A supervised learning approach to estimate the drivers impact on fuel consumption
A heavy-duty vehicle case study

GEORG ZETTERBERG WALLIN
MATTHIEU CRÉTIER
A supervised learning approach to estimate the drivers impact on fuel consumption
A heavy-duty vehicle case study

Georg Zetterberg Wallin - georgzw@kth.se
Matthieu Crétier - cretier@kth.se
Abstract

The aim of this Master thesis is to provide a statistical analysis of the factors influencing the fuel consumption, with a focus on the separation of the drivers’ performance. The study is focused on the long haulage trucks, which correspond to the application where the fuel consumption becomes of primary interest from the economical point of view. Further developments of the work leads to a graphical representation of the outcomes on a map, highlighting in particular the segments of the road network having the highest variation of the driver-influenced fuel consumption.

The analysis dataset created is the combination of data coming from different sources and additional features computed based on them. The datasources are providing respectively the vehicles’ operating data and configurations, the road network’s characteristics and the weather information.

The results obtained prove that it is possible to isolate the driver factor from the overall fuel consumption. This can be achieved by training a model composed by variables statistically chosen through a regression procedure. Further in the analysis the different driver factors are used in order to determine the fuel saving potential of the road stretches where the factors are computed. The results are gathered in multiple stages, based on the dimension of the dataset considered and the method used. Two methods are used to train the model: the least squares regression and the ridge regression. First the whole Swedish road network composed by primary roads is analyzed with least squares. 1195 road stretches belonging to this network present a defined and different than zero fuel saving potential varying between 0.003 and 83.71 l/100km. Then, a smaller portion of the same road network is analyzed after being provided with road slope information. The fuel saving potential estimated using ridge regression present values between 0.002 and 24.39 l/100km.

From the geographical point of view little can be deduced from the analysis of the complete network. The E4 provided with slope data, on the contrary, allows a better insight, especially using ridge.
Sammanfattning


Den datamängd som används i analysen är en kombination av data från olika källor. Dessa datakällor ger respektive fordonss driftdata, konfigurationer, vägnätverkets egenskaper samt väderinformation.

Resultaten visar att det är möjligt att isolera förarens påverkan från den totala bränsleförbrukningen. Det ästadkoms genom att en regressionsmodell anpassas till den data som inhämtats. I studien används en mängd bränsleförbrukningsfaktorer från olika förare för att bestämma bränslebesparingspotentialen till de vägsegment över vilka förarna färdats. Resultaten är presenterade i flera steg, baserat på mängden data som används och på metoden som utnyttjades. Två metoder används för att träna modellen: minstakvadratmetoden och ridge regressionen. Först analyseras det svenska vägnätverk som primärt består av motorvägar med minsta kvadrat regression. Av dessa vägsegment visar 1195 st. en bränslebesparingspotential större än noll. För dessa vägsegment varierar bränslebesparingspotentialen mellan 0,003 och 83,71 l/100 km. Sedan används en mindre del av samma vägnätverk och det analyseras efter att ha blivit försedd med information om väglutning. Bränsle besparingspotentialen uppskattas med hjälp av ridge regression och resultatet varierar mellan 0,002 och 24,39 l/100 km för de olika vägsegmenten.

Ur ett geografiskt perspektiv ger analysen av hela vägnätverket inga nya insikter som kan användas. Analysen av E4 försedd med väglutningsinformation ger däremot en bättre tolkningsbart resultat, särskilt vid användning av ridge regression.
## Contents

1 Introduction ................................. 1  
  1.1 Sustainability in transport ................... 1  
    1.1.1 Planetary boundaries ..................... 1  
    1.1.2 The role of transport in climate change ........ 2  
  1.2 Economy of a transport company .............. 3  
    1.2.1 An overview of the market situation .......... 3  
    1.2.2 The analysis of the logistics companies’ expenses .... 4  
    1.2.3 Real case scenario ....................... 5  
  1.3 Driver effects ............................. 6  
  1.4 Fuel consumption ............................ 7  
    1.4.1 Energy efficiency of vehicle propellants ........ 7  
    1.4.2 Fuel consumption conceptualization .......... 7  
    1.4.3 Energy dissipation in a road vehicle .......... 8  
    1.4.4 Driving performance effect on the motion resistance .... 9  

2 Method ...................................... 10  
  2.1 Data acquisition from database .............. 11  
    2.1.1 Acquisition of vehicle data and configurations .... 11  
    2.1.2 Acquisition of road data .................. 11  
    2.1.3 Acquisition of weather data ................ 12  
  2.2 Missing features computation ................. 13  
    2.2.1 Fuel consumption .......................... 13  
    2.2.2 Vehicle weight ............................ 14  
    2.2.3 Wind speed components .................... 14  
  2.3 Aggregating spatial data .................... 15  
  2.4 Statistical methods and phenomena .......... 16  
    2.4.1 Linear regression .......................... 16  
    2.4.2 Model selection ............................ 17  
    2.4.3 Multicollinearity .......................... 18  
    2.4.4 Ridge regression ........................... 19  
    2.4.5 Estimating confidence intervals using bootstrapping .... 20  
  2.5 Fuel consumption model ..................... 21  
    2.5.1 Predictors ................................ 21  
    2.5.2 Regression model ........................... 21  
    2.5.3 Stage one application of fuel consumption model .... 22  
    2.5.4 Stage two application of fuel consumption model .... 25
### Results

<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>3.1.1 Acquisition of vehicle data and configurations</td>
<td>27</td>
</tr>
<tr>
<td>3.1.2 Acquisition of road data</td>
<td>28</td>
</tr>
<tr>
<td>3.1.3 Acquisition of weather data</td>
<td>29</td>
</tr>
<tr>
<td>3.2.1 Predictors</td>
<td>30</td>
</tr>
<tr>
<td>3.2.2 Variables influence in the fuel consumption model</td>
<td>31</td>
</tr>
<tr>
<td>3.2.3 Stage one</td>
<td>32</td>
</tr>
<tr>
<td>3.2.4 Stage two</td>
<td>33</td>
</tr>
<tr>
<td>3.2.5 Stage three</td>
<td>35</td>
</tr>
<tr>
<td>3.2.6 Fuel saving potential’s geographical distribution</td>
<td>38</td>
</tr>
</tbody>
</table>

### Discussion

<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>4.1 Predictors</td>
<td>40</td>
</tr>
<tr>
<td>4.2 Comparison of approaches</td>
<td>41</td>
</tr>
<tr>
<td>4.3 Fuel saving potential’s geographical distribution</td>
<td>42</td>
</tr>
</tbody>
</table>

### Conclusions

43

### Appendices

A Fuel saving potential visualizations I
Nomenclature

\( \alpha_C \) Confidence level
\( \bar{f}_c \) Average fuel consumption
\( \bar{x} \) Mean
\( \bar{y} \) Mean value of the dependent variable
\( \beta \) Regression coefficient
\( \beta_0 \) Intercept regression coefficient
\( \ddot{x} \) Vehicle acceleration
\( \Delta \) Deviation from the reference point
\( \dot{x} \) Vehicle speed
\( \epsilon \) Error term
\( \hat{\beta} \) Estimation of the regression coefficient
\( \lambda \) Shrinkage coefficient
\( \sigma \) Standard deviation
\( \sigma^2 \) Variance
\( A \) Vehicle frontal reference area
\( ang(H) \) Heading angular direction
\( ang(W) \) Wind angular direction
\( C_x \) Aerodynamic drag
\( cvw_s \) Calculated vehicle weight of the shift
\( d \) Odometer
\( d_s \) Odometer of the shift
\( f \) Total fuel
\( f_r \) Rolling resistance coefficient
\( F_x \) Longitudinal force at the wheels
$f_{sp}$  Fuel saving potential
$g$  Gravity acceleration
$H$  Heading
$I$  Confidence interval
$I^+$  Confidence interval upper boundary
$I^-$  Confidence interval lower boundary
$m$  Vehicle mass
$m_j$  Vehicle rotating masses
$n$  Number of regression observations
$p$  Number of predictors
$p_{int}$  Interaction independent variable
$p_{lin}$  Linear independent variable
$p_{nlin}$  Non-linear independent variable
$R^2$  Coefficient of determination
$RSS$  Residual Sum of Squares
$t_{\alpha/2}(df)$  Quantile function of the t-distribution
$TSS$  Total Sum of Squares
$W$  Wind speed
$w_{ls}$  Wind longitudinal speed
$w_{ts}$  Wind transversal speed
$w_s$  Vehicle weight of the shift
$X$  Independent variable matrix
$x$  Independent variable
$Y$  Dependent variable vector
$y$  Dependent variable
$\alpha$  Road slope angle
$\rho_{air}$  Air density
1. Introduction

The driver behavior and its influence over the fuel consumption in heavy-duty vehicles has been the topic of several studies. Various attempts to improve the fuel economy through a corrective action of the driving behavior were carried out; it has been demonstrated that this can be achieved by following predetermined velocity profiles [1] or by assuming an anticipation behaviour [2]. These studies were limited to a small sample of drivers and vehicles, while this study aims at investigating this topic from a broader perspective. The objective of this thesis can be summarized by the following three goals:

- Separate the fuel consumption caused by the driver from other factors.
- Calculate the fuel saving potential and visualize it spatially.
- Identify geographic areas where the fuel saving potential is large.

The following sections of the introduction will present an overview of the reasons behind the strive towards an improvement in the fuel consumption. Moreover the question will be presented from different perspectives: from the environmental, economical, behavioural and energy dissipation point of view.

1.1 Sustainability in transport

1.1.1 Planetary boundaries

In order to understand and quantify the effect of human society on the earth system, the planetary boundaries framework can be used. Planetary boundaries is a concept that defines the safe operating space for humanity. The planetary boundaries defines geo-physical boundaries for humanity to stay within to ensure a continuing stable operation of the earth system. Trespassing these boundaries could cause sudden irreversible environmental changes that could seriously deteriorate and harm the human well-being. To quantify these earth system processes nine critical planetary boundaries are identified and presented in Figure 1.1. The nine planetary boundaries quantify different aspects of the safe operating space, but the interconnection and integration between the boundaries cannot be neglected. Trespassing the threshold for one of the nine boundaries can result in that other boundaries are brought closer to its critical threshold. To enable quantification of the planetary boundaries each earth system process is connected with

1
1.1. SUSTAINABILITY IN TRANSPORT

one or more control variables. These quantities are in the best case measurable or could otherwise be estimated in order to operationalize the safe operating space [3].

![Planetary Boundaries - A safe operating space for humanity](credit: Azote Images/Stockholm Resilience Centre [4].)

1.1.2 The role of transport in climate change

On September 25, 2015, the general assembly of the United Nations adopted the resolution *Transforming our world: the 2030 Agenda for Sustainable Development*. The agenda establish 17 global goals for sustainable development and 169 targets for all countries and stakeholders to work towards. For example, goal 12.2 declares the importance of efficient use of natural resources and in paragraph 27 sustainable transport is described as one of the important factors for building strong economic foundations for all member countries [5]. When emissions are concerned, 23% of the total energy related CO₂ emissions comes from the transport sector. The emissions from transport are predicted to double by 2050 [6]. If a European perspective is used, road transport make up 71.9% of the total greenhouse emissions caused by the transportation sector which is presented in Figure 1.2 [7].
The contribution of CO$_2$ from road vehicles is not solely from the actual vehicle operation. To grasp the full extent of CO$_2$ contribution the complete life cycle, including manufacturing and end of life after treatment of the vehicle must be considered [8]. In this thesis the full perspective will not be investigated. Instead focus is on the vehicle operation phase. One way to decrease the emitted CO$_2$ from road vehicles during operation, is to implement measures to decrease fuel consumption. Fuel consumption and the emission of CO$_2$ have a linear relationship, which is deduced from the hydrocarbonate content of gasoline and diesel fuels [9]. This means that lowering the fuel consumption will contribute to a decrease in CO$_2$ emissions.

### 1.2 Economy of a transport company

#### 1.2.1 An overview of the market situation

In the present global economy the transport sector has become a core business for its development. Economic opportunities are tightly related to the mobility of goods, this is such a close liaison that the efficiency of the transport system is often determining the profitability of an investment [10].

From an industrial point of view the focus on supply chain management and the adoption of the kanban [11] approach occurred in the last decades has pushed the industries to seek better services and to decrease their operational costs. The strive towards the reduction of inventories size, their centralization and the abandonment of stock-keeping
points in the supply chain increased the demand for the logistics companies [12]. The current scenario shows a considerable increase in the competition level among the transport companies: quick deliveries and low bills are common requirements that a logistics agent has to provide to keep its position in the market. From a goods' delivery point of view a transport company aims at being flexible and able to face emergencies and at taking advantage of its assets that are required to have a high uptime and to exploit fully their load capacity. While carrying out its tasks it needs to be profitable and, in order to achieve this status, particular attention is required in dealing with its operating costs.

The difference between revenues gained after providing a logistics service and costs faced during its performance corresponds to the operating profit. It is in the company interest to have this amount of money as high as possible. As a matter of fact it is used to bear the tax expenses, to invest in the company asset and, if the business is public, to pay its shareholders. Having that in mind a transport company has mainly two options: to increase the revenues either by guaranteeing a better service for a higher price or by leveraging on its image strength, and/or to decrease the operating costs. In the current market situation the first alternative has become unlikely: there are too many entities providing the same service for a very competitive price.

1.2.2 The analysis of the logistics companies’ expenses

Before planning any cut or saving strategy, the costs need to be detected and divided according to common sources. Several transport companies have a subdivision of the operating costs according to Figure 1.3.

![Figure 1.3: Operational costs of a transport company][13]

The subdivision mentioned can be considered standardized, yet some small differences can occur between companies of different size and operating in different areas. Fuel cost, employees wages, third part services and environment conditions are the main...
factors influencing the expenses of a logistic company. As is seen in Figure 1.3 the fleet itself is contributing to less than 15% of the total amount. The vehicles’ price has such a low leverage since it is the main contributor to the value creation for the logistics operation, moreover the initial investment to buy whichever truck is taken into account over the whole lifecycle of the vehicle. On the annual balance sheet of the transport company it is reported under the heading 'Asset depreciation'.

In the process of decreasing its operating costs a company should address its effort towards the parameters affecting the most the expenditures: fuel and drivers. The latter heading is not a viable alternative since in order to achieve serious improvements one would commit labor exploitation. Many countries pushed by the trade unions and the EU have carefully legislated against this threat to labor rights and in 2006 an European measure to regulate truckers working hours was approved [14]. Although wages and working conditions can not be addressed, the costs related to the manpower could be severely decreased in the future after the dawn of the autonomous vehicles era. For the previous reasons nowadays an intervention on the fuel consumption is the most profitable solution.

To decrease the costs related to the fuel purchasing is possible, but it exists the chance to encounter major issues. In fact the possibility to get a better fuel economy is strictly depending on several factors, among them are the traffic situation and the morphology of the road: neither in a congested road nor in steep uphill there is margin for important improvements. But in any other case it is very likely to cut down the expenses of the fuel supply.

### 1.2.3 Real case scenario

Table 1.1 reports the expenses related to the fuel supply of two important companies operating in the sector. The costs have been obtained from the annual report that they presented for the year 2014 [15] [16].

<table>
<thead>
<tr>
<th>Company</th>
<th>Year</th>
<th>Fuel Costs</th>
<th>Headquarter</th>
</tr>
</thead>
<tbody>
<tr>
<td>DHL</td>
<td>2014</td>
<td>EUR 848M</td>
<td>Germany</td>
</tr>
<tr>
<td>YRC</td>
<td>2014</td>
<td>USD 458M</td>
<td>USA</td>
</tr>
</tbody>
</table>

As it is possible to see from Table 1.1 the two agents had high fuel supply costs, whose magnitude is not surprising considered their size and the scope of their operations. There is a certain discrepancy between the two invoices which is simply caused by their size and to how widespread their services are.

Both DHL and YRC could benefit from a decrease in the fuel consumption. The previous affirmation can be clarified with a simple example based on the DHL case; inflation and price per barrel are assumed constant over time. In 2014 DHL paid 848
million euros to refuel its fleet. The average fuel consumption of all its trucks could realistically be of 28 l/100 km. In the Scania Fuel Efficiency Duel occurred in May 2011 its winner proved that it is possible to reach an economy of 25.7 l/100 km [17]. If the German courier through some internal politics and driving courses decreased average consumption to 27.5 l/100 km (98% of the hypothetical value in 2014) it would be able to save roughly 16.96 million euros.

1.3 Driver effects

The fuel consumption of heavy vehicles is affected by many factors, these factors can be arranged into five main components of the traffic system. These are Road (R), Vehicle (V), Driver (D), Environment (E) and Policy (P). Also the interactions between these components are considered as factors affecting the fuel consumption. This means that the factors that affect the vehicle fuel consumption are: R, V, D, RV, RD, VD, RVD, E and P [18]. Some of the characteristics of each component are presented in Table 1.2.

<table>
<thead>
<tr>
<th>Road (R)</th>
<th>Vehicle (V)</th>
<th>Driver (D)</th>
<th>Environment (E)</th>
<th>Policy (P)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Characteristics: Geometry, roughness, etc.</td>
<td>Dimensions, engine, weight, etc.</td>
<td>Behaviour, skills, etc.</td>
<td>Temperature, wind, altitude, etc.</td>
<td>Design related policies that affect fuel consumption.</td>
</tr>
</tbody>
</table>

If focus is put on the driver's possibility to affect the fuel consumption, the question of which driving behaviour is profitable arises. According to previous studies there are a few main metrics that describe the driver behaviour that affects the fuel consumption. If highway driving is concerned the most important factor is the vehicle velocity. Also the intensity and frequency of acceleration and braking are factors that have influence over the consumed fuel during driving. During zero velocity situations idling is a contributing factor as well as the frequency of stops done. A precise quantification of how large effect the driver, and other factors have on fuel consumption is hard to find in the literature [19]. Few studies regarding this area are taking environmental and road factors into consideration when comparing the effect on fuel consumptions for example driver training, which could introduce uncertainties in the results. Most studies also only have access to a small fleet of vehicles and drivers and are not committed in real world conditions which could limit the external validity of the results from those kinds of studies.
CHAPTER 1. INTRODUCTION

1.4 Fuel consumption

1.4.1 Energy efficiency of vehicle propellants

Fuel consumption is a term belonging to automotive applications, it derives from the need of describing the energetic performance of a vehicle during its duty time. This concept comes from the broader definition of fuel efficiency, which is applicable to several chemical processes that are meant to produce work. The most common way to perform this operation is with an engine, where the propellant’s chemical energy is transformed first into kinetic energy, then into work through the combustion process.

Worldwide it is possible to find several types of fuel, most of them need to be refined in order to be useful. Once this procedure is concluded though, several different propellants are created, all having a high specific energy density. The main automotive fuels are reported in Table 1.3 with their specific energy content.

<table>
<thead>
<tr>
<th>Fuel type</th>
<th>Low specific energy content</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gasoline</td>
<td>47 MJ/kg</td>
</tr>
<tr>
<td>LPG</td>
<td>51 MJ/kg</td>
</tr>
<tr>
<td>E85</td>
<td>33 MJ/kg</td>
</tr>
<tr>
<td>Diesel</td>
<td>48 MJ/kg</td>
</tr>
</tbody>
</table>

The values reported in Table 1.3 represent the amount of heat energy provided by one kilogram of propellant during a combustion whose output is CO\textsubscript{2} and steam H\textsubscript{2}O. Typically any petroleum based fuel has a high specific energy content that makes them extremely profitable for several applications, moreover they are found in nature at their liquid state allowing them to be easily stored. These two characteristics are the reason why they have been the most common propellants in automotive applications since the 19th Century.

1.4.2 Fuel consumption conceptualization

The main purpose of a vehicle is to move people and/or goods. Since a displacement is involved, the required work corresponds to force multiplied by distance. This leads to the usual representation of fuel consumption where the force applied to move the vehicle is disregarded and the focus is on the displacement: l/100km. This ratio is well-known in the automotive industry and it is reported in most of the cars’ dashboard nowadays. It corresponds to the liters of propellant required by the vehicle to perform 100 km at the driving conditions of the moment when the measure is taken. Since the concern on vehicle energy efficiency has grown in the last two decades, the fuel consumption
became a fundamental characteristic that every manufacturer has to provide in order to help the customer in its choice [21]. Usually more values of fuel consumption are provided because there are multiple driving scenarios affecting the vehicle performance. These are calculated following some standardized tests known as driving cycles.

The greatest advantage of describing the vehicle performance through the fuel consumption presented as l/100km depends on the fact that it is an indicator easy to understand. It clearly expresses the distance and the amount of liters required to cover it; and the related cost can be deducted from it. However it has a limited area of application, in particular it becomes meaningless when the vehicle is not moving. If that is the case, assuming that no Start-stop system is installed, idling time becomes quite relevant especially under the circumstance of traffic congestion. The fuel consumption index is no longer depending on the distance, but on the time: it is usually expressed as l/h. The fuel consumption at idle is represented by a defined number depending on the engine performance at its running condition. From the point of view of small motor vehicles fuel consumption and fuel consumption at idle can describe their operating cost in most situations. However while dealing with buses and trucks an important role is played by the weight of the vehicle, especially in the latter case where the load can vary between 5 and 44 tonnes, corresponding to the maximum limit imposed by EU on international traffic [22]. This high variations depends on the amount of goods stored in the trailer/container. Another indicator for the fuel consumption accounting for the weight is preferable, so the choice falls on l/(100km t), that corresponds to the litres of propellant used to carry 1 ton of goods over a distance of 100 kilometres.

### 1.4.3 Energy dissipation in a road vehicle

The use of fuel consumption as indicator is essential because of the limited efficiency of internal combustion engines and the high amount of losses occuring all along the driveline. Such flow of energy is depicted in Figure 1.4.

![Figure 1.4: Example on energy flows in a vehicle running on a highway][23].

From the example in Figure 1.4 it is seen that only 25% of the energy produced during the combustion is transmitted, and partly dissipated, in the driveline. The remaining
portion is used to produce the work able to overcome the resistance to motion of the vehicle, whose three components (inertia resistance, rolling resistance and aerodynamic resistance) may vary depending on the driving condition. Their correlation is expressed by the motion resistance equation and it represents the force that the vehicle must overcome in order to be set into motion.

\[ F_x = (m + m_j) \cdot \ddot{x} + m \cdot g \cdot (f_r \cdot \cos \alpha + \sin \alpha) + 0.5 \cdot C_x \cdot A \cdot \rho_{\text{air}} \cdot \dot{x}^2 \]  

(1.1)

In Equation 1.1 inertia force, rolling resistance force and aerodynamic force contribution is easy to detect. Among them the last two are affecting the fuel consumption the most. The rolling resistance is mostly caused by hysteresis in the tire materials which occurs when the wheel is rolling. All other physical phenomena, e.g. slide between the tire and the road, resistance due to air circulating within the tire and the air turbulence produced by the rotating tire, are less important. 90-95% of the rolling resistance is caused by internal hysteresis, 2-10% by friction between the tire and the ground and 1-3% is caused by air resistance [24]. Likewise, the aerodynamic resistance has two sources: one is the airflow around the vehicle body, the other the flow through its radiator and interior. The former is the dominant factor and it generates normal and shear stress on the whole vehicle body, caused by a gradient in pressure between front and wake of the vehicle.

1.4.4 Driving performance effect on the motion resistance

In the case of a heavy duty vehicle both rolling and aerodynamic resistance present an increase in magnitude with respect to a passenger vehicle. This is caused by the larger number of axles and the lack in compactness of the carriage. The first factor gives rise to a larger number of friction zones (rolling resistance) as well as higher air turbulence at the wheels (air resistance). On the contrary, the second one affects only the drag since it creates recirculation zones in the gap between the cabin and the trailers [25]. These issues can be tackled at a design stage; as a matter of fact the aerodynamics of the vehicle can be improved through the insertion of panels and the rolling resistance can be reduced by using more energy efficient tyres and by distributing the load more wisely [25].

Further actions can be taken during the performance of the truck’s duties. In fact as it is reported in the Pacejka Formula also the rolling resistance is affected by the vehicle speed and in particular by the slip of the wheels [24]. This assertion is supported by several studies whose purpose is to find the underlying connection between driving style and fuel consumption in vehicles. The driving styles assembles several parameters, among which the most relevant are maximum acceleration, maximum engine speed, average throttle position and speed’s standard deviation [26].
2. Method

The goals stated in the introduction can be achieved in several ways, among the others through simulations and field tests. Nevertheless the approach chosen for this thesis is data analytics. The study is focused on long haulage vehicles’ measurement points collected exclusively on high speed roads. Most likely the vehicles considered would belong to transport companies, whose operating costs are strongly affected by the total fuel consumption [13]. Therefore there is a huge interest in reducing the fuel consumption through the improvement of the driver performance. In particular this is possible in roads where the traffic congestion is limited, with a consequent reduction of harsh driving and idling situations. This together with the lack of traffic lights make the highways the best choice for this analysis.

The analysis procedure is divided in six steps as Figure 2.1 shows. It is possible to visualize the complete flow of operations, whose steps are performing several important operations leading to the final visualization of the outcome.

![Data processing procedure](image)

**Figure 2.1: Data processing procedure**

The process starts with the data acquisition from the data sources, then it continues with the calculation of the missing features, e.g. fuel consumption, and the aggregation of the data coming from the different sources. The core of the analysis consists in the training of the fuel consumption model, and it is divided in two parts: model selection and factor separation. Finally the outcome of the study is elaborated and visualized.


2.1 Data acquisition from database

2.1.1 Acquisition of vehicle data and configurations

The vehicle data is provided by a fleet management data warehouse and is considered the core of the working dataset. It corresponds to the operating information of the trucks provided with an internet connection.

The vehicle data available has different forms, depending on their content, namely snapshot and aggregated data. Table 2.1 is reporting some example of the data used.

<table>
<thead>
<tr>
<th>Snapshot data</th>
<th>Accumulated data</th>
</tr>
</thead>
<tbody>
<tr>
<td>Total fuel</td>
<td>Accumulated vehicle weight</td>
</tr>
<tr>
<td>Odometer</td>
<td>Driver id</td>
</tr>
<tr>
<td>Latitude &amp; longitude</td>
<td>Number of harsh brakes</td>
</tr>
<tr>
<td>Heading</td>
<td>Time overspeeding</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
</tr>
</tbody>
</table>

In particular the accumulated data is the sum of all measurements of a variable in a defined period of time. Usually the aggregation is triggered by the removal of the driver’s card from the reader. The vehicle operating data comes together with the design information, such as engine type, clutch system and wheels configuration among others.

2.1.2 Acquisition of road data

As described by Equation 1.1 also the road directly through its inclination angle $\alpha$ is influencing the force required to set the vehicle into motion and, hence, the fuel consumption. However the significance of the road is not only determined by its inclination; in fact the type of surface is playing a big role in the determination of the rolling resistance coefficient $f_r$. For these reasons road data is also collected.

OpenStreetMap has been chosen as source for the road data. The road data is publicly available online and it covers most of the Earth road network. OpenStreetMap data is downloaded for a bounding box covering the entirety of Sweden which is the main geographical scope for this study. The data is delivered in .shp (Shape files) [27]. Since in this thesis only the data regarding roads are of interest, it is separated from other data delivered by OpenStreetMaps. The road data is also structured in separate files depending on the road type. This enables an efficient usage of the data where only the road types of interest have to be considered during calculation and analysis. The list of road types are long but the most important are presented in Table 2.2.
Table 2.2: The most important road types [28].

<table>
<thead>
<tr>
<th>Road types</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Motorway</td>
<td>Fast, restricted access road</td>
</tr>
<tr>
<td>Trunk</td>
<td>Most important in standard road network</td>
</tr>
<tr>
<td>Primary</td>
<td>... down to...</td>
</tr>
<tr>
<td>Secondary</td>
<td>...</td>
</tr>
<tr>
<td>Tertiary</td>
<td>...</td>
</tr>
<tr>
<td>Unclassified</td>
<td>Least important in standard road network</td>
</tr>
<tr>
<td>Residential</td>
<td>Smaller road for access mostly to residential properties</td>
</tr>
<tr>
<td>Service</td>
<td>Smaller road for access, often but not exclusively to non-residential properties</td>
</tr>
</tbody>
</table>

In the obtained road data, the roads are divided into road segments of varying length. Each road segment has a number of properties that defines that specific element. In the data acquired for this study the road properties consist of the fields described in Table 2.3.

Table 2.3: Description of road properties.

<table>
<thead>
<tr>
<th>Property</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Name</td>
<td>The name of the road</td>
</tr>
<tr>
<td>Reference</td>
<td>The road identification e.g. &quot;E4&quot;</td>
</tr>
<tr>
<td>Type</td>
<td>The road type</td>
</tr>
<tr>
<td>One way</td>
<td>One way traffic only (y/n)</td>
</tr>
<tr>
<td>Tunnel</td>
<td>Going through a tunnel (y/n)</td>
</tr>
<tr>
<td>Max speed</td>
<td>The regulated maximum speed</td>
</tr>
</tbody>
</table>

2.1.3 Acquisition of weather data

As explained earlier environmental factors can have a large impact on the fuel consumption. In order to take these factors into account, environmental data is needed. There are two main types of data sources for weather data, both using data from weather stations. The first type derives from the interpolation of the data in order to enable the connection to positions other than the positions of the weather station. The second type of data comes from weather models that take measured weather data as initial values and then numerically estimate a large number of weather parameters at a large spatial scale.

After evaluating the availability and ease of use, the second type of data source is chosen. The weather model that generates the data is the Global Forecast System (GFS) which is produced by the National Centers for Environmental Prediction (NCEP) and published by the National Oceanic and Atmospheric administration (NOAA). The GFS model is composed by four standalone models covering atmosphere, ocean, land/soil and sea ice over the entire globe. The four models are connected in order to give a more accurate representation of the weather conditions [29]. The model data is structured in
grids where weather variables are constant over a grid cell. The resolution of the grid data that is used in this study is 0.5° in both longitude and latitude directions.

The model data is available in grid binary (GRIB) files. Since data from more recent years are used in this study that data is available in the second generation GRIB files (GRIB2). There are four datasets available per day, and the data is available from 2007-01-01 to the present date [30]. After downloading the needed GRIB2 files, they are managed using the wgrib2 software which has a lot of features, e.g creation of subsets by region and/or variables and data export [31]. In this case the grid data are regionally cropped and a selection of variables are exported to a readable .csv format. These files are then imported into the R environment where further analysis are conducted.

2.2 Missing features computation

The working dataset is the tool this analysis is based on. Thanks to the contributions of the different sources, it provides multiple pieces of information capable to represent an average truck ride. However some features are not reported by any source, for example the fuel consumption. Some calculations are required to complete the dataset with the four main features missing: fuel consumption, vehicle weight, front wind speed and side wind speed.

2.2.1 Fuel consumption

The fuel consumption is calculated as the ratio between total fuel and odometer. In particular, as reported in Equation 2.1, two consequent points (i and i − 1) are taken into account and the difference of their total fuel amount \( f \) and odometer distance \( d \) is computed. The resulting fuel consumption \( f_{ci} \) is not the instantaneous value of the two points, but it corresponds to the average value between the measurement i and i − 1.

\[
\bar{f}_{ci} = \frac{f_i - f_{i-1}}{d_i - d_{i-1}}
\]  

(2.1)

If there is no difference in total fuel between two consequent points, the feature is calculated using the first measurement presenting a variation in total fuel with respect to the initial value. To the measuring points located between them is assigned the same fuel consumption of points i and i − 1. Eventually there will be also point \( i + 1 \) that is contributing to the next calculation step. Point i will then have the same role as \( i - 1 \) in the previous step and it will be assigned a new value of fuel consumption.
2.2.2 Vehicle weight

As for the fuel consumption also the vehicle weight is computed as the average value \( w_{s,i} \) calculated over a determined period of time. The period of time can vary and it corresponds to the driver shift: the interval of time during which the same driver identification card is kept in the dashboard. The quantities that are used belong to the aggregated data group and are the accumulated calculated vehicle weight \( cvw_{s,i} \) and the odometer \( d_{s,i} \), as they appear in Equation 2.2.

\[
w_{s,i} = \frac{cvw_{s,i}}{d_{s,i}}
\]  

(2.2)

The accumulated vehicle weight is an aggregated value reporting a number depending on the load carried by the truck and the number of kilometers it ran during a driver shift. By applying Equation 2.2 the average vehicle weight during a shift is estimated by dividing \( cvw_{s,i} \) with the distance traveled during the shift \( d_{s,i} \).

2.2.3 Wind speed components

Among the weather data also the wind is reported, in particular inside the dataset the wind velocity is expressed by two numbers: speed \( W \) and angular direction \( \text{ang}(W) \). Since the focus is not the phenomenon itself, but the gust with respect to the vehicle, its components with respect to the vehicle reference system are calculated. Vehicle’s heading and wind angular direction do not share the same reference system. Both are based on the cardinal points of the Earth, but the former has the 0° angle that corresponds to the North Pole, while the latter’s reference system is rotated by 180°. The formula used are shown in Equations 2.3 and 2.4.

\[
w_{ls} = H \cdot W = |H||W| \cos(\text{ang}(H) - \text{ang}(W))
\]  

(2.3)

\[
w_{ts} = |H||W| \sin(\text{ang}(H) - \text{ang}(W))
\]  

(2.4)

The longitudinal component of the speed \( w_{ls} \) is obtained by the application of the scalar product of two vectors in Equation 2.3. \( H \) is the heading vector having length one and \( W \) is the wind speed vector. The angle inside the cosine operator corresponds to the angle of \( W \) with respect to \( H \). Similarly in Equation 2.4 the transversal component of the wind speed \( w_{ts} \) is calculated by applying the sine operator.

The interaction of the two vectors is illustrated in Figure 2.2.
2.3 Aggregating spatial data

The creation of the working dataset can be considered the foundation step of the analysis. At this point the four sets of data gathered from the sources interact with each other: vehicle data points, vehicle specification, road and weather data. The combination of the first two contributors is carried out without following any particular procedure; as a matter of fact both groups share a common information: the chassis number. Hence they are coupled using this feature.

A complete different approach is used for the aggregation of the vehicle data points to the road network and the weather grid. To complete this task their geographical position, that is determined through the coordinates, is used. The combination of coordinates and raw information is the mandatory first step of the aggregation. The vehicle data points, road network and weather data are treated in the form of, respectively, points, lines and grid elements. Some spatial data are treated before the actual interaction, in particular the road network lines are given a width in order to cover the road width, hence it becomes a network of polygons. This operation is done with the intention of letting more vehicle data points intersect with the road network, even if their GPS is reporting a slightly deviated position.
2.4. STATISTICAL METHODS AND PHENOMENA

Figure 2.3: Representation of spatial data.

Figure 2.3 shows an example of how the data looks like once they are in the spatial form, the grid corresponds to the weather data, the road data is represented by the white line on which it is possible to see some red points matching the vehicle data points. This procedure besides aggregating the information about road (Name, Max speed ...) and weather (Wind speed, Wind direction, Air temperature ...) performs another important action. It behaves as a filter for the points which are out of bounds, e.g. that lay outside the chosen roads, and for the roads that have no data points.

2.4 Statistical methods and phenomena

2.4.1 Linear regression

Linear regression refers to the modelling of the relationship between one dependent variable and one or more independent variables. If the regression only includes one independent variable it is referred to as simple linear regression and if there are several independent variables its referred to as multiple linear regression. The difference between simple and multiple linear regression can be seen in Equations 2.5 and 2.6 respectively. In both of the regression models, $\beta_0$ is the intercept which is the value of $y$ with the error subtracted when all of the independent variables are zero. $\epsilon$ is the error term which indicates the lack of an exact relationship between the dependent and the independent variables [32].

$$y_i = \beta_0 + \beta_1 x_{i1} + \epsilon_i$$

(2.5)

$$y_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_n x_{in} + \epsilon_i$$

(2.6)

The word linear in linear regression indicates that the regression model should be linear, i.e. there should be a linear relationship between the dependent variable and the
$eta$-values. However the definition does not restrict non linear relationships between the dependent and the independent variables [32], an example of this is presented as Equation 2.7. Also so called interaction terms are allowed in the scope of linear regression since its still linear in terms of $eta$-values. An interaction term is used to catch the combined effect of two independent variables on the dependent variable, an example is presented in Equation 2.8.

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i1}^2 + \epsilon_i$$ (2.7)  

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i1}x_{i2} + \epsilon_i$$ (2.8)

The independent variables in a regression model are usually of one of two types, either quantitative or qualitative. Qualitative variables compared to quantitative variables can only have a limited discrete number of possible values [33]. Qualitative variables could also be called categorical variables for that reason. One example of a qualitative variable is shown in Equation 2.9.

$$x = \begin{cases} 1, & \text{Category 1} \\ 0, & \text{Category 2} \end{cases}$$ (2.9)

In order to fit the data to the regression models a least squares approach can be used. The least squares method seeks to minimize the Residual Sum of Squares (RSS) which is given by Equation 2.10 [33]. In other words find values $\hat{\beta}_0 \cdots \hat{\beta}_p$ that minimizes the RSS function. Where $\hat{\beta}_0 \cdots \hat{\beta}_p$ are the estimates of the unknown constants $\beta_0 \cdots \beta_p$.

To determine how well the regression fit the data, the $R^2$ statistics can be calculated using the RSS defined in Equation 2.10 and the Total sum of squares (TSS) defined in Equation 2.11, the calculation is presented in Equation 2.12.

$$RSS = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = \sum_{i=1}^{n} (y_i - \hat{\beta}_0 + \hat{\beta}_1 x_{i1} + \cdots + \hat{\beta}_p x_{ip})^2$$ (2.10)

$$TSS = \sum_{i=1}^{m} (y_i - \bar{y})^2$$ (2.11)

$$R^2 = 1 - \frac{RSS}{TSS}$$ (2.12)

### 2.4.2 Model selection

Among $p$ different numbers of available predictors to use in the linear least squares model it is far from certain that every one of the $p$ predictors contribute to the model in the best possible way. It is possible that some of the predictors are not associated
at all with the response variable. In order to develop the best possible model from the
set of available predictors a variable selection process can be performed. For the linear
least squares model, model selection by subset selection is explored.

The most general form of subset selection is Best subset selection which computes
all models containing all possible combinations of predictors. This mean calculating
and comparing $2^p$ models. For obvious reasons this becomes computational heavy for
a large number of predictors e.g. $2^{50} = 1.1259e + 15$. For this reason this section
will concentrate on stepwise selection methods instead. The first method presented is
Forward selection which is a computational effective alternative to best subsection
selection. Forward selection considers a much smaller set of models than best subset
selection which makes it feasible from a computational point of view. Forward selection
start by considering the null model which contains no predictors and adds the predictors
one at a time, until all predictors are included in the model. At each step the predictor
which contributes to the lowest RSS, i.e. the best model fit, gets added to the model.
As a result of the nature of forward selection, a model $\mathcal{M}_{i+1}$ contains all the predictors
of the model $\mathcal{M}_i$ as well as one additional predictor. This means that forward selection
can fail to find the optimal model with $i + 1$ predictors since the optimal model doesn’t
necessary include all the predictors from the previous model. The opposite method
to forward selection is Backward selection, which uses a similar approach to model
selection but starts with all predictors in the model. In the backwards selection process
the predictor with the largest p-value (least significance) is removed from the model.
This is then done in an iterative fashion until the model only contains one predictor.
The last of the stepwise methods is a hybrid approach which is called Sequential
replacement which starts with no predictors in the same manner as with forward
selection. Stepwise a new predictor is included in the model but any variable that
does not contribute to the fit is excluded, hence it is the combination of forward and
backward selection [33].

2.4.3 Multicollinearity

Multicollinearity is a phenomenon that can occur in linear regression when there is
high correlation between two or more independent variables. The main consequences
of multicollinearity in linear regression models are concerning the following problems with
the least squares coefficient estimates: wrong signs, instability to slight changes in the
data and also false non significance. In the case of perfect multicollinearity one or more
columns of the $X$ matrix in the linear regression model written in matrix notation in
Equation 2.13 is a linear combination of one or more of the other columns. This means
that the least squares estimates $\hat{\beta}$ cannot be estimated since the matrix $(X^TX)$ from
Equation 2.14 is not invertible. From the linear regression perspective this can occur
when two or more independent variables correlates perfectly or when a independent
variable shows zero variation around it’s mean value.

$$ Y = X\beta + \epsilon \quad (2.13) $$
\[ \hat{\beta} = (X^T X)^{-1} X^T Y \]  
\[ (2.14) \]

One method to find and diagnose multicollinearity is to use the Generalized variance inflation factors (GVIF) [34]. Using the GVIF method one GVIF value is achieved for each independent variable in the model. In order to compare variables with different degrees of freedom, e.g. categorical variables one can adjust the GVIF value according to $GVIF^{1/(2 \cdot df)}$ [34].

### 2.4.4 Ridge regression

For data with high correlations among the variables it might not be possible to solve Equation 2.14 due to the badly conditioned $(X^T X)$ matrix. Even if it is solvable high correlation introduces large variances for the estimates of the regression coefficients. To account for this the linear least squares model is extended with a shrinkage term as presented in Equation 2.15. $\lambda$ is the shrinkage coefficient that determine the amount of shrinkage[35]. This extension of ordinary least squares regression is called Ridge regression and works by adding a penalty to the sum of the regression coefficients which can more clearly be seen in Equation 2.16. The larger the value of $\lambda$ the more shrinkage occurs i.e the coefficients goes towards zero. The solutions of the ridge regression (Equation 2.15) are not equivalent under scaling of the input variables, therefore it’s common to standardize the variables prior to regression[35]. In this context standardizing a variable is to subtract with its mean and divide by its standard deviation, as Equation 2.17 shows.

\[ \hat{\beta}_{ridge} = (X^T X + \lambda I)^{-1} X^T Y \]  
\[ (2.15) \]

\[ \hat{\beta}_{ridge} = \arg\min_{\beta} \left\{ \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 + \lambda \sum_{j=1}^{p} \beta_j^2 \right\} \]  
\[ (2.16) \]

\[ x_x = \frac{x - \bar{x}}{\sigma_x} \]  
\[ (2.17) \]

There are several methods to choose the value of $\lambda$, but one approach is to use k-fold cross validation. In k-fold cross validation the data set is split into k equally sized parts (folds) as seen in Figure 2.4. The model is then trained upon $k - 1$ of these parts and the last part is used as a validation set, the prediction error is then calculated on the model predicting the validation part. This procedure is done for all k parts.

| 1 | 2 | 3 | ... | k |

Figure 2.4: k-fold data split.
2.4.5 Estimating confidence intervals using bootstrapping

In order to assess the uncertainty in the Ridge regression coefficients’ estimate a bootstrapping approach can be used. The bootstrapping method consists of sampling $n$ number of data points from the training data set with replacement $R$ number of times. This means that $R$ number of data subsets are achieved, on which the regression model is trained. The result is $R$ estimates of the regression coefficients, which can be seen in Figure 2.5. Using the bootstrap samples, confidence intervals for the regression coefficients’ estimate can be computed. In this study two different methods for the estimation of confidence intervals from bootstrapped estimates are considered. The first method is denoted as the **Percentile** method and is carried out by using the $100 \cdot \alpha_C$ and $100 \cdot (1 - \alpha_C)$ percentiles directly from the bootstrap distribution, where $\alpha_C$ is the confidence level [36]. The second method considered for confidence interval estimation is the **Bias-corrected accelerated percentile (BCₐ)** method. The BCₐ method shows enhanced performance over the Percentile method for many cases, but it suffers from a higher calculation load since more calculation steps are carried out [36].

![Figure 2.5: Graphical representation of the bootstrapping procedure.](image-url)
2.5 Fuel consumption model

In order to estimate the drivers effect on fuel consumption, separated from other factors, a statistical regression model is developed. The model can be used to predict fuel consumption from a set of variables. The actual resulting fuel consumption prediction is not the main focus of this study, instead the effect of each predictor on the outcome is of higher importance. To explain the drivers effect on the fuel consumption the model development is divided into three different stages, where each stage adds more complexity to the model. In the first stage only drivers of the same vehicle on a specific road stretch are considered. This simplifies the situation since vehicle configuration parameters and variables describing the road are held constant. In stage two, data from several vehicles are added to the regression and also predictors connected to the vehicle configuration. At the same time the sample of considered roads is enlarged to all the main roads of Sweden. In the last stage predictors that describe the road are added.

2.5.1 Predictors

The predictors in the model can be divided into four groups depending on their role in the model. To catch the driver behaviour a qualitative variable here named *driver id* is used. The variable contains a unique number, representing a specific driver in the data. Since the variable is qualitative, dummy variables are used to represent each driver in the dataset. These dummy variables are $P_1 = \{d_1 \cdots d_{n-1}\}$ where $n$ is the number of unique drivers in the used dataset. The first group of predictors then contains the driver dummy variables. The second group of predictors contains all other predictors that are suspected to have a linear relation with the dependent variable in the model, i.e. predictors $P_2 = \{p_{lin,1} \cdots p_{lin,m}\}$. To catch non linear behaviour in the prediction, variables with exponents greater than one are contained in the third group $P_3 = \{p_{nlin,1} \cdots p_{nlin,q}\}$. In the last group predictors catching interaction effects between different predictors are contained, $P_4 = \{p_{int,1} \cdots p_{int,r}\}$.

2.5.2 Regression model

The fuel consumption model is a multiple linear regression model containing the predictors contained in the groups $P_1 \cdots P_4$. The foundation of the model can be written in the summarized form seen in Equation 2.18, where the $\beta$-values are real valued constants and $\epsilon$ is the error term.

$$f_{c,q} = \beta_0 + \sum_{i=1}^{z} \beta_i d_{iq} + \sum_{j=1}^{m} \beta_{(z+j)} p_{lin,jq} + \sum_{k=1}^{w} \beta_{(z+m+k)} p_{nlin,kq} + \sum_{l=1}^{r} \beta_{(z+m+w+l)} p_{int,lq} + \epsilon_q$$  \hspace{1cm} (2.18)
As explained earlier, the purpose of the fuel consumption model is to estimate the fuel consumption caused by the behaviour of a specific driver. $\beta_i$ gives the fuel consumption deviation from $\beta_0$, and gives the fuel consumption explained by the driving behaviour of driver $d_i$ as shown in Equation 2.18. This enables that the fuel consumption of several drivers can be compared and analysed in an objective sense, without the effects of the factors introduced in the fuel consumption model.

In order to get the best possible estimation of the $\beta$-values when using least squares regression given the data that the model is trained upon, the predictors must be chosen in a way which achieves the smallest RSS. In this study the predictors are chosen using the sequential replacement selection method and the final model having the least RSS is chosen. To ensure the best possible outcome from the model selection, it is executed using the data points connected to a road which is assessed to be representative in terms of the amount of data. If Ridge regression is used instead the model selection procedure is not committed, instead all predictors are included in the model.

### 2.5.3 Stage one application of fuel consumption model

In the stage one fuel consumption model application, the model learning procedure is repeated for each road segment. This means that if there are $r$ number of road segments, a total of $r$ fuel consumption models are trained. Each model contains the same predictors, chosen in the model selection. From the trained models the intercept and regression coefficients for the driver variables are extracted and used in the fuel saving potential analysis. The regression coefficients are reported with a confidence interval calculated using the Student’s t-distribution. Equation 2.20 shows the quantities involved in the calculation of the confidence interval of a regression coefficient according to the Student’s t-distribution approach. The individual driver coefficient is the deviation from the intercept, so the complete fuel consumption factor for a driver is given by Equation 2.19.

\[
f_{c,d_i} = \beta_0 + \beta_i
\]

\[
I_{\beta_i} = \left( \beta_i - \frac{t_{\alpha/2}(n - p - 1)\sigma_{\beta_i}}{\sqrt{n}}, \beta_i + \frac{t_{\alpha/2}(n - p - 1)\sigma_{\beta_i}}{\sqrt{n}} \right)
\]

Each driver coefficient is output by the regression function with their confidence interval, as seen in Figure 2.6. These coefficients have positive or negative sign, depending on the driver behaviour with respect to the intercept.
Figure 2.6: Example of drivers’ confidence interval in a road segment.

From the output it is possible to determine the fuel saving potential of the road segment. The fuel saving potential is a measure of the difference between drivers with high and low fuel consumptions on a road segment. The calculation procedure chosen is divided in four parts. The first step consists in finding the best driver and set that driver as the reference point of the analysis, represented in Figure 2.6 by the coefficient marked with a circle. The best driver is here defined as the driver having the lowest confidence interval upper limit located below zero. If none of the drivers fulfills this condition, the intercept becomes the reference point and a reference value of zero is used. The intercept value in this section of the analysis is always set to zero, as in Figure 2.6 where this change is represented by the purple mark on the x axis. In the second step the deviation of each driver from the reference point is calculated according to Equation 2.21. Where $I_0^+$ is the upper confidence bound for the reference driver and $I_i^-$ is the lower confidence bound for the $i^{th}$ driver. In Figure 2.6 this is represented by the dashed lines connected to the chained line.

$$\Delta_i = I_i^- - I_0^+ \quad (2.21)$$

The deviation of the $i$-driver corresponds to the difference between the reference point’s confidence interval upper limit and the $i$-driver’s confidence interval lower limit. If $\Delta_i$ is negative it means that the $i$-driver’s confidence interval is overlapping the reference one, hence it is not possible to exclude that the two considered values are the same. Then, the deviation $\Delta_i$ is set to 0. This case occurs also in Figure 2.6, in particular to the confidence interval marked with the rectangle. In the same road segment the deviation from the reference value can vary considerably and some values can lay far away from the rest of the population. The third step consists in the computation of the value representing the fuel saving potential of the road segment considered. The
fuel saving potential is solely based upon the distribution of \( \Delta \) values on the road segment considered. In the distribution, two boundaries are defined based on two chosen percentiles: 10th and 80th. Below the 10th percentile lays 10% of the population, in this case the drivers with the lowest fuel consumption. Above the 80th percentile is 20% of the population located, these are the drivers with the highest deviation from the reference point, i.e. the largest fuel consumption. An example of the distribution is seen in Figure 2.7, where the distribution is shown both in discrete and approximated continuous form. The two boundaries determined by the percentiles are shown through the two red dashed lines.

\[ fsp = \Delta_{80th} - \Delta_{10th} \] (2.22)
2.5.4 Stage two application of fuel consumption model

During stage two the working dataset sees an important increase in size; as a matter of fact more roads and vehicles are considered in the analysis. The number of different predictors grows, both quantitative and qualitative variables are added, as well as interaction terms.

As for stage one, the calculation of the confidence interval is required. For the purpose of the study the confidence interval of the $\beta$-value should be as small as possible. As shown by Equation 2.20 its limits are strictly related to the standard deviation, which depends on the variance of the measurements’ population. The variance $\sigma^2$ is described by the general formula in Equation 2.23, where $\bar{x}$ is the mean value of the measurements, $x_i$ is a generic element belonging to the population and $n$ corresponds to the total number of elements.

$$\sigma^2 = \frac{1}{n} \sum_{i=1}^{n} (x_i - \bar{x})^2$$  \hspace{1cm} (2.23)

In order to obtain results that have a low variance, a limitation on the minimum amount of observations per driver is introduced. At first drivers having less than 20 measurements in the whole dataset are removed. Then the requirement becomes more restrictive and also the drivers having less than 20 measures in the segment where the training of the model is carried out are removed.

During the pre-processing of the dataset two main tasks are carried out: the qualitative variables, such as driver id and day section, are encoded as dummy variables. In parallel the quantitative variables are standardized, according to Equation 2.17. This operation is performed because of practical reasons. In fact, by doing so, the intercept term is interpreted as the expected value of fuel consumption $f_c$ when the predictors are set to their means. Otherwise, the intercept would be interpreted as the expected value of $f_c$ when the predictors are zero, which may never happen as in the case of the vehicle weight. Table 2.4 shows the example of how generic values of Slope and Pressure change after their standardization.

<table>
<thead>
<tr>
<th>#</th>
<th>Slope[%]</th>
<th>Pressure[Pa]</th>
<th>Slope std[%]</th>
<th>Pressure std[Pa]</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>2</td>
<td>102201</td>
<td>-0.22962947</td>
<td>1.3057480</td>
</tr>
<tr>
<td>2</td>
<td>2</td>
<td>101268</td>
<td>-0.22962947</td>
<td>0.4187129</td>
</tr>
<tr>
<td>3</td>
<td>-2</td>
<td>101768</td>
<td>-0.43838353</td>
<td>0.8940801</td>
</tr>
<tr>
<td>4</td>
<td>2</td>
<td>99794</td>
<td>-0.22962947</td>
<td>-0.9826695</td>
</tr>
<tr>
<td>5</td>
<td>7</td>
<td>100398</td>
<td>0.03131311</td>
<td>-0.4084260</td>
</tr>
<tr>
<td>6</td>
<td>-7</td>
<td>98905</td>
<td>-0.69932611</td>
<td>-1.8270166</td>
</tr>
<tr>
<td>7</td>
<td>30</td>
<td>101265</td>
<td>1.23164898</td>
<td>0.4158607</td>
</tr>
<tr>
<td>8</td>
<td>30</td>
<td>100276</td>
<td>1.23164898</td>
<td>-0.5244155</td>
</tr>
<tr>
<td>9</td>
<td>-30</td>
<td>101998</td>
<td>-1.89966198</td>
<td>1.1127489</td>
</tr>
<tr>
<td>10</td>
<td>30</td>
<td>100402</td>
<td>1.23164898</td>
<td>-0.4046230</td>
</tr>
</tbody>
</table>
Standardizing before regression is a useful operation also in the case when the considered variables have different scales, as in Table 2.4. The standardization of the variables does not only exist in order to provide them with the same order of magnitude, but it positively affects the regression procedure itself. As a matter of fact it is an important step in the avoidance of multicollinearity [37].

The regression result is post-processed and further requirements for the considered segments and the accepted drivers are set. Among the goals of the analysis there is the determination of the fuel saving potential; for this achievement the fuel consumption coefficients coming from two different drivers are required. Thus directly after the regression, the segments with less than two drivers are disregarded.

2.5.5 Stage three application of fuel consumption model

During the third stage the dataset is coupled to a new information: the slope. At this point the aim is to have also the road contribution in the regression model, in particular through the slope. The road slope is derived from a map containing topographic information corresponding to the E4 from Luleå to Helsingborg. The dataset is then downsized by removing the data not laying on the road segments where vehicle data have been recorded for the time period of interest.

Multiple approaches are used at this stage, as well as different methods to estimate the coefficients. At first, the same approach as stage one and two is chosen: the measurements are grouped according to their road segments and a fuel consumption model is trained for each of them. Stage two and stage three have different outcomes; as a matter of fact the latter shows also the road contribution thanks to the $\beta$-value of the slope.

At the same time the Ridge regression is applied in order to covalidate the outcomes. This supplementary analysis is chosen based on its suitability for the cases where the regression coefficients have a large variance. In fact Ridge regression counteracts the collinearity of the predictors and it provides results with a shrunk variance. Together with Ridge also the bootstrapping method is used in order to assess the uncertainty of the training result and compute its confidence interval. The application of this method is limited to the approach that sees the training of the fuel consumption model occurring for every road segments being part of the E4. The reason for this choice is the high calculation load linked to the estimation of the confidence intervals using bootstrapping.
3. Results

3.1 Data acquisition

3.1.1 Acquisition of vehicle data and configurations

The acquisition of data and the creation of the dataset are two operations that use the combination of snapshot and accumulated data. It can be graphically represented based on the coordinates that every measurement point carries. An example can be seen in Figure 3.1, where several data points plotted on a map carries different pieces of information. In fact for the same point it is shown that the popup icon in every subfigure in Figure 3.1 can display, among others, the id of the driver, the speed and the weight of the vehicle.

(a) Driver id #.
(b) Longitudinal speed [km/h].
(c) Vehicle weight [t].

Figure 3.1: Graphical representation of the dataset.
3.1.2 Acquisition of road data

As described in the Method section, the road data downloaded from OpenStreetMap are separated by road type. In the analysis, the road network is built up by the road types \{ Motorway, Trunk, Primary \}. The complete road network used in the analysis is presented in Figure 3.2a. In Figure 3.2b the road stretch used in the last stage of the study is shown, where the road speed limit and its slope are considered.

(a) Road network used in stage two.  
(b) Road network used in stage three.

Figure 3.2: Sweden road network.

For the final phase in the fuel consumption modelling only the E4 between Luleå and Helsingborg is considered. For this road stretch high resolution altitude and slope data are acquired. A plot of these entities is presented in Figure 3.3.

Figure 3.3: Altitude and slope data for the E4 between Luleå and Helsingborg.
3.1.3 Acquisition of weather data

The weather data acquired consist of a very large data set covering 4 predictions a day for the entirety of the year 2014. For the sake of interpretation one prediction of each variable is shown in Figure 3.4.

Figure 3.4: Prediction of weather variables for 2014-01-10, 12:00 AM.
3.2 Fuel consumption model

As it is reported in Method chapter, the study has been performed in three stages, where several datasets and methods have been investigated and approached. In total four groups of fuel consumption models have been created; these four groups are sharing a road stretch that has reached the final step of the procedure, i.e. the determination of its fuel saving potential. The road segment is shown in Figure 3.5.

![Figure 3.5: Road segment common for every fuel consumption model groups in this work.](image)

The segment in Figure 3.5 belongs to the E4 and it is located in the Jönköpings län, tangent to the tätort of Skillingaryd. Its total length is of approximately 14.5 km.

3.2.1 Predictors

The predictors have been determined by the model selection and they correspond to the available variables able to develop the best possible model for least squares regression. The road stretch chosen to carry out the model selection is shown in Figure 3.6. As for the segment in Figure 3.5, also this stretch belongs to the E4 and it is located in the Jönköpings län.
The total length of the road shown in Figure 3.6 is of approximately 24 km. Thanks to its length this road segment has a lot measurements, so that the model selection’s result can be considered valid also for the other segments.

Table 3.1: Determined predictors.

<table>
<thead>
<tr>
<th>Type</th>
<th>Variables</th>
</tr>
</thead>
<tbody>
<tr>
<td>Qualitative</td>
<td>Driver id, Time of the day, Emission level</td>
</tr>
<tr>
<td>Quantitative</td>
<td>Vehicle weight, Elevation from the sea, Atmospheric pressure, Humidity, Temperature, Water equivalent accumulated snow depth, Frontal wind speed, Side wind speed, Slope</td>
</tr>
<tr>
<td>Combined</td>
<td>Humidity &amp; Temperature, Driven axles &amp; Total axles, Engine stroke volume &amp; Engine hp</td>
</tr>
</tbody>
</table>

Not all the predictors reported in Table 3.1 are used in every segment; as a matter of fact in some roads there may be variables that do not vary, e.g. *Time of the day*, and therefore can not be used to train the model.

### 3.2.2 Variables influence in the fuel consumption model

The training of the fuel consumption model does not simply rank the drivers based on their behaviour, but it gives also a measure of how the different predictors listed in Table 3.1 are influencing the dependent variable. The extent of this influence has been analysed for the road stretch in Figure 3.5, in particular the models trained in stage three are considered. Table 3.2 shows some of the variables’ coefficient that belong to the trained model.
Table 3.2: Example of computed predictors.

<table>
<thead>
<tr>
<th>Variable</th>
<th>Ridge [l/100km]</th>
<th>Least Square [l/100km]</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intercept</td>
<td>94.37</td>
<td>60.58</td>
</tr>
<tr>
<td>Vehicle Weight</td>
<td>0.18</td>
<td>2.6</td>
</tr>
<tr>
<td>Front Wind Speed</td>
<td>-0.1</td>
<td>-0.3</td>
</tr>
<tr>
<td>Slope</td>
<td>0.12</td>
<td>0.43</td>
</tr>
<tr>
<td>Driven axles &amp; Total axles</td>
<td>0.3</td>
<td>-20.20</td>
</tr>
<tr>
<td>Time of the day [08:00-12:00]</td>
<td>-1.83</td>
<td>-2.1</td>
</tr>
<tr>
<td>Temperature</td>
<td>-0.07</td>
<td>-1.07</td>
</tr>
</tbody>
</table>

Table 3.2 is listing the coefficients for quantitative, qualitative (in italic) and combined variables. Also the Intercept is reported. Some of the coefficients are positive, others are negative. If a variable has a negative numerical coefficient, e.g. Temperature, it means that at the increase of the variable value, the fuel consumption decreases. The qualitative variables are handled differently. For example, both in Ridge and in least squares when Time of the day [08:00-12:00] is present there is a decrease in the fuel consumption corresponding to the coefficient of the value itself.

### 3.2.3 Stage one

The application of the procedure according to what reported in the Method chapter leads to the identification of roads presenting a fuel saving potential. Both because of the limited amount of measurements and because of the smaller number of variables considered at this stage of the study, only five roads present a fuel saving potential. Their values are reported in Figure 3.7, where it is possible to see that three out five roads (Id = 4269586, 8031446, 345706184) have coefficients linearly dependent on others.

![Figure 3.7: Fuel saving potential of one single truck using least squares regression.](image-url)
Jönköpings län’s E4 road segment

Figure 3.8 shows two plots. Figure 3.8a reports the $\beta$-values assigned to each driver with their respective confidence interval. Its x-axis lists all the unique drivers who have driven the analyzed truck on the mentioned road segment; the y-axis has l/100km as unit measure and it corresponds to the $\beta$-values themselves. The driver id who has no confidence interval and the $\beta$-value equal to zero corresponds to the intercept of the fuel consumption trained model. The population of drivers has the coefficients varying between -2 l/100km and +6 l/100km. Moreover Figure 3.8a allows a first evaluation of the confidence intervals; excluding the intercept they have an average span of around 4 l/100km.

On the contrary, Figure 3.8b shows the deviation of each drivers’ $\beta$-value from the reference point that is represented by the best driver of the population. Hence, the higher the deviation is, the worse is the performance of the driver. It can be observed that eight drivers out of nineteen deviate from the reference value. The maximum deviation is 3.89 l/100km. The fuel saving potential of the road is calculated according to the procedure shown in Figure 2.7 and it corresponds to 1.14 l/100km.

![Image](image.png)

(a) Confidence interval of the drivers.  (b) Drivers’ delta from the reference value.

Figure 3.8: Stage one group - linear least squares regression.

3.2.4 Stage two

As described in the Method chapter, the stage two approach builds upon a large set of data containing many vehicles, more drivers and an increased road network. The resulting fuel saving potential for the different road segments containing an adequate number of observations are presented in Figure 3.9. It shows the fuel saving potential for each road segment in a polar coordinate manner where the length of each line corresponds to the fuel saving potential of the corresponding road segment. The color indicates whether the model on a road segment had issues with aliased coefficients i.e. perfect multicollinearity among some of the variables. Red color indicates the presence of aliased coefficients and green indicates absence of the mentioned issues. The road segment identification number is located at the outer rim of the plot.
Figure 3.9: Fuel saving potential [l/100km] using least squares regression.

Jönköpings län’s E4 road segment

Figure 3.10 presents the same content as Figure 3.8 but for the second stage approach with several vehicles involved. The majority of the driver population presents coefficient estimates that are contained in the interval -30 l/100km to +20 l/100km. Three drivers show considerably larger values for both the coefficient estimates and the confidence intervals. To note is that the confidence intervals of the previously mentioned drivers are overlapping all other drivers’ estimate and therefore they can not be distinguished from any of the other drivers, including the reference level. For this reason they show zero delta in Figure 3.10b. The fuel saving potential represented by the orange line in Figure 3.10b is calculated to 1.63 l/100km.
3.2.5 Stage three

A portion of the resulting fuel saving potential of the roads analyzed in stage three by using the linear least square is reported in Figure 3.11. The plot in Figure 3.11 is reported using polar coordinates in order to properly show the fuel saving potential of all 177 roads that are the outcome of the analysis carried out during stage two and stage three. Among all the roads reported in Figure 3.11 only twenty have not aliased predictors. The maximum value of fuel saving potential computed is 46.7 l/100km and the minimum is less than 0.01 l/100km.
Jönköping’s län’s E4 road segment

Despite the inclusion of the slope information Figure 3.12a looks similar to Figure 3.10a. In fact it has the same three drivers whose confidence interval is extremely wide. Among them the largest is reported by driver 601580, its lower limit is -44.39 l/100km and its higher limit is 522.96 l/100km.

Driver 599930, 601580 and 605029 shows a high value of their beta coefficient. However because of their wide confidence interval, they have no deviation from the reference value. This can be seen from Figure 3.12b. The inclusion of the predictor *slope* leads to have thirty drivers presenting a deviation from the reference point. The fuel saving potential of the road reaches 2.88 l/100km as it is depicted by the orange line in Figure 3.12b.

![Graph](image)

(a) Confidence interval of the drivers.  
(b) Drivers’ delta from the reference value.

Figure 3.12: Stage three group - linear least squares.

Figure 3.13 shows the fuel potential of 177 roads, determined using the Ridge regression. The results are consistently different with respect to Figure 3.9 and Figure 3.11, in particular the color chosen is not anymore an index dividing the roads based on the presence of aliased predictors. As a matter of fact Ridge regression prevents the appearance of aliased coefficients through the addition of the shrinkage term as described by Equation 2.16. The maximum value of fuel saving potential found in Figure 3.13 is 24.39 l/100km and the minimum is 0.01 l/100km.
Figure 3.13: Fuel saving potential using Ridge model with road slope.

Jönköpings län’s E4 road segment

Figure 3.14 presents the same content as Figure 3.10 and 3.12, but it shows the results obtained while using a different regression method: Ridge. The $\beta$-coefficients reported in Figure 3.14a have their confidence interval varying between 2.22 and 21.37 l/100km.

The shrunk confidence intervals allow the determination of sixty-six drivers who have a consistent deviation from the reference point. The maximum value is found for driver 725073 who presents a deviation from the population’s best element of 35.17 l/100km. The fuel saving potential of the road determined with the Ridge regression is 8.3 l/100km.
3.2. FUEL CONSUMPTION MODEL

Table 3.3: Fuel saving potential for different methods and stages.

<table>
<thead>
<tr>
<th>Method</th>
<th>Stage</th>
<th>Fuel saving potential [l/100km]</th>
<th>Comment</th>
</tr>
</thead>
<tbody>
<tr>
<td>least squares</td>
<td>1</td>
<td>1.1</td>
<td>One vehicle</td>
</tr>
<tr>
<td>least squares</td>
<td>2</td>
<td>1.6</td>
<td>Several vehicles</td>
</tr>
<tr>
<td>least squares</td>
<td>3</td>
<td>2.9</td>
<td>With slope</td>
</tr>
<tr>
<td>Ridge</td>
<td>3</td>
<td>8.3</td>
<td>With slope</td>
</tr>
</tbody>
</table>

3.2.6 Fuel saving potential’s geographical distribution

The fuel saving potential computed can be graphically represented in several ways. If the aim of the analysis is to compare the different roads, either a bar plot or a radial plot can work. But to get a better oversight over the road network a geographical approach is preferred. The map in Figure 3.15 is a perfect example of how the analysis results can be efficiently displayed to provide a geographical description of the results.

Figure 3.15 shows the roads presenting a fuel saving potential computed through the Ridge approach. The portion of E4 shown is located in southern Sweden. As it is possible to deduce by using the scale in Figure 3.15, the fuel saving potential has its lowest values around 0.1 and its peak over 20 l/100km. Among the segments represented there are many located in the Jönköping län presenting a fuel saving potential value close to the maximum.
Figure 3.15: Stage three on southern Sweden E4 - Ridge.
4. Discussion

4.1 Predictors

The determination of the relevant predictors that occurred during the model selection procedure is not supervised. The unique variable that has been forced in the model is the driver identification number, but all the others variables are determined upon their aptitude to decrease the Residual Sum of Squares.

Since the fuel is the primary energy source in a truck, the fuel consumption is directly related to the energy dissipation. A relation between the predictors listed in Table 3.1 and the variables reported in Equation 1.1 can be proven. In particular most of the terms that are in Equation 1.1 are clearly represented by some factors, such as the vehicle weight $m$ and the slope of the road $\tan\alpha$. Others, like the frontal wind speed and the atmospheric pressure, are contributing in the aerodynamic force. The number of driven axles, total axles, engine horse power and cylinder stroke volume can be defined as the primary causes for losses in the engine and in the driveline, according to the scheme in Figure 1.4. Also the emission level is indirectly influencing the driveline efficiency and the fuel consumption. Finally it can be also assumed that temperature, humidity and water equivalent accumulated snow depth, namely the weather related predictors, are related to the road surface condition and to a certain extent the rolling resistance coefficient $f_r$. Time of the day is unrelated from Equation 1.1, instead it is accounting for the influence of the traffic condition the truck had to face during its duty time.

The number assigned to each predictor depends on the road population and also on the method chosen, see Table 3.2. In general it is possible to say that the choice of the Ridge regression as method to train the fuel consumption model leads to smaller numerical coefficients. In Ridge the magnitude of the coefficients becomes a measure of the importance of the predictors for the model; so from Table 3.2 the slope coefficient (0.12) is more important than the temperature coefficient (-0.07). Another important coefficient is the front wind speed (-0.1), the negative sign may lead to the misconception that an increase in front wind speed intensity could decrease the fuel consumption. In reality it is the other way around, since this variable is taken with a negative sign with respect to the vehicle reference system; when the wind blows against the vehicle frontal area its velocity is negative. The same explanation is not valid for the intercept since its value corresponds to the average of all the fuel consumption values used as reference to train the population, according to Ridge. A major difference between Ridge and least
squares reported in Table 3.2 is seen for Driven axles & Total axles where the coefficient is not simply shrinking, but it presents also a change in sign, this is probably caused by the least squares model being affected by multicollinearity.

4.2 Comparison of approaches

In the results chapter the results of the different approaches to estimate the fuel saving potential are presented. If Figures 3.8, 3.10, 3.12 and 3.14 are examined, clear differences can be seen. These differences are propagated to the fuel saving potential, which is summarized in Table 3.3. The results from stage one seems to be the most restrictive in terms of fuel saving potential when the E4 road segment is analysed. When interpreting the results one must have in mind that stage one contains the least amount of data of the cases, i.e. less amount of drivers that contribute to the fuel saving potential. Since the fuel saving potential is based on the distribution of fuel consumption deltas, the results are definitely affected by the number of drivers included in the model. The fuel saving potential of stage two shows an increase in comparison to stage one. This seems reasonable in the light of the previous arguments. Stage two both add more vehicles and drivers, as well as data about vehicle configuration.

The results of stage three are interesting for several reasons, additional data about the road is added to give the model a better resolution and a better fit to the data. Also two different regression techniques are investigated, least squares and Ridge regression. If the radial fuel saving potential plots in Figures 3.9 and 3.11 are observed one can see that many models are suffering from issues with perfect collinearity or aliasing. The data has shown to suffer from multicollinearity which implies that the variance of the regression coefficients gets seemingly large when estimated using least squares regression. One of the main reasons for using Ridge regression is to minimize the variance of the regression coefficients under the presence of correlated data. If once again the information in Table 3.3 are considered one might wonder why the results of Ridge regression differs so much from the results of least squares. One explanation to this behaviour is that Ridge lowers the variance of each of the coefficient estimates. This results in confidence intervals that are smaller in comparison to those estimated using least squares. If the confidence intervals are shrunk then the separation of the coefficients of the different drivers is more likely to occur. If more drivers can be compared individually then the fuel saving potential has an potential to increase compared to the case where least squares regression is used.

One of the possible downsides of using Ridge regression in this thesis is the increased calculation load when estimating the confidence intervals. Since the confidence intervals play a main role in this study it has been a necessity to estimate them also for the case of Ridge regression. This results in, that models for a limited number of roads have been trained and used in the comparison with the least squares regression. The route used for the comparison is the E4 between Luleå and Helsingborg.
4.3 Fuel saving potential’s geographical distribution

Figure 3.15 proves the achievement of the study’s last two goals; as matter of fact the fuel saving potential has been spatially plotted based on the roads where it is computed. Since the fuel saving potential is computed using different methods it is possible to compare its geographical representation. The average trend shows the fuel saving potential computed by using the Ridge regression presenting a decrease in fuel saving potential with respect to least squares. This judgement is based on the analysis of the outcomes of the model training during stage three, where the measurements laying on the segments belonging to the E4 are analysed by using the two different methods. The mean value of the fuel saving potential of the 269 roads trained with Ridge is 3.91 l/100km, while least squares’ 243 roads show a result of 6.47 l/100km. The maximum values follows this trend and they are 24.39 l/100km and 46.7 l/100km respectively for Ridge and least squares. These pieces of information can be retrieved from Figure A.1 and A.2a. An interesting aspect showing the advantage of Ridge is the number of roads itself; in fact this approach managed to train the model for 269 roads, 26 more than least squares. This proves that Ridge is capable to determine smaller variation of $\beta$-values among the drivers.
5. Conclusions

This study set out to explore the idea of estimating the fuel consumption connected to the driving behaviour of drivers. To achieve this other factors affecting the fuel consumption are needed to be identified and accounted for. Finally, also the the concept of fuel saving potential were set out to be explored. More specific the goals of this thesis consisted in separating the fuel consumption caused by the driver from other factors. Secondly, calculate the fuel saving potential and visualize it spatially. Finally, geographic areas where the fuel saving potential was large were to be identified.

The importance of minimizing the fuel consumption have never been larger than now, both in the light of the economic aspects of running a transport business but also, even more important to contribute to a global more sustainable world. The first step in understanding how the fuel consumption can be decreased is to gain objective information on the contributing factors. Some factors affecting the fuel consumption are more easily controlled than others. The most fundamental factor to change is how the vehicle is controlled by the driver and that has been the focus of this study.

The results of this thesis show that it is possible to objectively report the drivers contribution to the fuel consumption, when the driver is considered as a part of a driver population on a specific road segment. This gives a measure to compare drivers that have driven on a common road segment and estimate the fuel saving potential of the road segment considered. By reconnecting the estimated fuel saving potentials to the spatial properties of the road network it is also possible to visualize the fuel saving potential in a way that makes it more understandable than figures in a table. Using these visualizations it is also possible to identify individual and groups of road segments that have a large fuel saving potential. These insights can be used as the foundation for further analysis of fuel consumption related issues.

The theoretical implication of this thesis is a new method on how to approach evaluation of drivers in the context of fuel consumption. Also, the introduced concept of fuel saving potential gives a new way of thinking about analysis of fuel consumption in a spatial context. This contribution to the theoretical framework in the area of vehicle driver analysis is an addition to a foundation which future work can leverage from. As explained in both the Method and the Results chapters several factors have contributed to limiting the scope of this thesis. High calculation load have limited the estimation of large scale fuel saving potentials using larger data sets and a more extensive road network.
Future work in this topic consists in applying the method considered in this thesis at a larger scale. This includes using more data and a larger road network. Also the weather data can be improved by finding another source with an increased grid resolution and a higher sampling frequency. Future work could also consist in developing the modelling approach further, to decrease the calculation load and increase the accuracy in separating the drivers fuel consumption factors from each other. Finally an exploration of the road properties that affect the fuel saving potential is valuable to gain a greater interpretation of the fuel saving potential itself.

Generally when discussing the fuel consumption connected to the driving behaviour an effort is made to understand what behaviour affects the fuel consumption and in a second stage how much. This thesis has contributed to an alternative way of thinking regarding driver related fuel consumption. This thesis presents a method which is not based upon assumptions of which driver behaviours that affects the fuel consumption. Instead, drivers can be compared upon the fuel consumption related to their driving behaviour and the fuel saving potential of the roads they have driven upon can be estimated.
References


http://files.shareholder.com/downloads/YRCW/1459836582x0x816524/

http://www.dpdhl.com/content/dam/dpdhl/Investors/Events/Reporting/
2015.

http://www.scania.com/group/en/scania-fuel-efficiency-duel-


Safety and Fuel Economy.*
https://www.aa.co.nz/assets/about/Research-Foundation/Ecodrive/


[23] National Research Council of the National Academies. *Tires and Passenger Ve-
hicle Fuel Economy.*


driven and expert analysis approach to determining parameters affecting fuel
economy”. In: *Elsevier* Transportation Research, Part D: Transport and Envi-
ronment (2014).

[27] *ESRI Shapefile Technical Description.*
REFERENCES


Appendices
A. Fuel saving potential visualizations

Fuel saving potential - road network visualization

Figure A.1: Stage three on Sweden E4 - Ridge.
Fuel saving potential [l/100km]:

(a) Sweden.

(b) Southern Sweden.

Figure A.2: Stage two - linear least squares.
APPENDIX A. FUEL SAVING POTENTIAL VISUALIZATIONS

Figure A.3: Stage three - linear least squares with slope.