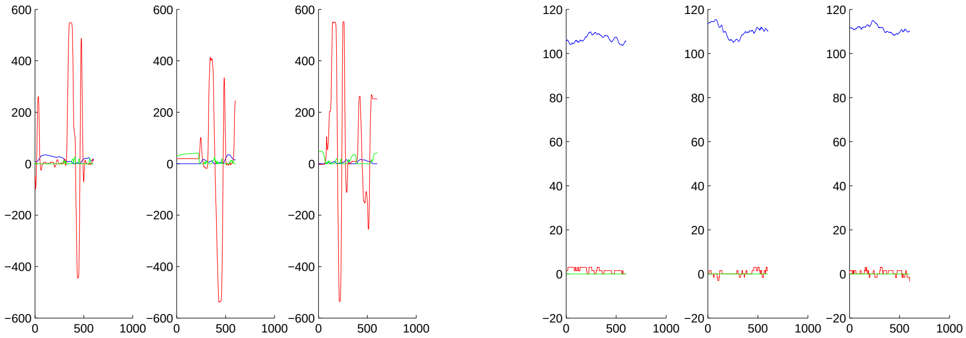
(b)

Figure B.1: The two closest matches for two steering events obtained using the covariance matching algorithm. The leftmost event is the original. The red signal is the steering angle, the blue is the vehicle speed and the green is the brakepressure.



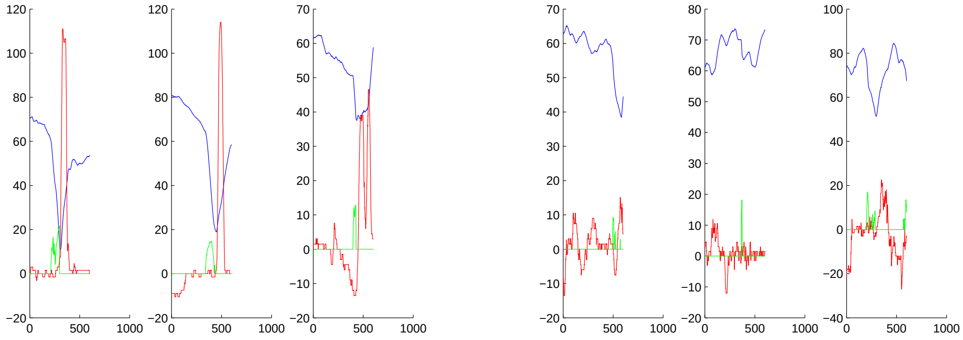
(a) (b)

Figure B.2: The two closest matches for two steering events obtained using the covariance matching algorithm. The left event is the original. The red signal is the steering angle, the blue is the vehicle speed and the green is the brakepressure.



(a) (b)

Figure B.3: The two closest matches for two equilength segments obtained using the covariance matching algorithm. The left event is the original. The red signal is the steering angle, the blue is the vehicle speed and the green is the brakepressure.



(a)

(b)

Figure B.4: The two closest matches for two equilength segments obtained using the covariance matching algorithm. The left event is the original. The red signal is the steering angle, the blue is the vehicle speed and the green is the brakepressure.