Master of Science Thesis in Electrical Engineering & Industrial Engineering and Management
Department of Electrical Engineering & Department of Management and Engineering, Linköping University, 2016

Improving knowledge of truck fuel consumption using data analysis

Sofia Johnsen and Sarah Felldin

LINKÖPING UNIVERSITY

Master of Science Thesis in Electrical Engineering & Industrial Engineering and Management

Improving knowledge of truck fuel consumption using data analysis

Sofia Johnsen and Sarah Felldin

LiTH-ISY-EX--16/4961--SE

Supervisor: **Johan Dahlin**
ISY, Linköping University
Promporn Wangwacharakul
IEI, Linköping University
Jan Melin
Volvo Group Trucks Technology

Examiner: **Martin Enqvist**
ISY, Linköping University
Mattias Elg
IEI, Linköping University

Division of Automatic Control & Division of Quality and Business Development
Department of Electrical Engineering & Department of Management and Engineering
Linköping University
SE-581 83 Linköping, Sweden

"Logic will get you from A to Z; imagination will get you everywhere."

Albert Einstein

Abstract

The large potential of big data and how it has brought value into various industries have been established in research. Since big data has such large potential if handled and analyzed in the right way, revealing information to support decision making in an organization, this thesis is conducted as a case study at an automotive manufacturer with access to large amounts of customer usage data of their vehicles. The reason for performing an analysis of this kind of data is based on the cornerstones of Total Quality Management with the end objective of increasing customer satisfaction of the concerned products or services.

The case study includes a data analysis exploring how and if patterns about what affects fuel consumption can be revealed from aggregated customer usage data of trucks linked to truck applications. Based on the case study, conclusions are drawn about how a company can use this type of analysis as well as how to handle the data in order to turn it into business value.

The data analysis reveals properties describing truck usage using Factor Analysis and Principal Component Analysis. Especially one property is concluded to be important as it appears in the result of both techniques. Based on these properties the trucks are clustered using k-means and Hierarchical Clustering which shows groups of trucks where the importance of the properties varies. Due to the homogeneity and complexity of the chosen data, the clusters of trucks cannot be linked to truck applications. This would require data that is more easily interpretable. Finally, the importance for fuel consumption in the clusters is explored using model estimation. A comparison of Principal Component Regression (PCR) and the two regularization techniques Lasso and Elastic Net is made. PCR results in poor models difficult to evaluate. The two regularization techniques however outperform PCR, both giving a higher and very similar explained variance. The three techniques do not show obvious similarities in the models and no conclusions can therefore be drawn concerning what is important for fuel consumption.

During the data analysis many problems with the data are discovered, which are linked to managerial and technical issues of big data. This leads to for example that some of the parameters interesting for the analysis cannot be used and this is likely to have an impact on the inability to get unanimous results in the model estimations. It is also concluded that the data was not originally intended for this type of analysis of large populations, but rather for testing and engineering purposes.

Nevertheless, this type of data still contains valuable information and can be used if managed in the right way. From the case study it can be concluded that in order to use the data for more advanced analysis a big-data plan is needed at a strategic level in the organization. The plan summarizes the suggested solution for the managerial issues of the big data for the organization. This plan describes how to handle the data, how the analytic models revealing the information should be designed and the tools and organizational capabilities needed to support the people using the information.

Acknowledgments

Our first thanks goes to our supervisor at Volvo Group Trucks Technology, Jan Melin, for giving us this opportunity, teaching us all about trucks and guiding us through the impressive and sometimes confusing world of Volvo. We would also like to thank the many employees at Volvo Group who generously gave of their valuable time for meetings, trainings and interviews. Without your contribution we would not have been able to perform the data analysis or written this report. Especially to the team at Vehicle Productivity – all our lunches, fikas and subsequent discussions were always a nice break from MATLAB.

The academic guidance given by Promporn Wangwacharakul and Johan Dahlin, our supervisors from Linköping University, as well as Mattias Elg and Martin Enqvist, our examiners, has been important. Without your feedback and ideas, this report would not be as interesting to read. Your enthusiasm for our idea to make a multidisciplinary master thesis considering this is not very usual has been encouraging.

Finally, we would like to express our gratitude to our friends and families for supporting us during the writing of this thesis and during our time at Linköping University. Thank you for all the encouraging pep talks when we thought we would never finish, for telling us to keep going and never losing hope on us.

Göteborg and Stockholm, June 2016
Sarah Felldin and Sofia Johnsen

Contents

List of Figures

List of Tables

Notation

Abbrevation	Meaning
CFFU	Customer Fuel Follow-Up
CLT	Central Limit Theorem
CV	Cross-validation
EM	Expectation maximization
EUROFOT	European Field Operation Test
FA	Factor Analysis
GCW	Gross combination weight
GTT	Group Trucks Technology
LAT	Logged Vehicle Data Analysis Tool
LS	The Least Squares Method
LVD	Logged Vehicle Data
MSE	Mean Squared Error
OEM	Original quipment manufacturer
PC	Principal Component
PCA	Principal Component Analysis
PCR	Principal Component Regression
PTO	Power take-off
RSS	Residual Sum of Squares
SVM	Support Vector Machine
TCO	Total cost of ownership
TQM	Total Quality Management

1
Introduction

This chapter describes the importance of knowledge about customer needs and how big data can be used in order to increase this knowledge. It further describes the purpose and scope of this study and discusses previous work in the field in order to explain the approach.

1.1 Knowledge about the needs of the customer

In markets with high competition customer satisfaction is important to keep loyal customers and to attract new. A company needs to provide its customers with the right product or service according to their needs in order to contribute to their customers' productivity and thereby increase their satisfaction.

In order for a company to be able to meet their customers' needs it is essential to find out what the customers want and to have detailed knowledge about the customers' habits and desires. However, it is not enough to ask the customers what they need, because what they think they need might not always correspond to their actual usage and they might not be able to link up their needs with the latest technological opportunities. [Bergman and Klefsjö, 2010]

By knowledge, imagination and innovation a company can surprise the customers and fulfill needs the customers did not know they had until they were provided with a new product or service. This is to exceed the customers' expectations. The needs of the customers are constantly changing and it is therefore important for a company to not only fulfill the present needs of the customers, but also their future needs. [Deming, 1986]

The Volvo Group is one of the world's leading manufacturers of trucks, buses, construction equipment and marine and industrial engines. Trucks is the biggest part of the business with more than 200,000 trucks sold in 2014 [Volvo Group, 2014] and Volvo Group Trucks is a good example of a company that can use data generated by their products in order to find information about how their customers actually use their products. This information could then be used in product development at Volvo Group Trucks Technology (GTT) to improve the products according to the needs of the customers.

In mature markets fuel costs dominate the total cost of ownership (TCO) while the acquisition price still plays a dominant role in the TCO over the lifetime of a truck in emerging markets. However, fuel costs, service and repair are becoming increasingly important in the emerging markets because of increasing oil prices over time and the technology becoming more and more complex which creates a need for qualified technicians. Fuel costs are the largest TCO component in mature markets like Western Europe (30 percent) that can be influenced by the manufacturer. [KPMG International, 2011]

Volvo Group Trucks needs to know how their customers use their trucks in order to improve the trucks and services they provide today to increase customer satisfaction. Since decreasing fuel costs are an important key to increased profitability of commercial truck operators, fuel consumption is one of the key features Volvo Group Trucks are working with to improve on their trucks. An analysis of the customers' actual usage of the trucks with respect to fuel consumption could be done in order to improve the customers' experience of this feature. However, this is not done today on a larger scale to see correlation of different types of customer usage to fuel consumption. To further increase customer satisfaction it is also important to exceed their needs by identifying new fuel saving opportunities to create business value. Large sets of customer usage data, so called big data, can be a huge asset for a company in their quest of fulfilling their customers' needs. Taking this into account, this thesis aims at extracting the value adding information residing in Volvo Group Trucks's databases containing big data of customer usage of their trucks.

Big data has a large potential in diverse industries, in everyday lives, in research, in governmental activities etc. There are various examples of the value of big data for healthcare, urban planning, intelligent transportation, environmental modeling, smart materials, machine translation between natural languages, education, computational social sciences, systemic risk analysis in finance, homeland security, computer security, and so on. There are also examples of the value of big data for energy saving, through unveiling patterns of use. [Jagadish et al., 2014]

Although the value and potential benefits of big data are real and significant, there are many technical challenges that must be dealt with in order to fully realize this potential, which are not only connected to the volume of the data, but also its variety, velocity and veracity [Jagadish et al., 2014].

The terms *data mining* and *machine learning* are today often mentioned together

with big data. Data mining is the process of discovering patterns in the data whereas machine learning is the field from which many of the techniques used in data mining are taken from [Witten, 2011]. Both data mining and machine learning, and what resides within these fields, can therefore be considered as being part of the data analysis techniques used when dealing with big data.

To gain value adding knowledge about the products using these data analysis techniques a wider approach is taken including both an evaluation of different techniques as well as an organizational perspective of how to handle and use the data. The technical challenges connected to big data need to be considered in order to reveal the wanted knowledge.

Since big data has such large potential if handled and analyzed in the right way, and since Volvo Group Trucks has access to large sets of customer usage data, an analysis of how to fully capitalize the data is made in this thesis. The in-depth analysis of the data itself in regard to fuel consumption of trucks uses different data analysis techniques for handling the technical challenges of big data. Many organizations have access to or can collect customer usage data today and many also use data-driven insights in their decision making process, which is a fast growing trend in recent years [IBM, 2014]. Therefore this analysis of how to exploit such data is a good example of big data analytics and could be applied on many organizations also in different industries. The application on fuel consumption of trucks is a case study in the automotive industry which shows that value adding information in the form of patterns in customer usage and how they affect a certain performance can be revealed.

1.2 Purpose & objectives

The purpose of this thesis is to evaluate how and which data analysis techniques that can be used to extract value adding information from large amounts of aggregated customer usage data. Moreover, this thesis will investigate how a company needs to handle the data in order to be able to use this value adding information to increase customer satisfaction of their products and services.

1.2.1 Research questions

The purpose is investigated in the form of performing a case study at a truck manufacturing company and the data used is logged vehicle truck data, see Chapter 5. The focus lies on extracting information about what affects fuel consumption from this logged vehicle data and how and if it differs depending on how the trucks have been used. In the case study the following research questions are answered:

1. Can relevant patterns linked to truck applications be found in logged vehicle data?

2. If so, which factors influence these patterns the most?

3. Is the importance of the fuel consumption parameters the same for all applications?
4. If not, in what way do they differ?
5. To what extent can logged vehicle data be used to draw these conclusions?
6. What technical issues of big data need to be taken into consideration in the data analysis?
7. How can managerial challenges concerning this kind of data be pinpointed by using the Total Quality Management cornerstones?

1.3 Previous research

Prytz et al. [2013] stated that analyses of large amounts of data on-board trucks is getting more and more achievable but until the technology is mature enough, analyses of already existing data is required. Furthermore, the study described an investigation of how on-board truck data can be used for predicting truck compressor failure by investigating data mining techniques. Grubinger [2008] discussed how knowledge can be extracted from logged truck data, especially concerning differences in the operating environment of the trucks by using unsupervised learning methods. In a later published article by Grubinger et al. [2009], the possible knowledge extraction from the information available in real-world logged data from Volvo long haul trucks and the problem with handling this vast amount of data was further analyzed with recommendations for an automatic application. Here, Grubinger et al. again stressed the importance of differences in the operating environment of the trucks, together with the customer usage, especially to find trucks which had been used differently than what they were specified for.

Customer usage has been shown to have a significant impact on fuel consumption and several studies have been made to find the key factors. Important "big-picture items" in heavy-vehicle fuel consumption was found to be vehicle configuration, traffic congestion, speed limits, payload factors, and use of regenerative braking [Hunt et al., 2011]. The four latter of these factors are external and can be connected to customer usage. Ribeiro et al. [2013] presented an innovative model for estimating instant fuel consumption from a smartphone's GPS data alone. However, Alessandrini et al. [2006] stated that not only the drive cycle affects the fuel consumption, but also driver behavior, and suggested a new definition of driver behavior linked to the way the driver uses the pedals.

McGordon et al. [2011] stated that one of the major influences of real-world fuel economy is driver behavior, but that this is difficult to study. Their model was a simulation driver model based on data obtained through real-world data and showed that logging can provide a good representation of real-world driving behavior in terms of the vehicle speed.

In a paper which is a part of an on-going Ph.D. thesis, Carpatorea et al. [2014]

asserted that rich datasets of actual vehicle usage are available and a data-mining approach can be used to not only validate earlier hypotheses, but also discover unexpected influencing factors. This study focused on how driver behavior affects fuel consumption and presented a *base value*, which will be used to categorize drivers' performance more accurately than previously used methods in order to exhibit different driver and fuel consumption characteristics.

In another study, Factor Analysis was used to reduce the initial 62 parameters to 16 independent driving pattern factors. Regression analysis on the relation between driving pattern factors and fuel-use and emission factors showed that nine of the driving pattern factors had considerable environmental effects for cars. [Ericsson, 2001]

From data gathered in real traffic conditions using advanced vehicle location systems in cars, one conventional and one hybrid electric, driving features have been extracted to investigate their influence on fuel consumption and emissions. Two superior driving features, "energy" and "idle time percentage" were found and used for clustering of driving segments. [Montazeri-Gh et al., 2011]

Other studies have been made concerning the use of machine learning methods when processing large amounts of data. Hsu et al. [2009] have established regular rules for the development of sizing systems of body types from the anthropometric data of females effectively by using a fuzzy clustering-based data mining approach. Cho et al. [2009] have used a classifying algorithm based on Support Vector Machine (SVM) and k-means Clustering in order to classify vehicles based on radar signals. Furthermore, there are several studies made on data mining applications for quality improvement in the manufacturing industry [Köksal et al., 2011].

Furthermore, decreasing fuel consumption of trucks does not only affect the original equipment manufacturers (OEM) and the commercial truck operators, but the whole society since the world is facing a number of global challenges, including climate change. There is a widespread scientific consensus that the global climate is changing and that human activity contributes significantly by greenhouse gas emissions, which are mostly caused by the burning of fossil fuels, including coal, gas, oil and diesel, in industry, transport, agriculture and other vital economic sectors [World Meteorological Organization, 2013].

Population growth, augmenting industrialization and urbanization in combination with economic growth place increasing demands on the use of the finite capital of the planet. Resource efficiency and finding ways to reuse materials and energy in product life cycles is increasingly important for the transport industry. A sustainable transport sector must therefore respond by improving fuel efficiency of heavy vehicles. [Volvo Group, 2013]

The real-world fuel consumption of vehicles is becoming increasingly important for manufacturers as well as consumers [McGordon et al., 2011]. There are examples from the automotive industry of reasons for shifting towards manufacturing a sustainable product, e.g. the shortage of fossil resources and the resulting oil

price increase, new legal requirements which penalize environmental pollution and the changing behavior patterns of consumers [Hetterich et al., 2012]. The relevance and pressure to substitute fossil materials with renewable ones can be expected to increase and will not only be due to the potential decline of resources, but more notably as a result of customer demand [Hetterich et al., 2012].

1.4 Approach

The novel approach of this thesis is to derive value adding information from large amounts of customer usage data already being logged today. In the case study performed this is applied to find patterns in customer usage of trucks and link the involved parameters to fuel consumption. This thesis also includes a Total Quality Management (TQM) and organizational perspective with focus on how the studied organization can use the customer usage data to increase customer satisfaction, which put together is a novel and holistic approach compared with previous research. In Figure 1.1 these two perspectives of data analysis techniques and TQM are applied on big data.

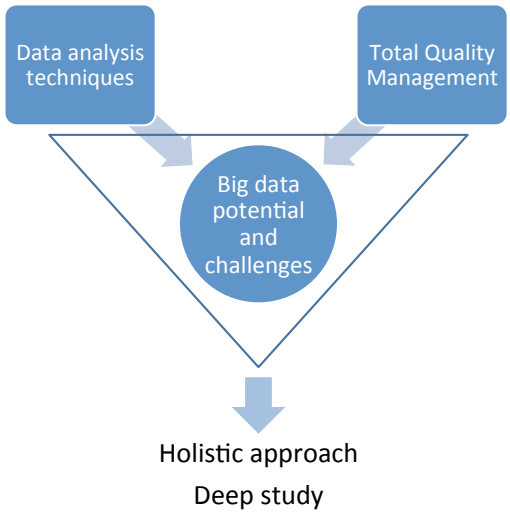

Figure 1.1: *The approach of this thesis contains two different perspectives applied on big data, resulting in a holistic approach containing a deep study.*

TQM was chosen as basis for the organizational perspective since it is a holistic concept of how to combine values, methodologies, tools and people in order to increase customer satisfaction, see Chapter 3. This fits well in this thesis since the goal is to create value from this data by integrating the usage of data analysis techniques in the organization aswell as making the techniques suited for the organization. In turn, the data analysis techniques used were chosen so that the

big data could be utilized, addressing the different challenges of big data, which are further described in Chapter 4. This gives the thesis a holistic approach as well as permitting a deep study in the form of an extensive analysis of the big data.

This deviates from for example the work in Grubinger [2008] and Grubinger et al. [2009], where the purpose instead was to find vehicles with usage deviating from what they were originally specified for. Ribeiro et al. [2013] also had a different approach since their fuel consumption model was based only on smartphone GPS data and for cars. Moreover, Ribeiro et al. did not use existing logged vehicle data, but generated new data from an on-board device. The purpose of this thesis is also different from the one by Carpatorea et al. [2014] where the development of a *base value* was connected to the performance of an individual driver as well as the operational environment. The findings and methods in Ericsson [2001] and Montazeri-Gh et al. [2011] are highly related to the work in this thesis, but had other aims and the data was extracted from cars in urban traffic.

1.5 Scope of study

The scope of this study is to investigate how large amounts of logged vehicle data can be used to its full potential including data analysis techniques to analyze the data and the application of the TQM cornerstones to deal with the managerial challenges of the big data. The scope does not include implementation or testing of this methodology in the organization.

Interviews were conducted with employees about customer needs and feedback that were already known to the organization. These interviews served as the basis for the knowledge about the needs of the customers of Volvo Group Trucks for this thesis and therefore a further investigation of customer needs was not included in the scope of this study.

Furthermore, the focus of this thesis was to analyze a set of parameters which were suspected to have an impact on fuel consumption. Which parameters to include were partly decided upon by conducting interviews with employees with long experience of trucks and partly by a literature review of previous work in the field, which was briefly described in Section 1.3.

The term customer usage does not include individual driver behavior in this study, but refers to patterns in larger populations.

2

The need at Volvo Group Trucks Technology

This chapter introduces Volvo Group Trucks Technology where this thesis was conducted and explains the setting and specific background of the thesis connected to the company such as the Volvo Group vision of sustainable transport solutions, Volvo Group Trucks' strategy, a presentation of some available databases and what type of data they contain.

Volvo Group is one of the world's leading manufacturers of trucks, buses, construction equipment, and marine and industrial engines. Trucks is the biggest part of the business with more than 200,000 trucks sold in 2014 [Volvo Group, 2014] and which makes Volvo Group one of the largest truck manufacturers in the world. Their portfolio of truck brands includes the Volvo, Mack, UD, Renault, Dongfeng and Eicher brands. With a portfolio this wide, Volvo Group can attract customers in different market segments. All of the brands offer efficient and economic solutions for a wide range of applications for long-haul, regional and city distribution and construction purposes. [Volvo Group, 2013]

Volvo Group Trucks Technology (GTT) has the global responsibility for the Volvo Group technology research, engine development, product design and all the technology and product development linked to truck operations, as well as supporting the products in the aftermarket [Volvo Group Trucks Technology, 2016]. This includes development of on-board and off-board (back office) applications designed for improving the fuel efficiency of the trucks according to the philosophy of the Volvo Trucks brand: "Every drop counts", see Section 2.1.

2.1 Sustainable transport solutions

Volvo was founded in 1927 with the mission to build vehicles with the core values quality and safety. In 1972, Volvo added care for the environment as a core value which put them in the forefront of the industry. At the time when this thesis is written Volvo Group's vision is to become the world leader in sustainable transport solutions. [Volvo Group, 2013]

For further reading about Volvo Group's core values quality and environmental care, see the Volvo Group's Environmental Policy and the Volvo Group's Quality Policy [Volvo Group, 2012a,b].

Sustainable transport solutions consist of, according to Volvo, three dimensions: environmental, economic and social sustainability. Environmental sustainability implies energy-efficient transport solutions with very low emissions of carbon dioxide, particulate matter, nitrogen oxides and very low levels of noise. [Volvo Group, 2013]

Economic sustainability is the second dimension of sustainability and means that in order to contribute to high productivity in the transport system, the company must provide the customer with the right product or service [Volvo Group, 2013]. This is an important part of this thesis since it aims to improve knowledge of truck fuel consumption in order to increase customer satisfaction. Meeting the customer's needs can very well be combined with developing environmentally sustainable products. Improving fuel efficiency is a good example of when these go hand in hand. Reducing fuel consumption in heavy trucks benefits both the fleet owners and the environment through lower fuel costs and fewer emissions [Volvo Group, 2013].

In order to combine improving fuel efficiency in trucks and providing the customer with the right product, it is important to understand how fuel efficiency coincide with the business and operations of the customers. Many of Volvo Group Trucks' customers are haulage contractor companies. Figure 2.1 illustrates how a haulage contractor company can drive fuel efficiency in their organization. It all starts with the vision and mission of the company [Söderman, 2014]. However, it is important to remember that the main mission of a haulage contractor is not to consume as little fuel as possible; in that case the best solution would be to not drive trucks at all. Their main mission is to deliver the right goods in the right time to their customers, see e.g. [Skoogs Åkeri och Logistik, 2014], [Andreassons Åkeri, 2014] or [Foria]. To be fuel efficient is however in their interest, since it will save fuel costs. In order to be fuel efficient, the haulage contractor needs to state this in their vision or mission somehow, otherwise it will not be a prioritized matter [Söderman, 2014].

When fuel efficiency is stated in the vision and mission of the company the next step is to enable their drivers to drive fuel efficiently. The company therefore needs to educate their drivers in eco-driving, but it does not end there. In order to show the employees that fuel efficiency really is an issue the company believe

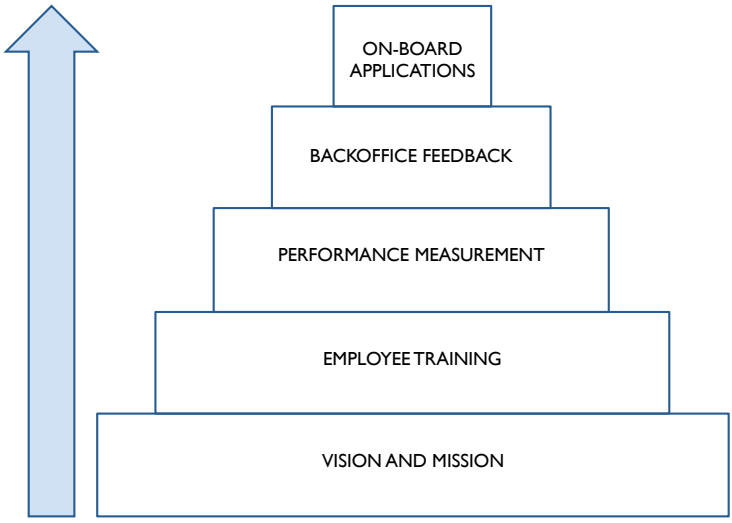

Figure 2.1: *The means for driving fuel efficiency in a haulage contractor company [Söderman, 2014].*

is important, some kind of performance measurement system needs to be put in use [Söderman, 2014]. A performance measurement system enables informed decisions to be made and actions to be taken because it quantifies the efficiency and effectiveness of past actions through the acquisition, collation, sorting, analysis and interpretation of appropriate data [Neely et al., 2002]. Therefore the truck needs to be able to provide measurements and transfer them to the back office, which is the fleet management in this case. One of their tasks is to evaluate each driver's performance and to give feedback to the drivers. Here it is important to figure out incentives for the drivers for striving to be fuel efficient, and also to consider that all drivers may not be motivated by the same incentives [Söderman, 2014].

On top of this are all the on-board applications available in today's trucks such as automated gearbox, cruise control, etc. These applications are designed to help an unexperienced eco-driver to be able to become at least an intermediate eco-driver, as well as to make the driving experience comfortable [Söderman, 2014]. This together with the previous steps in the pyramid are the means a haulage contractor have to drive fuel efficiency in their company.

Volvo Group Trucks provides services which measure the performance of the truck and the driver and send this information via telematics[1] to the back office. There, the software evaluates the performance of the driver and gives the

[1] Telematics is the integrated use of telecommunication and informatics used for application in vehicles and control of vehicles on the roads.

fleet management tools to give feedback to the driver. Volvo trucks also have a number of on-board applications to help the driver drive more fuel efficiently such as the automated gearbox I-Shift and the Cruise control function. Another fuel saving invention is the I-See system which stores the gradient of the roads the truck drives on and shares it to other Volvo trucks via the I-See cloud. When approaching a hill, the truck automatically downloads the topography information of the hill and uses this information to optimize the I-Shift transmission, the engine and the speed to maximize the use of the truck's own kinetic energy in order to save fuel [Volvo Trucks, 2014b]. The fuel savings of these applications are for I-Shift reduced fuel consumption of up to 7 percent compared with a manual gearbox, I-See which can save up to 5 percent fuel and Fuel Advice which is a coaching service used for continuous follow-up and feedback of drivers' fuel efficiency performance and help transport companies cut fuel consumption with up to 5 percent [Volvo Trucks, 2014a]. These are some examples of innovations that GTT already have developed in order to make both the trucks more fuel efficient as well as helping the customers to use them fuel efficiently.

The focus is shifting from product to service based operations in the automotive industry, as well as in many others [Prytz et al., 2013]. There is now a trend where customers are buying services rather than goods, and correspondingly, Volvo Trucks has shifted its core activity from manufacturing trucks to "creating trouble-free transport". Therefore it is extremely important to focus on handling and improving service quality [Bergman and Klefsjö, 2010]. This is also amplified by the fact that service and repair costs are increasing, especially in emerging markets [KPMG International, 2011], as discussed in Section 2.1. Volvo Group offer therefore their customers service contracts where all service and repair is included for a fixed price – hence selling "trouble-free transport".

There is an example of this "trouble-free transport" in Volvo Trucks in North America who are offering a service called Remote Diagnostics. It is designed to benefit the customer with real uptime management and real downtime protection. Remote Diagnostics provides proactive diagnostics and repair planning assistance with detailed analysis of diagnostic fault codes. The service includes taking care of the whole information chain in order to secure the uptime for the customer: Driver to Vehicle, Vehicle to Volvo, Volvo to Decision-Maker, Decision-Maker to Dealer, Dealer to Driver in order to get the truck back on the road as soon as possible. This is made possible by diagnostics fault codes (data created to explain a fault) being sent automatically when they occur in the truck via telematics to the Volvo back office. [Volvo Trucks Support Services, 2012]

2.2 Volvo Group Trucks strategy 2013–2015

To be able to take steps in the direction of the Volvo Group vision, Volvo Group Trucks have set up their strategy for 2013–2015. This strategy shows that Volvo Group Trucks have some focus areas which go in line with the purpose of this thesis in the areas of creating business value by focusing on customers and driv-

ing innovations for new business opportunities, where energy-efficient transport solutions play an important part for the future.

Volvo Group Trucks are convinced that their success is based on being the best at solving their customers' problems and strengthening their operational performance. This is considered a key factor in building customer loyalty and becoming their customers' preferred business partner and therefore, customer focus is important for Volvo Group Trucks. [Volvo Group Trucks, 2013]

Another important focus for Volvo Group Trucks is to capture profitable growth opportunities. Volvo Group Trucks want to retain and strengthen their position as a profitable and global player in the truck industry. This is crucial given that high volumes help them achieve economies of scale and maintain their priority position among suppliers and retailers. The potential for new business, and for expanding the current offering, in areas such as vehicle productivity and vehicle management have been recognized and should be put into business value. [Volvo Group Trucks, 2013]

Finally, environmental concerns, political demands, megacities and fuel prices are driving regulation and green technology. To be able to anticipate and act on changing market demands and shifts in technology, and have the capacity to rapidly bring new solutions to market is important for Volvo Group Trucks. One focus is to improve fuel efficiency through vehicle optimization, diesel efficiency and electromobility[2]. Volvo Group Trucks have stated that they need to pursue fuel-efficiency improvements and optimization of their vehicles and the existing diesel engine platform, and that they also must continue to develop hybrid solutions and alternative drivelines. [Volvo Group Trucks, 2013]

To commercialize alternative fuel technology by launching concepts or products in all regions is also important for taking steps in the right direction. This is about not only inventing new ideas, but also to turn them into commercial viable products and put them into market. In order for this to succeed, Volvo Group Trucks want to work in close partnership with customers and providers of infrastructure and alternative fuels. [Volvo Group Trucks, 2013]

This strategy shows that Volvo Group Trucks has focused on the issue of fuel efficiency and identified that it needs to be dealt with from different perspectives and in collaboration with other stakeholders such as customers and providers of infrastructure and alternative fuels. It also shows that these three areas are strongly connected and that it is not only possible, but perhaps even necessary to combine them to reach success: creating business value when capturing profitable growth opportunities by increasing customer satisfaction through innovating and commercializing energy-efficient transport solutions, which this thesis is meant to contribute to.

[2]The electromobility market includes fully electric vehicles and machines – powered or propelled solely by an electric motor – as well as hybrids, which have two sources of power [Volvo Group, 2013].

2.3 Databases of logged truck data

In order to drive research and development of Volvo Group Trucks vehicles, GTT have various sorts of data sources available containing logged truck data. One of them contains a very large population of data logged from all Volvo Group vehicles in use, which is today about 2 million vehicles. This data is called Logged Vehicle Data (LVD) and this is also the name of the database. It contains information about usage and performance of the vehicles and is on aggregated form, containing e.g. accumulated distance, time and fuel for truck related parameters. The most usual way this data is downloaded to the database is when the truck comes in for service. The technician working on the truck connects it to a device called TechTool to see what needs to be fixed. While the technician is connected to the truck, TechTool downloads the data logged in the truck and transfers it to the database, which is stored centrally. Each data entry of downloaded information from a truck is called a reading. Each truck can log about 8,000 signals, but usually the readings only contain 200-600 signals since the technician disconnects TechTool when the service is done and does not wait for the reading to finish, which is probably one of the main reasons why so much data had to be removed when arranging the data, see Chapter 6. The vehicles can also upload their LVD to the database via telematics, but these readings are not as extensive as the ones using TechTool. There is a tool called Logged Vehicle Data Analysis Tool (LAT) in which employees with access can extract data from LVD after making a selection of a population of readings from a specified group of vehicles. With this population different plots can be made in the tool depending on which signals were included in the selected readings. Originally these parameters were created to be used in early engineering testing in product development.

Dynafleet is another database which really is a tool for customers to get information from the trucks in their fleets. It is Volvo Trucks' own fleet management system, used by some of their customers, and stores information gathered from the tachograph[3] and the engine management system. In this database data is uploaded more frequently than in LVD, but it contains less signals.

For research purposes, GTT have two databases containing a small number of trucks. European Field Operation Test (EUROFOT) is a pan-european research project involving multiple vehicle manufacturers and research institutes with the goal to test intelligent vehicle systems for developing safer trucks. From this project, data from vehicles of various brands are available, whereof 30 are Volvo trucks. Customer Fuel Follow-Up (CFFU) is a GTT project and has data logged from 15 trucks. The data is time sampled with a sampling rate of 10Hz and contains 500 and 200 signals respectively.

Clearly, all these databases contain big data and require analytic models well developed for the purpose of the research GTT wishes to perform, see Table 2.1. However, LVD was chosen for the data analysis in this thesis since knowledge

[3] A tachograph is a device that automatically records the speed and distance of a vehicle, together with the driver's activity selected from a choice of modes.

about real customer usage of different kinds of trucks and different kinds of customers was sought. Since LVD contains all trucks out in the field, and since the data is logged using electrical architecture already in place in all trucks, it would be extremely powerful if information about customer usage and fuel consumption could be extracted from this data.

There are ongoing initiatives at GTT and Volvo IT[4] which are aiming at industrializing vehicle data retrieval for fuel efficiency among other things. This thesis can be an input to these initiatives who need some research concerning the issues of processing big data in the context of fuel efficiency and taking advantage of customer usage information.

[4]Volvo IT delivers industrial IT solutions, telematics services and consulting services, both to other parts of Volvo Group as partners, and to other customers [Volvo IT, 2014].

Table 2.1: *Descriptions of the different databases available at GTT and what they contain.*

Database	Data	Parameters	Usage at Volvo	In this thesis
LVD	Customer usage data, aggregated. Downloaded when at service. 8,000 signals possible, but usually 200-600 signals.	Accumulated distance, time, fuel for truck related parameters e.g. regarding engine, gear modes etc.	Mostly for examining single vehicles, not larger populations.	Data analysis.
Dynafleet	Truck information sent via telematics updated regularly after a certain time or distance.	From the tachograph and the engine management system.	Fleet management system for customers. GTT (Advanced Technology & Research) use it restrictedly for research.	Future work.
EUROFOT	From trucks of various brands, whereof 30 Volvo Trucks. Time sampled data with sampling rate of 10Hz, contains 500 signals.	Both truck related parameters and videos observing driver behavior.	Research for developing safe trucks.	Future work.
CFFU	From 15 trucks. Time sampled data with sampling rate of 10Hz, contains 200 signals.	Advanced measurements related to fuel consumption.	Research for testing fuel consumption in trucks used in the field, one kind of truck.	Future work.

3

Total Quality Management

This chapter discusses different definitions of Total Quality Management but focuses on one. This definition is built on the cornerstones focus on customers, focus on processes, base decisions on facts, improve continuously, let everybody be committed and committed leadership. Together they constitute the theoretical framework upon which the analysis is based in Chapter 9.

Quality Management has become an all-pervasive management philosophy finding its way into most sectors of today's business society [Sousa and Voss, 2002]. Several studies have tried to synthesize the vast Quality Management literature and the agreement in the literature on what constitutes Quality Management indicates that it as a field has indeed matured and is laid down on solid definitional foundations [Sousa and Voss, 2002].

Total Quality too has generated a great interest in many business sectors, such as manufacturing, service, health care, education and government around the world [Dean and Bowen, 1994]. Total Quality has been defined by Dean and Bowen [1994] as:

> *"A philosophy or an approach to management that can be characterized by its principles, practices, and techniques."*

The three principles of Total Quality is according to Dean and Bowen [1994] customer focus, continuous improvement and teamwork.

Total Quality and Quality Management could be combined into one concept called Total Quality Management, as by Oakland [2003], whose Total Quality Management model brings together a number of components of the quality ap-

proach, including quality circles (teams), problem solving and statistical process control (tools), quality systems such as ISO 9000 (systems) with the processes of the organization at the core of the model. For organizations to be successful or not in their quality approaches, culture, good communication, and most of all commitment from not only senior management but from everyone in the organization are vital [Oakland, 2003].

Bergman and Klefsjö [2010] have also brought Total Quality Management into one concept and defined it as:

> *"A constant endeavor to fulfill, and preferably exceed, customer needs and expectations at the lowest cost, by continuous improvement work, to which all involved are committed, focusing on the processes in the organization."*

Working with Total Quality Management means working with active prevention, change and improvement rather than inspection and repair, since quality work is a continuous process and not a one-time project. Total Quality Management can be seen as a holistic concept where values, methodologies and tools are combined to create increased internal as well as external customer satisfaction at as low resource consumption as possible. The improvement work shall rest on a culture based on the values *focus on customers*, *focus on processes*, *base decisions on facts*, *improve continuously*, *let everybody be committed* and *committed leadership*, which are the cornerstones of Total Quality Management, see Figure 3.1. [Bergman and Klefsjö, 2010]

Total Quality Management therefore fits well as a theoretical model in this thesis since the purpose is to extract value adding information from customer usage data and use it to increase customer satisfaction. In order to achieve this the Total Quality Management theory can be used to highlight how to make the usage of data analysis techniques in analyzing the data to be well integrated with and suited for the organization.

These different definitions are in fact quite similar and capture the same philosophy. For example, the principles of Dean and Bowen [1994], the different components of the Total Quality Management model by Oakland [2003] and the values of Bergman and Klefsjö [2010], are in fact quite the same, only divided into more distinguished parts by Bergman and Klefsjö [2010]. These cornerstones are similar to the different parts of the Volvo Group Quality Policy, see Volvo Group [2012b]. The Volvo Group Quality Policy can be described to contain the parts *focus on customers*, *focus on processes*, *improve continuously* and *let everybody be committed* from Total Quality Management, and therefore the definition of Bergman and Klefsjö [2010] and the six cornerstones are chosen as the framework for this thesis.

Figure 3.1 shows the cornerstones of Total Quality Management and how they interrelate. Together they constitute the quality based theoretical framework of this thesis.

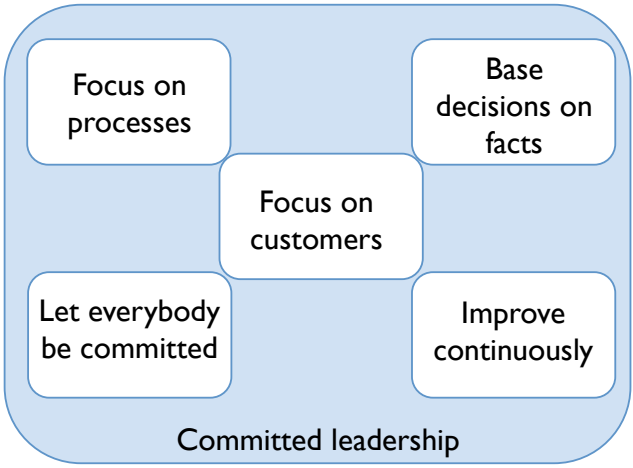

Figure 3.1: *The cornerstones of Total Quality Management [Bergman and Klefsjö, 2010].*

3.1 Focus on customers

There are several definitions of the concept of quality. This thesis follows the definition of Bergman and Klefsjö [2010] for products, which is

> *"The quality of a product is its ability to satisfy, and preferably exceed, the needs and expectations of the customers."*

Consequently, quality is a relative term and often depends on the competition on the market, which means that quality has to be valued by the customers and put in relation to their expectations and needs [Bergman and Klefsjö, 2010].

A definition of the customer is also required, since the customer concept is essential in this definition of quality. From the definition of quality, it can be concluded that the one who decides the quality of the product is the customer, which also is supported by Deming [1986] and Juran [2010]. This can be rephrased into the definition used in this thesis, which is based on Bergman and Klefsjö [2010] and Witell [2007];

> *The customers are defined as those for whom we want to create value.*

According to this definition, an organization has several kinds of customers, for example the product or service may be purchased by one person, used by someone else, and its quality can be decided by a third person, but they are all customers since the product or service will bring them different kinds of value [Deming, 1986, Witell, 2007].

Focusing on customers is the center cornerstone of Total Quality Management and implies finding out what the customers want and need and to systematically

try to fulfill these needs when developing and manufacturing a product or a service so that they will provide a better living for the customer in the future, i.e. create value for the customer [Bergman and Klefsjö, 2010, Deming, 1986]. A customer focused organization should therefore work actively to make the whole organization generate, spread and act based on customer information [Witell, 2007]. Consequently, quality also involves finding out how the customers experience the product as well as all other contacts between the organization and the customer, and to feed this information back into the company process and into the design of the product to bring about improvements [Bergman and Klefsjö, 2010, Deming, 1986]. Deming [1986] stressed the importance of also fulfilling the future needs of the customer, since these needs are constantly changing, and that this cannot be done by asking the customer, but by knowledge, imagination, innovation, risk, trial and error.

3.1.1 External and internal customers

As mentioned above, the quality of a product is valued by the customer, which refers to the external customer outside the organization, since it always is the external customer who judges the quality of an organization's products and therefore, the degree of customer satisfaction is the ultimate measurement of quality [Bergman and Klefsjö, 2010].

It was also mentioned above that an organization has several kinds of customers. External customers are extended to those who live in the environment that is influenced by the organization, its products or production, and society at large [Bergman and Klefsjö, 2010]. According to Juran [2010], since quality is defined by the customers and customers are driven by societal problems, quality now includes safety, no harm to the environment, low cost, ease of use etc. To succeed, all organizations must focus on attaining sustainable organizations [Juran, 2010].

Focusing on the customers does not only involve external customers, but within the company, every employee has an internal customer [Bergman and Klefsjö, 2010]. Ishikawa stated according to Bergman and Klefsjö [2010] that

"The next process is our customer."

This means that the employees of a company constitutes a chain of internal customers and suppliers, each meeting the needs of the next link in the value creating chain [Bergman and Klefsjö, 2010]. Internal and external customers are connected via this customer-supplier chain, since it starts with an external supplier and ends with an external customer outside the organization, with internal customers linking them together and at the same time creating value [Oakland, 2003].

In Total Quality Management, which is focused on external customers, it is important not to forget the internal customers. The needs of the employees must also be satisfied so that they can do a good job and be motivated [Bergman and Klefsjö, 2010]. Internal customer satisfaction and employee motivation are two very connected issues and it can be argued that the key to motivation and qual-

ity is for everyone in the organization to have well-defined customers, since this facilitates fulfilling the needs of the next link in the chain, which also prevents failure to travel all the way to the external customer [Oakland, 2003]. Working with Total Quality Management basically becomes a way to enable employees to do a good job and feel proud of their performance, which creates a foundation for future external customer satisfaction [Bergman and Klefsjö, 2010].

The interdependency between having satisfied employees and attaining high external customer satisfaction is supported in several scientific investigations, for example one from the International Service System with a correlation as high as 0.89. Another study demonstrated a statistical connection as to how employee satisfaction affects external customer satisfaction, and also that employee satisfaction indeed leeds to increased productivity. In a Danish study investigating four hotels and the whole chain from employee satisfaction, through customer satisfaction to financial results, the conclusion was that the higher the degree of employee satisfaction, the higher external customer satisfaction, which in turn leads to higher gains. [Bergman and Klefsjö, 2010]

Focus on customers is highly relevant for this thesis since increasing mainly external but also internal customer satisfaction is the whole reason for doing an in-depth analysis of the data.

3.2 Focus on processes

Everything we do is a process according to Oakland [2003], whose definition is as follows

> *"A process is the transformation of a set of inputs into outputs that satisfy customer needs and expectations, in the form of products, information or services."*

According to [Bergman and Klefsjö, 2010], most organized activities can be regarded as a process, which is defined as

> *"... a network of interrelated activities that are repeated in time, whose objective is to create value to external or internal customers".*

Moreover, the process transforms certain inputs, such as information, materials and knowledge, into certain outputs in the form of numerous kinds of goods and services, which are transferred to somewhere or to someone – the customer, see Figure 3.2 [Bergman and Klefsjö, 2010, Oakland, 2003]. The purpose of the process is to produce an output that satisfies its customers while using as little resources as possible [Bergman and Klefsjö, 2010]. An organization consisting of people and their relationships, resources and tools, supports the process [Bergman and Klefsjö, 2010]. In order to produce an output that meets the customer requirements, it is necessary to define, monitor and control the inputs of the process, which in turn may be supplied as output from an earlier process [Oakland, 2003]. To minimize resources and to satisfy customers it is important

to identify the suppliers of the process and to provide clear signals about what is needed in the process [Bergman and Klefsjö, 2010]. There resides a transformation process at every supplier-customer interface, and every single task throughout an organization must be viewed as a process in this way [Oakland, 2003].

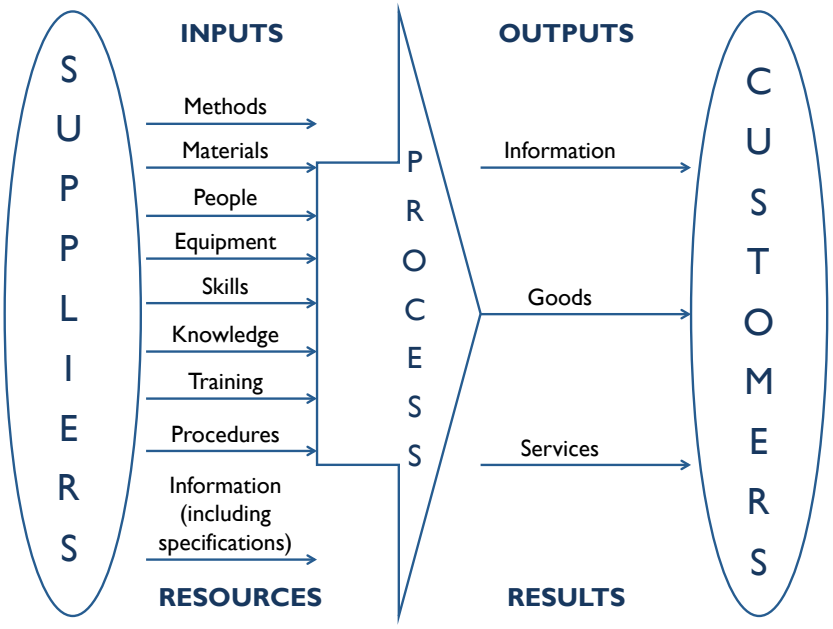

Figure 3.2: *A process transforms certain inputs from suppliers into certain outputs to customers with the purpose of satisfying the needs of the customers with as little resource consumption as possible [Bergman and Klefsjö, 2010, Oakland, 2003].*

Each process can be analyzed by examining its inputs and outputs, which will determine some of the actions necessary to improve quality [Oakland, 2003]. The process generates data that indicate how well it satisfies the needs of the customers [Bergman and Klefsjö, 2010]. With statistical tools and models, it is possible to draw conclusions from the process history about its future results, and to recover the necessary information to improve the process [Bergman and Klefsjö, 2010].

Once it is established that the process is capable of meeting the requirements of the customer, it must be ensured that the process continues to do so, which brings a requirement to monitor the process and the controls on it [Oakland, 2003]. By shifting the view of the process, the need to ask the "inspection question" has moved to focus attention on the inputs of the process in order to make sure they are capable of meeting the requirements, and have replaced a strategy of detection with one of prevention [Oakland, 2003]. According to Bergman and Klefsjö [2010], the *process view* means not only looking at every single piece of data, such

as a measurement result or a customer complaint, as a unique phenomenon, but instead regard it as a part of the statistics that can provide information about how well the process is working and how it can be improved. Therefore, it is essential to look at data over time [Bergman and Klefsjö, 2010].

Furthermore, in order to predict the output of the processes, they need to be stable and repeatable by being standardized. Standardized tasks and processes are the foundation for continuous improvement and employee empowerment since by standardizing today's best practices, they capture the learning up to this point and this way, standards provide a launching point for true and lasting innovation. [Liker and Meier, 2006]

Focus on processes is an important part of the theory for this thesis since it is important to look at the data analysis with the process view in order to identify patterns instead of looking at individual observations as isolated events. It is also important to identify that in order to control the process of analyzing the data and its outputs, one must also pay attention to its inputs, i.e. the used data.

3.3 Base decisions on facts

To base decisions on facts which are well-founded and to not let random factors have decisive importance is an important cornerstone of Total Quality Management and an important element in modern quality philosophy [Bergman and Klefsjö, 2010]. Numbers and information should always form the basis for understanding, decisions and actions in order to constantly improve the ways processes are operated [Oakland, 2003]. This requires knowledge about variation and the ability to distinguish between "natural variation" and variation due to identifiable causes [Bergman and Klefsjö, 2010]. Factual data of both numerical and verbal character is needed and an organization must gather, structure, analyze and decide upon different kinds of information [Bergman and Klefsjö, 2010]. To be able to focus on customers, systematic information about the needs, requirements, reactions and opinions of the customers is required [Bergman and Klefsjö, 2010]. In order to satisfy their customers, the organization must of course understand not only the needs of the customers, but also the ability of its own organization to meet them [Oakland, 2003]. It is also important to have sufficient knowledge about the product before releasing it on the market [Bergman and Klefsjö, 2010] and after releasing it, test it in service and find out what the customers think about it [Deming, 1986].

To have a strategy for making decisions based on facts in relation to manufacturing is also important [Bergman and Klefsjö, 2010]. Earlier, it was common to collect many facts and take a lot of measures, store them in files, tapes or discs without ever using them to make simple statistical analyses and draw conclusions about the manufacturing process, which could have been an excellent basis for variation reduction within the production process, and thus for improving quality [Bergman and Klefsjö, 2010]. However, in recent years, a global trend towards more data, more computer-operated analysis of it, with decision making more ori-

ented to being based on facts have been observed [Davenport et al., 2010]. Still, research suggests that 40 percent of major decisions are based on the manager's intuition instead of facts [Accenture, 2008].

Basing decisions on facts requires actively searching for relevant information, which then needs to be compiled in order to be analyzed. From this analysis, conclusions are drawn, which are used for making improvements. [Bergman and Klefsjö, 2010]

Oakland [2003] mentioned that a system for data-gathering, recording and presentation is essential, which should include to record data, use data, analyze data and act on the results. He stated that all processes should be measured and all measurements should be recorded; if data is recorded and not used it will be abused; data analysis should be carried out by means of some basic systematic tools; and that recording and analysis of data without action leads to frustration [Oakland, 2003].

Base decisions on facts is an essential cornerstone for this thesis and also an important motivation to why this data analysis should be performed. Information about the customer must be the ground for making decisions about how to develop products to satisfy the customer, and therefore customer usage data is excellent to use for in-depth analysis for this purpose.

3.4 Improve continuously

The market changes constantly with technological advances and so does the demands of the external customers. Therefore, continuous quality improvements of products and services produced by the company are essential [Bergman and Klefsjö, 2010]. To improve continuously is an important component in a successful quality strategy, since anyone who stops improving in this evolving environment soon stops being good [Bergman and Klefsjö, 2010]. If the company succeeds to improve continuously, improved quality means delivery of those features of the product or service that respond better to customer needs, and will therefore have an effect on the revenues of the company [Juran, 2010]. One strategy of organizations who have been successful in their quality initiatives is to engage in continuous innovation and process improvement since it has been recognized that quality is a moving target and therefore there is no end to improving processes [Juran, 2010].

There are also internal drivers for a company to improve continuously, and that is the reduction of costs. Improved quality means fewer errors, fewer defects and fewer field failures and therefore a lower "cost of poor quality" [Juran, 2010]. One of Deming's [1986] 14 points of management clearly state that a company should

> *"Improve constantly and forever the system of production and service, to improve quality and productivity, and thus constantly decrease costs".*

Deming [1986] studied the management of companies in Japan and observed in 1948 and 1949 that improvement in quality generates naturally and inevitably improvements in productivity as variations are reduced, as seen in Figure 3.3.

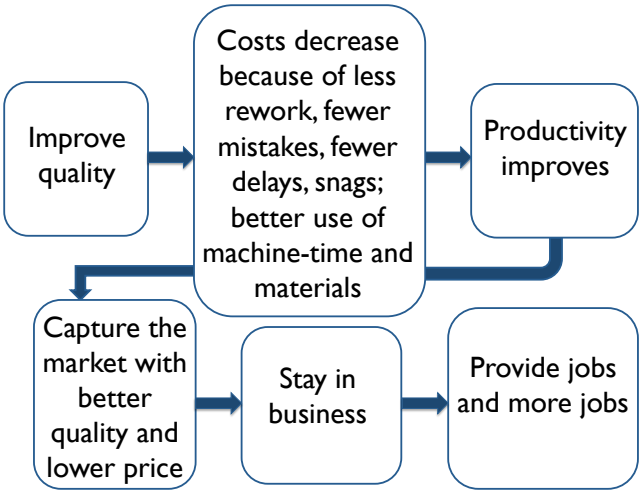

Figure 3.3: *The chain reaction from improved quality [Deming, 1986].*

According to Juran [2010] the pursuit of high quality transforms a culture and these transformational changes are a result of an organization's relentless pursuit to be the best in quality and implementing a systematic method to get there. Designing and continuously improving the quality of an organization's goods and services creates high stakeholder and employee satisfaction, which enables the organization to sustain the pursuit for the long term [Juran, 2010].

The basic rule of continuous improvement says that it is always possible to improve products, processes and methodologies while using fewer resources, i.e. to achieve higher quality at lower costs, which is similar to the quote

"There is a better way. Find it."

by the American inventor, scientist and businessman Thomas Edison (1847–1931). This brings a mental picture of that everything can be done better than how it is done today; better in the sense of providing better customer benefit and better in the sense of using less resources to do so, which results in a win-win effect for customers, employees and the company. [Bergman and Klefsjö, 2010]

3.4.1 Improve continuously and focus on processes

Continuous improvement follows immediately after having achieved stable processes, since by standardizing todays's best practices, they capture the learning up to this point. The task of continuous improvement is then to improve upon this standard by continuously using tools for determining the root cause of inefficiencies or slowness and making them visible. By doing this, opportunities to

continually learn are revealed. [Liker and Meier, 2006]

The Japanese term *kaizen* means improvement for the better and is a customer focused strategy for continuous improvement that includes everyone in the organization, which has generated a focus on processes and acknowledged people's process-oriented contributions for improvements rather than only focusing on the results of people's performance [Imai, 1986]. This can very well include both big and small improvements, since improvements can be made by advancement leaps as well as slower progression [Bergman and Klefsjö, 2010]. Both innovation and kaizen is needed for a company to survive and grow [Imai, 1986].

3.4.2 Improve continuously and focus on customers

As mentioned above, focus on customers is a center cornerstone of Total Quality Management and is closely connected to continuous improvement. If the customer is satisfied with a product, the customer is likely to continue to consume this product and therefore the needs and expectations of the customers are an excellent basis for improvement work in trying to get additional and more satisfied customers [Witell, 2007]. The idea is to, by combining customer focus and continuous improvement, get a holistic view and to use the needs of the customers as a starting point to systematically work to create continuous and systematic improvements in the processes and results of a company [Witell, 2007].

Therefore, improve continuously is also relevant in this thesis. There are always ways to improve and the data analysis of customer usage data can be used over time to achieve customer focused ameliorations.

3.5 Let everybody be committed

For the quality work to be successful, it is essential to create conditions for participation in the work with continuous improvement [Bergman and Klefsjö, 2010]. There are three important components that are key to facilitate for employees to be committed and to participate, which are *communication*, *delegation* and *training*. Communication is important since information is needed for a person to take responsibility and to understand the importance of his or her task for the goals of the whole organization [Bergman and Klefsjö, 2010]. Training is also important for a person to take responsibility, and the employee must have a chance to feel commitment, professional and personal pride, and responsibility, to be able to do a good job [Bergman and Klefsjö, 2010]. Also Ishikawa [1985] shares this view and stated that once a subordinate is educated on a one-to-one basis through actual work with the superior, authority can be delegated to the subordinate together with the freedom to do his or her job. In this way the subordinate will grow [Ishikawa, 1985].

In order for the data analysis of customer usage data to actually be used to increase customer satisfaction, the organization needs to let everybody be committed and train employees in using the right tools, and therefore this is relevant to

this thesis.

3.6 Committed leadership

If the company's work with Total Quality Management is to be successful it must be built on the top management's continuous and consistent commitment to quality issues and it cannot be emphasized too much how important strong and committed leadership is to create a culture for successful and sustainable quality improvements [Bergman and Klefsjö, 2010].

According to Bergman and Klefsjö [2010], one of the greatest experts in the quality field Joseph M. Juran stated

> *"To my knowledge, no company has attained world-class quality without upper managment leadership."*

The top management have to include quality aspects in the company vision and take actions to support this vision, but also actively take part in the improvement process to show the employees that quality is as important as they say [Bergman and Klefsjö, 2010].

Many research studies show how important it is to create commitment and engagement from the members of the staff and that managers at all levels of the company are credible, clear and good at communicating and work well as good examples [Bergman and Klefsjö, 2010]. The middle management have a particularly important role to play, since they must not only understand the principles of Total Quality Management, but also explain them to the people they are responsible for, and ensure that their own commitment is communicated, and only then will these principles spread throughout the organization [Oakland, 2003].

One of Deming's [1986] 14 points of management is about adopting and instituting leadership and Deming stated that the job of management is not supervision, but leadership. Deming [1986] also stressed that the transformation of Western style of management requires that managers be leaders. This goes in line with what Imai [1986] called *process focused leadership*, which is a leadership style focused on people, in contrast to a leadership style only focused on results. In process focused leadership, a manager must support and stimulate contributions to improve the way in which employees execute their tasks [Imai, 1986].

Committed leadership is the whole basis for the other cornerstones to work at all in the organization. It is also important for the customer usage data, the database and the analysis of it to be identified as significant for the organization's work with customer satisfaction and therefore highly relevant for this thesis.

4

Big data and analytics

In this chapter, the research and theory on big data and analytics is described as well as its benefits and challenges based on a 5Vs model: Value, Volume, Variety, Velocity and Veracity, and also ethical aspects of privacy. How big data can be handled in an organization is also discussed.

4.1 Big data

Data has increased in a large scale in various fields over the last 20 years [Chen et al., 2014]. According to a report from the International Data Corporation, the amount of data created and replicated surpassed 1.8 zettabytes[1] in 2011, a number that will grow by a factor of nine in just five years [Gantz and Reinsel, 2011]. It has been estimated that the amount of data generated worldwide in 2013 was going to reach four zettabytes [VSAT, 2013], however there are others who suggest that the total data stored on the Internet was over one yottabyte[2] in 2013 [Facts Hunt, 2014].

Big data refers to large amounts of data. However, huge *volume* is not its only feature. This section will explain the potential and benefits as well as the challenges of big data by describing *volume* as well as the other so called Vs of big data; *variety*, *velocity*, *veracity* and *value* [Chen et al., 2014, Jagadish et al., 2014].

Today, many agree on the importance and the potential of big data. However, there is a difference of opinion of its definition.

[1] SI (metric) definition: 1 zettabyte = 1 trillion gigabytes = $(10^{12})10^9$ bytes = 10^{21} bytes
[2] SI (metric) definition: 1 yottabyte = 1 trillion terabytes = $(10^{12})10^{12}$ bytes = 10^{24} bytes

4.1.1 Definition of big data

A general definition have been stated by Chen et al. [2014];

> *"In general, big data refers to the datasets that could not be percieved, acquired, managed, and processed by traditional IT and software/hardware tools within a tolerable time."*

However, scientific and technological enterprises, research scholars, data analysts, and technical practitioners have different definitions of big data because of different concerns, which will be shown in the following definitions.

Apache Hadoop develops open-source software for the distributed processing of large datasets and defined in 2010 big data as

> *"... datasets which could not be captured, managed, and processed by general computers within an acceptable scope."* [Chen et al., 2014]

In 2011, a report from McKinsey Global Institute stated that big data is part of every sector and function of the global economy and that it is

> *"The next frontier for innovation, competition and productivity."* [Manyika et al., 2011]

According to Manyika et al. [2011], big data refers to datasets whose size is beyond the ability of classic database software tools to capture, store, manage and analyze. This definition implies that the size of datasets that qualify as big data will change over time as technology advances. Also, the definition of big data can differ between sectors and applications since different tools and different kinds of datasets are available.

As early as in 2001, Doug Laney defined a 3Vs model of data management: Data Volume, Velocity and Variety and the opportunities and challenges that comes with it [Laney, 2001]. While these three are important, they fail to include other features of big data such as Veracity [Jagadish et al., 2014]. Therefore, this report extends the 3Vs model into 5 Vs in order to describe both the benefits and challenges of big data: *Volume*, *Variety*, *Velocity*, *Veracity* and *Value*, see Figure 4.1.

4.1.2 Potential and benefits of big data

Value

As mentioned in Section 1.1, there are many examples of the value and potential of big data if used creatively and efficiently for various industries. Since this thesis is a case study at GTT, the value of big data is here focused on the value for global manufacturing companies.

In the extensive report from McKinsey Global Institute in 2011, the transformative potential of big data is discussed for five different domains, of which global manufacturing is one [Manyika et al., 2011]. This study examines multiple subsectors, covering both discrete and process manufacturing, from basic sub-sector

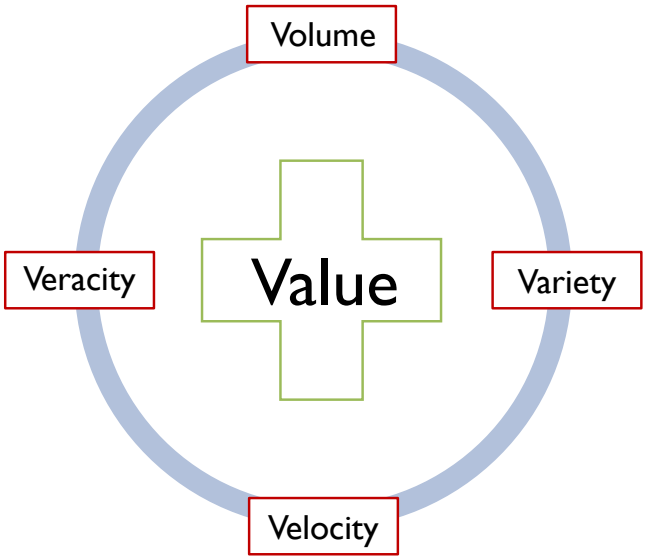

Figure 4.1: *The 5 Vs of big data, where Value represents the benefits and Volume, Variety, Velocity and Veracity represent the challenges of big data.*

manufacturing such as consumer goods and food, to advanced sub-sector manufacturing such as automotive and aerospace. The analysis focuses primarily on the core activities of a manufacturer, i.e. research and development, supply chain and manufacturing processes, and less on adjacent processes such as marketing and sales. Since the purpose of this thesis is to use customer usage data in order to get knowledge about the customer and increase their satisfaction, more focus is put on the activities closer to the customer.

This report states that although manufacturing historically has been a productivity leader, big data can still help extend gains. Manufacturers have tremendous potential to generate value from the use of large datasets by leveraging big data across the value chain. These gains will come from improved efficiency in design and production, further improvements in product quality and better meeting customer needs through more precisely targeted products and effective promotion and distribution. [Manyika et al., 2011]

Research and development

Big data can be used in research and development by enabling and improving product life cycle management and open innovation. For example, the product development time can be reduced significantly, as for Toyota, Fiat and Nissan who have all cut new-model development time by 30 to 50 percent. For many years, manufacturers have implemented IT systems to manage the product life cycle including computer aided-design, engineering, manufacturing, and product development management tools, and digital manufacturing. However, the large

datasets generated by these systems have tended to remain trapped within the different systems. By instituting product life cycle management platforms that can integrate datasets from several systems to enable effective and consistent collaboration, manufacturers could capture a significant big data opportunity to create more value. This kind of platform could enable so called "co-creation", which is to bring together internal and external inputs to create products and could be especially useful for fields where a new product is assembled from many different components from many different suppliers around the world. This kind of platform can enable co-creation of designs between OEM s and suppliers, and can also enable extensive experimentation at the design stage, which is especially useful since decisions made in the design stage typically drive 80 percent of manufacturing costs. [Manyika et al., 2011]

Manufacturers are increasingly relying on outside inputs through innovative channels to drive innovation and develop products that address emerging customer needs. Some manufacturers are inviting external stakeholders such as customers and different external experts, including academic and industry researchers, to submit ideas for innovations or even collaborate on product development via web-based platforms. These open innovations have shown to be very successful, but one key problem is how to extract the valuable ideas from the potentially large number of inputs these platforms can give. By applying big data techniques such as automated algorithms, this issue can be resolved. For example BMW has developed an "idea management program" to help evaluate ideas which has cut the time taken to identify high-potential ideas by 50 percent. [Manyika et al., 2011]

Product design

Design to value means to systematically design new products according to information about customer needs extracted from different customer data sources. Market research has for a long time provided customer data as input into the production process, but many manufacturers have yet to extract important insights from increasing volume of customer data to enhance existing designs and help develop specifications for new models and variants. World leading manufacturers conduct conjoint analyses[3] to uncover how much customers are willing to pay for certain features and to understand which features are most important for success in the market. These efforts are supplemented by additional quantitative customer insights mined from sources such as point-of-sales data and customer feedback and some companies use newer sources of data such as customer comments in social media and sensor data that describe actual customer use. [Manyika et al., 2011]

Supply chain

Big data can also be used to handle the volatility of demand which is a critical issue for many manufacturers. By the improved use of their own data, manufactur-

[3]Conjoint analysis is a statistical technique that involves providing a controlled set of potential products and services to elicit end users' preferences through which an implicit valuation of individual elements that make up the product or service can be determined [Manyika et al., 2011].

ers can improve their demand forecasting and supply planning. Real-time data can help manage demand planning across extended enterprises and global supply chains, while reducing defects and rework within production plants. [Manyika et al., 2011]

Production

In production there are also big-data levers. Case studies in automotive, aerospace and defense, and semiconductors show that advanced simulations taking inputs from product development and historical production data can reduce the number of production-drawing changes as well as the cost of tool design and construction. This enables realization of substantial reductions in assembly hours, cost savings and even improved delivery reliability. [Manyika et al., 2011]

The growth of Internet of Things[4] applications allows manufacturers to optimize operations by embedding real-time, highly granular data from networked sensors in the supply chain and production processes. These data enable process control and optimization to reduce waste and maximize yield, and even allow for innovations in manufacturing that have not been possible before, such as nano manufacturing. [Manyika et al., 2011]

Marketing and sales

There are also many possibilities to use large datasets in the marketing, sales and aftermarket service activities. For example, using sensor data from products once they are in use is an increasingly important application for manufacturers to improve service offerings. [Manyika et al., 2011]

One of the top trends in commercial vehicles telematics for 2014 is that big data analytics and business intelligence will strengthen OEM and aftermarket telematics vendors offerings in predictive maintenance and diagnostics reducing the average repair time by 25-50 percent and reduction in warranty costs by 2-3 percent resulting in higher customer satisfaction. [Tare et al., 2014]

Also truck manufacturers use large data sets to improve their products. For example Man Trucks use data collected from the driving cabs of the trucks they sell to gain information about how those vehicles can be driven more safely and efficiently, which they call "Trucknology" and has provided Man Trucks with a competitive customer proposition [Bartram, 2013].

To conclude, the usage of big data has a huge potential which is real and significant. However, in order to realize this into actual value, there are a number of technical as well as organizational challenges to deal with.

4.1.3 Challenges of big data

The other Vs of big data chosen to be treated in this report are connected to technical features which make big data challenging to handle.

[4] "Internet of Things" refers to sensors and actuators within networks of physical objects [Manyika et al., 2011].

Volume

The first thing that comes to mind when thinking about big data is of course its size. In the past, the problem of increasing data volume has been mitigated by processors getting faster according to Moore's Law[5][Jagadish et al., 2014]. Now however, there is a fundamental shift in motion as a result of that data volume is increasing faster than CPU speeds and other compute resources [Jagadish et al., 2014]. Pervasive sensors and computing are generating data at unprecedented rates and scales compared with the relatively slow advances of storage systems and the data scale becomes increasingly huge with the generation and collection of massive data [Chen et al., 2014]. One of the most pressing challenges of big data is that the current storage system could not support such massive data and therefore, an important principle related to the analytical value should be developed to decide which data shall be stored and which data shall be discarded [Chen et al., 2014]. However, as enterprises come to see information as a tangible asset, they become reluctant to discard it [Laney, 2001]. Typically, increases in data volume are handled by purchasing additional online storage, but as data volume increases, the relative value of each data point decreases proportionately, this resulting in a poor financial justification for slightly increasing online storage [Laney, 2001].

The move toward cloud computing, which is now joining different tasks with varying performance goals into very large clusters, is another dramatic shift in motion [Jagadish et al., 2014]. Cloud computing is utilized to meet the requirements on infrastructure for big data, e.g. cost-efficiency, elasticity, and smooth upgrading/downgrading [Chen et al., 2014]. However, this level of sharing resources on expensive and large clusters stresses grid and cluster computing techniques from the past, and requires new ways of determining how to run and execute data processing jobs so that the goals of each task can be met cost-effectively while also dealing with system failures [Jagadish et al., 2014].

Generally, there is a high level of redundancy in datasets. To reduce the indirect cost of the entire system on the premise that the potential values of the data are not affected, redundancy reduction and data compression is effective. For example, most data generated by sensor networks are highly redundant, which may be filtered and compressed at orders of magnitude. [Chen et al., 2014]

The volume of the data used in this thesis is managed when arranging the data and also when differentiating the usage of the trucks, see Chapter 6 and Chapter 7.

Variety

Effective data management has a problem with the variety of incompatible data formats, non-aligned data structures and inconsistent data semantics [Laney, 2001]. Variety refers to heterogeneity of data types, which include semi-structured and

[5]Moore's Law stated that the number of transistors on an affordable CPU will double in every two years. This means that processor speeds, or overall processing power for computers will double every two years [Moore's Law].

unstructured data such as audio, video, webpage, and text, as well as traditional structured data [Chen et al., 2014]. It also refers to the heterogeneity of data representation and semantic interpretation [Jagadish et al., 2014]. To summarize, datasets defined as big data have certain levels of heterogeneity in type, structure, semantics, organization, granularity and accessibility [Chen et al., 2014].

Since machine analysis algorithms expect homogeneous data and are poor at understanding nuances, data needs to be carefully structured [Jagadish et al., 2014]. Data representation aims to do this so that the data can be meaningful for computer analysis and user interpretation [Chen et al., 2014]. However, an improper data representation will reduce the value of the original data and may even obstruct effective data analysis, thus in order to avoid this, efficient data representation shall reflect data structure, class and type, as well as integrated technologies [Chen et al., 2014].

Not only the primary data needs to be structured, but there is also a challenge to automatically generate the right metadata to describe the data logged, since for example details concerning specific conditions may be required in order to interpret the primary data correctly. To be able to do this automatically it is necessary to minimize the human burden, but logging information about the data at its birth, which is called data provenance, is not useful unless this information can be interpreted and carried along through the data analysis. For example, a processing error at one step can render subsequent analysis useless, but with suitable provenance, all subsequent processing that depends on this step can easily be identified. [Jagadish et al., 2014]

The variety of the data used in this thesis is managed when arranging the data and also when differentiating the usage of the trucks, see Chapter 6 and Chapter 7.

Velocity

Velocity refers to the rate at which data arrive as well as the time frame in which they must be acted upon [Jagadish et al., 2014]. The timeliness of data collection and analysis etc. must be rapidly and timely conducted, in order to maximumly utilize the commercial value of big data [Chen et al., 2014].

As data volume grows, real-time techniques are needed to summarize and filter what is to be stored, since in many situations it is not economically viable to store the raw data. Another usual need is to find elements in a very large dataset that meet a specified criterion. A way to speed this process up, as opposed to searching the whole dataset, is to use index structures which are created in advance to find qualifying elements quickly and can be used in for example traffic management systems to be able to suggest route alternatives. [Jagadish et al., 2014]

Since the data used in this thesis is on aggregated form and not downloaded frequently, this issue is not handled in the data analysis of this thesis.

Veracity

Veracity denotes the unreliability, inconsistency and incompleteness of big data. The various sources increasingly providing information are of varying reliability. Uncertainty, errors, and missing values are usual and must be handled, but even after error correction has been applied, some incompleteness and some errors in data are likely to remain. This incompleteness and these errors must be correctly managed during data analysis, which is a challenge. One way to make progress is suggested by recent work on managing and querying probabilistic and conflicting data. [Jagadish et al., 2014]

Veracity of the data was handled in this thesis especially when arranging the data before starting the data analysis itself. Incomplete data was removed from the population and also outliers. A further description can be seen in Chapter 6.

Ethical aspects

There are also other increasing concerns connected to big data, which includes ethical aspects of privacy and data ownership. There is a great public fear regarding the inappropriate use of personal data, particularly through linking of data from multiple sources. To realize the potentials of big data, managing privacy effectively, which is both a technical and sociological problem, must be addressed jointly from both perspectives. [Jagadish et al., 2014]

This is also a concern when analyzing customer usage data collected from trucks, since the privacy of the driver in the judgement of how the driver handles the vehicle can be harmed. However, when using aggregated data such as in LVD used in this thesis, it is impossible to see patterns of individual drivers and their specific behavior. The objective is on the contrary to find patterns in large populations to draw conclusions about groups of trucks used in similar ways. Therefore, this issue is not considered to be especially addressed in the data analysis performed in this thesis, but still needs to be emphasized to be a concern connected to big data.

4.1.4 Organizational big data handling

The technical challenges of big data mentioned above, as well as organizational challenges often make companies pursue costly or ineffective solutions or make them paralyzed into action. To handle these challenges in order to exploit the data, companies need to create a plan for how data, analytics, frontline tools, and people come together to create business value, see Figure 4.2. [Biesdorf et al., 2013]

A big-data plan should address similar issues to those of a strategic plan: a company needs to choose the internal and external data they will integrate; select, from a long list of potential analytic models and tools, the ones that will best support their business goals; and build the organizational capabilities needed to exploit this potential [Biesdorf et al., 2013].

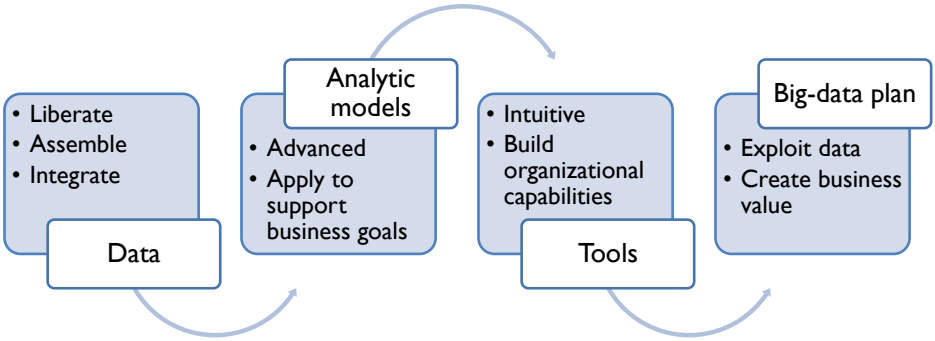

Figure 4.2: *Data, analytic models and tools are the three parts of a big-data plan.*

Data

A strategy for assembling and integrating data is essential [Biesdorf et al., 2013]. Critical data may be residing in legacy IT systems, siloed horizontally across business units or vertically by function, or outside the company in unstructured forms such as social-network conversations [Biesdorf et al., 2013]. To fully capitalize the data, a company needs to have a free interchange of data among different functions and business units such as marketing and sales, research and development, and production [Manyika et al., 2011]. Many of the levers of big data also require access to data from different players in the value chain, for example data from suppliers are necessary to optimize production planning [Manyika et al., 2011]. A big-data plan may highlight a need for the massive reorganization of data architectures over time, among other things implementing data-governance standards that systematically maintain accuracy [Biesdorf et al., 2013].

Analytic models

However, integrating data alone does not generate value, but rather advanced analytic models are needed to enable data-driven optimization or predictions. A plan must identify where models will create additional business value, who will need to use them, and how to avoid inconsistencies and unnecessary proliferation as models are scaled up across the organization. Nevertheless, it is important not to include too many variables since this will create complexity while making the models harder to apply and maintain. [Biesdorf et al., 2013].

Tools and organizational capabilities

The output of the models may contain a lot of information, but it will only be valuable if managers and frontline employees understand and use them. To obtain this, intuitive tools that integrate data into day-to-day processes and translate modeling outputs into tangible business actions are needed. It is common that companies fail to complete this step in their planning, resulting in managers and operational employees not using the new models, whose effectiveness obviously

falls. [Biesdorf et al., 2013]

On the other hand, even a big-data plan can disappoint when organizations lack the capabilities and the right people [Biesdorf et al., 2013]. Manufacturers must tackle organizational, cultural, and talent challenges to maximize the benefits of big data [Manyika et al., 2011].

First, companies need to invest in IT, since the rising volume of data from new sources requires a new level of storage and computing power. Also investments to develop interfaces and protocols to share data effectively across extended enterprises are needed. These investments will be costly, but the long-term payoff should outweigh the costs. [Manyika et al., 2011]

Big-data planning is at least as much a management challenge as a technical one, and there is hard work of getting business players and data scientists to work together to find the best solutions [Biesdorf et al., 2013]. Therefore, achieving success will require strong leadership and a cultural shift to establish the mind-sets and behaviors to breech today's silos [Manyika et al., 2011]. A strong leadership is also needed to engage the organization in using the tools [Biesdorf et al., 2013].

Also, there is a shortage of talent with the right experience and deep analytical expertise for handling this level of complexity [Manyika et al., 2011]. However, companies need not only to recruit new talent, but also to remove organizational obstacles that today prevent such individuals from making maximum contributions [Manyika et al., 2011]. A good rule of thumb for planning purposes is a 50-50 ratio of data and analytics investments of data and modeling to training [Biesdorf et al., 2013].

4.2 Analytics

In recent years a global trend towards more data, more computer-operated analysis of it, with decision making more oriented to being based on facts have been observed [Davenport et al., 2010]. However, research suggests that 40 percent of major decisions are based on the manager's intuition instead of facts [Accenture, 2008].

Organizations can use big data as a real source of competitive advantage and the benefit of combining the progress made in big data and advanced analytics is no longer in doubt [Biesdorf et al., 2013]. Applying this, an organization is able to know what is really working. With an extensive data analysis it is possible to establish whether implemented changes really are causing desired effects in the business, or whether they are simply caused by random statistical variations [Davenport et al., 2010]. Cutting costs and improving efficiency is also possible with analytics, which can be used for optimization of asset requirements [Davenport et al., 2010].

To be able to detect patterns in the vast amount of customer and market data coming up is very powerful, and can be used to anticipate changes in market con-

ditions. Predictive models can anticipate market changes and enable companies to act quickly to avoid costs and eliminate waste. [Davenport et al., 2010]

By using analytics it is also possible to leverage previous investments in IT and information to get more insight, faster execution, and more business value in many business processes [Davenport et al., 2010]. Finally, this builds a basis for improving decisions over time. If clear logic and explicit supporting data is used to make a decision, it is possible to examine the decision process more easily and try to improve it [Davenport et al., 2010]. This kind of data-driven business brings about greater transparency into how operations actually work, better predictions, and faster testing [Biesdorf et al., 2013].

5
Research methodology

This chapter contains a discussion on the method of case study as well as a summary of the data analysis made in this thesis, which is further described in Chapters 6, 7 and 8.

5.1 Setting

The setting and background at GTT for this thesis is described in Chapter 2. In Section 2.3 the databases containing logged data from the trucks are presented. Several different databases were available, but the LVD database was to be used in particular. The initial problem the thesis authors were presented with and which GTT were facing was that this database was not used to the desired extent. This group at GTT, working with fuel consumption, did not work with any of the databases available described in Section 2.3 but knew that LVD first and foremost had large potential. The idea of combining the databases in some way would have even more potential since they contain different kinds of data. This scope was however considered as too large and therefore the thesis was initiated with the intention of exploiting the customer usage data in LVD in some way to gain knowledge connected to fuel consumption, which by extension was meant to be used in product development in order to increase customer satisfaction.

5.2 Pre-study

Since the thesis did not have a pre-set outline and the thesis initiators at GTT did not want to bias the authors in the findings of their solution to the problem, a pre-study was necessary in order to establish what could be done, how it could

be done and to further explore why it should be done. The pre-study set out from the fact that GTT had this database of customer usage data available and was interested in finding patterns in this data and investigating how different for GTT known parameters affect fuel consumption for different kinds of usage patterns. A review of previous research, which is briefly described in Section 1.3, was made in order to answer the questions of what and how, and also touched the question of why in reviewing research about fuel consumption and climate changes. This motivation for the thesis is further described in Chapter 1.

Interviews with different key persons at GTT were also conducted in order to learn more about the data and the databases and about how the engineers work with these. As a result of this pre-study, the purpose, research objectives, ideas of theoretical framework and what data analysis techniques to use for the thesis were formed. The theoretical framework and the data analysis techniques were chosen based on the purpose of the thesis, on the accessible data which were derived from real customer usage of the vehicles, and on previous research showing that it had been done before but in other contexts and purposes.

5.3 Case study

The method of this thesis is a case study of GTT, which can be seen as an example of why and how a company should make use of their databases with big data of customer usage by using data analysis techniques as analytical models. The specific data analysis made in this thesis can in turn be seen as a deeper part of the case study and an example of an analysis that an organization who wants to use customer usage data to analyze what usage factors effect a certain performance of their product can implement. In Section 5.3.1 a short presentation of the data analysis is given and further details are described in Chapters 6, 7 and 8.

As a result from this case study, conclusions are drawn about how and which data analysis techniques can be used to extract value adding information from large amounts of aggregated customer usage data. Conclusions are also drawn about how a company needs to handle the data in order to be able to use this value adding information to increase customer satisfaction of their products and services.

The method of performing a single-case study was chosen because of the depth of this study, which not only studied the managerial challenges of big data, but also the technical challenges and explored how a big data analysis could be performed at a truck manufacturer. Naturally, for a given set of available resources, the fewer the case studies, the greater the opportunity for depth of observation [Voss et al., 2002]. Since this study contained these both dimensions, having a both holistic and deep approach, a single-case study was chosen considering the limited time available.

Deep understanding of the actors, interactions, sentiments, and behaviors occurring for a specific process through time should be seen as the principal objective

by the case study researcher [Woodside, 2010]. In contrast to the survey approach, a case study can study things in detail. The reason for concentrating efforts on one case rather than many is that there may be insights to be gained from looking at the individual case that can have wider implications and, importantly, that would not have come to light through the use of a research strategy that tried to cover a large number of instances, such as a survey approach [Denscombe, 2007].

Case research can provide results that can have very high impact. Free from the rigid limits of questionnaires and models, the results of case research can lead to new and creative insights and have high validity with practitioners, which are the ultimate users of research [Voss et al., 2002]. Using triangulation with several methods of data collection, the validity can be increased further [Voss et al., 2002]. Triangulation can be used to achieve deep understanding in case study research and often includes: (1) direct observation by the researcher within the environments of the case, (2) probing by asking case participants for explanations and interpretations of "operational data" and (3) analyses of written documents and natural sites occurring in case environments [Woodside, 2010]. One of the strengths of the case study approach is that it invites the researcher to use a variety of sources, a variety of types of data and a variety of research methods as part of the investigation [Denscombe, 2007].

Triangulation has been pursued in this study by conducting interviews with different key persons having information and experience of the use of the different databases, by studying internal documentation about the company and their databases, and by direct observation where the researchers have worked with the data and the tools available in order to perform the data analysis.

Another benefit of the case study approach is that it not only allows for in-depth study, but also allows for the researcher to deal with the case as a whole, in its entirety, and thus have some chance of being able to discover how the many parts affect one another and in this respect, case studies tend to be holistic rather than deal with isolated factors [Denscombe, 2007]. It has also been recognized that case study research is particularly good for examining the how and the why questions [Voss et al., 2002], which is the case in this thesis.

5.3.1 Data analysis outline

A summary of the data analysis made in this thesis can be seen in Figure 5.1. The data processing and analysis is designed so that the challenges of big data, described in Chapter 4, are addressed in order to extract value adding information.

By using data analysis techniques, such as dimensionality reduction, clustering and model estimation, and combining them in an overall process of multiple steps, a way of differentiating groups of trucks being used in the same way is found. These groups, with trucks originating from a large population, are then analyzed to see what affects the fuel consumption the most. The implementation is made using MATLAB [MATLAB, 2014].

The data analysis is divided into three steps and each step contains smaller tasks,

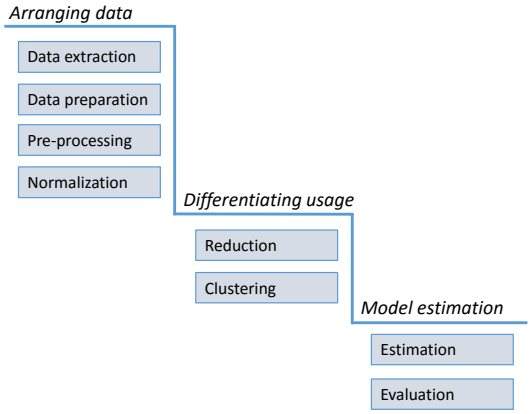

Figure 5.1: *An overview of the different steps of the data analysis and what their purpose is.*

as seen in Figure 5.1. **Arranging data** is the first step and includes extracting a population, choosing interesting parameters related to fuel consumption and putting together a dataset for further analysis, addressing the issues of *variety*, *veracity* and *volume*. Arranging data is necessary for the following steps and to handle big data, but is not the focus of the data analysis. This step is further described in Chapter 6. In the second step **differentiating usage**, the parameters are reduced and groups of trucks are found based on patterns in the reduced parameters. Here *volume* and *variety* are handled. Techniques such as Factor Analysis, Principal Component Analysis, k-means Clustering and Hierarchical Clustering are used. This is further described in Chapter 7. The last step, **model estimation**, describes how fuel consumption and what affects it varies among the groups of trucks. The idea is to estimate a simple model for fuel consumption, containing only the relevant parameter information. The model estimation techniques covered are Principal Component Regression, the Lasso and Elastic Net. This step is covered in Chapter 8.

5.4 Research methodology criticism

There are however a number of challenges in conducting case research. It is time consuming, it needs skilled interviewers, and care is needed in generalizing from a limited set of cases and in ensuring rigorous research [Voss et al., 2002].

Because of the limited resources available, a single-case study approach was chosen for this thesis. Other delimitations also had to be made as the work progressed due to the limitation of resources, which are described in Section 1.5.

Interviews were conducted, but not as the primary source and not as the primary

method of data collection. The primary source was the researchers' direct observations. The interviews served as sources of information so that the researchers were able to do the direct observation of the databases and data, since there were some tools needed to access these. During these interviews, other information was also unraveled, such as how the interviewees experience the tools and their experience of how others use them. This information helped build the context and motivation of the study.

Regarding the issue of generalizing from case study findings, there are some criteria for doing this: (1) although each case is in some respects unique, it is also a single example of a broader class of things, (2) the extent to which findings from the case study can be generalized to other examples in the class depends on how far the case study example is similar to others of its type, and (3) reports based on the case study include sufficient detail about how the case compares with others in the class for the reader to make an informed judgement about how far the findings have relevance to other instances [Denscombe, 2007].

6

Arranging data

This chapter includes a description of the criteria the choice of population is based on. How the criteria need to be combined to yield a population suitable for further analysis is explained. A thorough pre-processing to remove outliers and handle missing values is made to differentiate subgroups in the population.

6.1 Data extraction

Since there is a large number of models and truck configurations with different specifications in the database, attention has to be put on the heterogeneity of the population. The data analysis is developed to be used on all kinds of trucks in the whole heterogeneous population of the database. However, if the trucks in the chosen population are too similar, there is a risk that the patterns of variations in usage are less obvious. Focus is primarily on including trucks which have a high business benefit, i.e. the trucks Volvo Group Trucks produce and sell the most, but also trucks of which the knowledge about their configurations is easily accessible. To get a reasonable size of the population, dealing with the issue of *volume*, it has to be limited by some criteria, see Table 6.1. These requirements and criteria result in a population containing long haul trucks of various configurations.

The selection of parameters which are assumed to be related to fuel consumption is mainly based on interviews performed with a number of GTT employees with long experience of truck fuel consumption. The final selection includes parameters such as vehicle speed, engine torque, weight, topography as well as use of gears and cruise control. Moreover, an exploratory approach is to some extent used, which is a philosophy of data analysis where the data is approached with-

Table 6.1: *The chosen configurations from which the population of trucks are selected.*

Configuration dimension	Selected value
Brand	Volvo
Model	FH, FM
Engine emission level	Euro 5, Euro 6
Minimum operating hours	500 h
Engine version	340 - 700 HP
Gear shifting system	Automated

out making any assumptions about it. That is, letting the data itself tell the examiner about the phenomena [Martinez and Martinez, 2005]. A second selection of parameters is made based on what parameters are available in the database. Those are parameters logging the use of brakes, the fuel consumption for various driving modes, when the trucks have been driven manually and when the trucks have been coasting. A summary of the 28 chosen parameters can be seen in Table 6.2.

All of the parameters of interest are not available as single values; gross combination weight (GCW, also mentioned as vehicle weight), road slope, vehicle speed and engine torque are logged and stored in a vector format. Each parameter has an aggregated distance, time or fuel consumed stored for each element in the vector and also a corresponding axis giving information about what the aggregated quantity corresponds to. Figure 6.1 shows a graph of the GCW parameter for one specific truck, where the percentage of the total distance driven for each weight class is calculated. The road slope parameter is stored in a similar way, but with aggregated distances for different slope intervals.

6.2 Data preparation

Since all of the included parameters need to be on the same form to assure that the issue of *variety* of the data is handled, new single value parameters are extracted from the parameters on vector format.

Two ways for handling of parameters on vector format are tested. One of them extracts new variables from summary statistics, such as skewness, kurtosis, variance, lower and upper quartile. This is successfully used in the work by Wang et al. [2006], to handle clustering of time series data of different lengths, and in the work by Nayak et al. [2010], where the goal is to detect skin pathology by classifying normal and pathological conditions. Another, more straightforward way, is to divide the vector into as many single values as there are elements in the vector and add these to the population.

The approach using summary statistics is initially tested. However, this approach is not a good choice when the data is already on histogram form and therefore

Table 6.2: *Initially chosen parameters. The parameters of size* 28×1 *are vectors containing 28 values.*

Description	Size
Main log brake distance	1
Main log coasting distance	1
Main log econ distance	1
Main log pedal distance	1
Main log PTO[1] distance	1
Main log brake time	1
Main log cruise distance	1
Main log cruise fuel	1
Main log cruise time	1
Main log drive distance	1
Main log drive fuel	1
Main log drive time	1
Main log idle fuel	1
Main log PTO fuel	1
Main log coasting time	1
Main log economical time	1
Main log idle time	1
Main log pedal time	1
Main log PTO time	1
Main log economical fuel	1
Main log pedal fuel	1
Distance in top gear	1
Fuel in top gear	1
Time in top gear	1
Total fuel consumption	1
Top Gear -1 Mode Total Distance	1
Vehicle weight distance	28×1
Road slope distance	28×1

the second approach is used. In Figure 6.1 it is clear that many of the weight classes do not correspond to the truck usage, having no values stored for some weight classes, and therefore new weight intervals are formed, as can be seen in Figure 6.2. The same conclusion can be drawn from the road slope vectors and new intervals are formed as can be seen in Figure 6.3. How these weight and road slope intervals are formed is summarized in Table 6.3.

Neither vehicle speed nor engine torque is included in the final choice of parameters due to logging or data handling problems. The corresponding information to these parameters, explaining what each element in the vector corresponds to, do not have the same number of elements and therefore the data cannot be interpreted correctly.

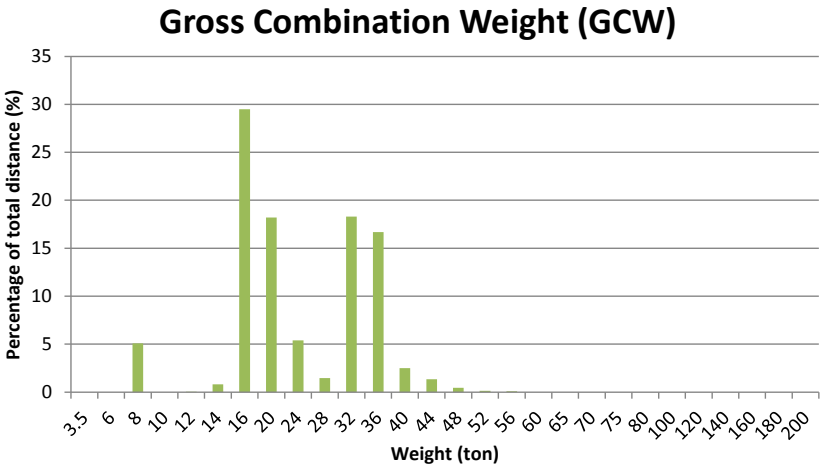

Figure 6.1: *An example of how one of the feature vectors, GCW, is stored in the database. An accumulated distance is stored for 28 weight classes, ranging from 3.5 to 200 tons. For simplicity, the percentage of the total distance instead of the accumulated distance for each weight class is shown in this graph.*

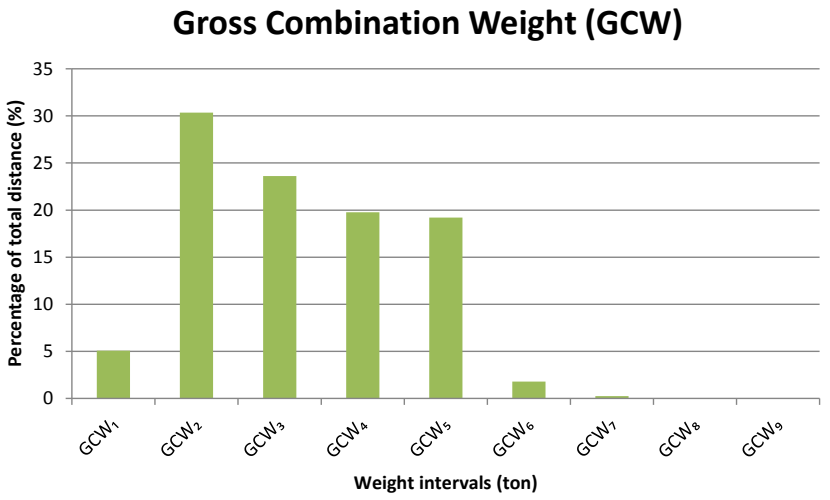

Figure 6.2: *An example of how one of the feature vectors, GCW, is modified before beeing included. Weight intervals containing several weight classes are formed.*

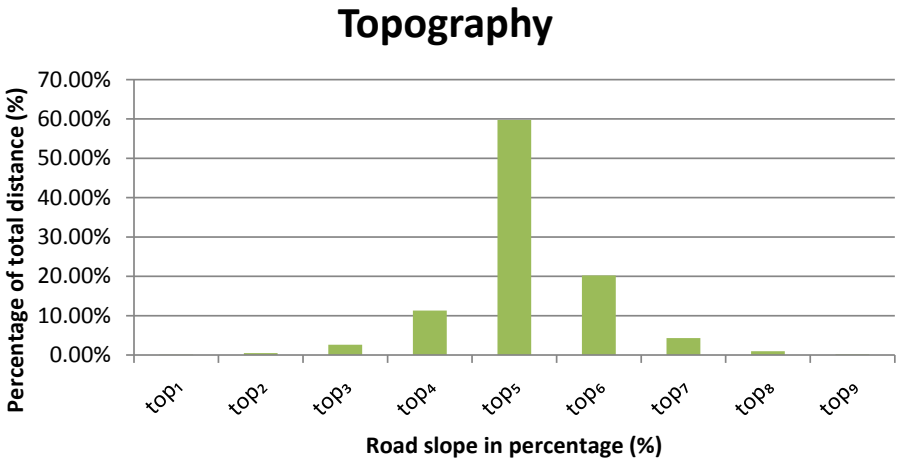

Figure 6.3: *An example of how one of the feature vectors, road slope, is modified before beeing included. Road slope intervals containing several road slope classes are formed.*

Table 6.3: *Final choice of normalized parameters included in the population, originating from the feature vector parameters.*

Parameter	Notation
GCW 3.5 to 10 tons percentage	GCW_1
GCW 12 to 16 tons percentage	GCW_2
GCW 20 to 24 tons percentage	GCW_3
GCW 28 to 32 tons percentage	GCW_4
GCW 36 to 40 tons percentage	GCW_5
GCW 44 to 48 tons percentage	GCW_6
GCW 52 to 56 tons percentage	GCW_7
GCW 60 to 65 tons percentage	GCW_8
GCW 70 to 200 tons percentage	GCW_9
Road gradient -20% to -11% percentage	top_1
Road gradient -10% to -8% percentage	top_2
Road gradient -7% to -5% percentage	top_3
Road gradient -4% to -2% percentage	top_4
Road gradient -1% to 1% percentage	top_5
Road gradient 2% to 4% percentage	top_6
Road gradient 5% to 7% percentage	top_7
Road gradient 8% to 10% percentage	top_8
Road gradient 11% to 20% percentage	top_9

6.3 Pre-processing

Since the data contains redundancies together with inconsistent and incomplete data, a thorough pre-processing of the data is required to handle the issue of *veracity*. The two main pre-processing steps are outlier removal and handling of missing values in the data. 65 percent of the population, i.e. 6,306 trucks, are removed due to missing parameter values. This is further described in Section 6.4. After that 33 percent of the remaining trucks, i.e. 1,124 trucks, are classified as outliers and therefore removed as well. In total only 23 percent of the initial population is left.

To get an idea of what kind of pre-processing steps are needed, a first step is to visualize the original data using scatter plots. A visualization easily shows the need of outlier removal and an example of this can be seen in Figure 6.4 for the parameter *Idle Time*. The same procedure is repeated for all parameters.

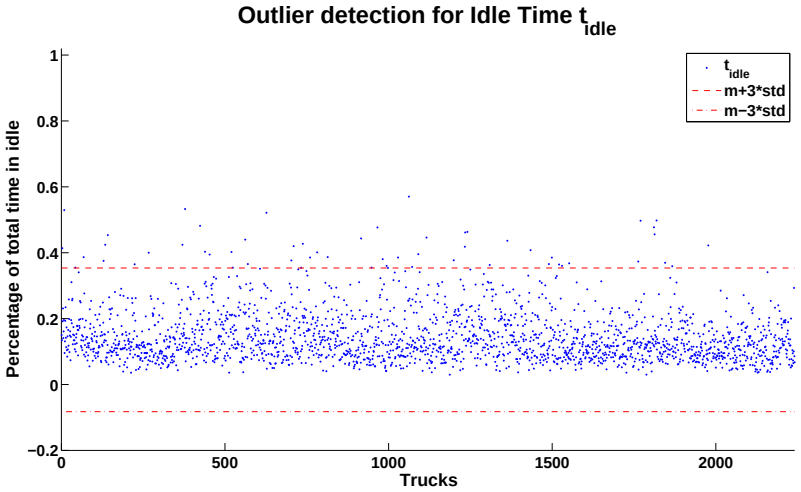

Figure 6.4: *The idle time parameter for the entire population where each point represents one truck. The two dotted lines represent the interval from which outliers are defined. Trucks outside the interval are removed from the population.*

An observation x_{ij} for a specific parameter j is defined as an outlier if it deviates more than three times the standard deviation, σ_j from the mean, μ_j, of the parameter, according to

$$|x_{ij} - \mu_j| > 3\sigma_j \tag{6.1}$$

If an observation is outside this interval, all observations originating from this truck are removed from the population since the data analysis described in Sections 7.1 and 7.2 otherwise will include the missing parameter values in their algorithms. The result would then be based on if the data has been logged cor-

rectly and not on how the trucks are actually used, which is not the purpose of this analysis.

6.4 Normalization

The algorithms in the data analysis described later in Sections 7.1 and 7.2 require the parameters to be comparable. The parameter values are therefore normalized so that the parameters represent a percentage of the total physical quantity rather than their given physical quantity. The normalizing factors are total time, distance and volume. The final set of normalized parameters included in the population can be seen in Table 6.3 and Table 6.4.

A visualization of the data after this step easily shows if the parameter values are acceptable or if there is a need to go back to the previous pre-processing step described in Section 6.3. None of the values should be larger than one since that would mean they represent more than 100 percent. The values above 100 percent are therefore classified as outliers and removed since they indicate that something is incorrect with the data, either in the logging or in the handling of the data. For example the parameter distance in top gear percentage, d_{top1}, is normalized with the total distance driven by the truck. If the total distance is smaller than the distance logged for the top gear, a value larger than 100 percent would be given. This indicates thus that something is incorrect with the logged data and this truck is therefore removed.

The data analysis described in Section 7.1 also requires a normalization, known as *whitening*, where the data is given zero mean by simply subtracting the mean of each parameter from each data point x_n, according to

$$y_n = x_n - \bar{x} \tag{6.2}$$

where $\bar{x}$ is the sample mean [Bishop, 2006].

Table 6.4: *Final choice of normalized parameters included in the population, originating from single value parameters.*

Parameter	Notation
Brake distance percentage	d_{brake}
Coasting distance percentage	d_{coast}
Economical driving distance percentage	d_{econ}
Pedal driving distance percentage	d_{pedal}
PTO distance percentage	d_{pto}
Cruise distance percentage	d_{cruise}
Drive distance percentage	d_{drive}
Distance in top gear percentage	d_{top1}
2nd highest gear distance percentage	d_{top2}
Brake time percentage	t_{brake}
Coasting time percentage	t_{coast}
Economical driving time percentage	t_{econ}
Pedal driving time percentage	t_{pedal}
PTO time percentage	t_{pto}
Cruise time percentage	t_{cruise}
Drive time percentage	t_{drive}
Idle time percentage	t_{idle}
Time in top gear percentage	t_{top1}
Coasting fuel percentage	f_{coast}
Economical driving fuel percentage	f_{econ}
Pedal driving fuel percentage	f_{pedal}
PTO fuel percentage	f_{pto}
Cruise fuel percentage	f_{cruise}
Drive fuel percentage	f_{drive}
Idle fuel percentage	f_{idle}
Fuel in top gear percentage	f_{top1}
Average fuel per kilometer	f_{total}

7

Differentiating usage

This chapter describes how properties explaining truck usage could be found using dimensionality reduction techniques and how, based on these properties, clusters could be discovered. The clusters are supposed to represent groups of trucks being used in the same way.

7.1 Dimensionality reduction

To keep an open mind about what affects fuel consumption, as many as 43 parameters were chosen for further investigation, see Chapter 6. Since some of the chosen parameters only differ in terms of units, such as the two parameters logging the distance the truck has used cruise control (d_{cruise}) and the time the truck has used cruise control (t_{cruise}), they are most likely correlated. Dimensionality reduction is then used to find underlying structure in the data while at the same time reduce dimensions. This can be seen as an alternative to comparing the end result when different subsets of parameters believed to be influencing fuel consumption have been chosen. The latter can lead to loss of useful information as some parameters are excluded [Martinez and Martinez, 2005].

More precisely, finding underlying structures in the data means that new variables are constructed, which are functions of the original ones. In our case, these new variables can be seen as properties describing truck usage. The new variables can then be used to transform the original data by projecting the data onto the new variables [Martinez and Martinez, 2005, James et al., 2013].

There are many approaches to dimensionality reduction. In this thesis, Principal

Component Analysis (PCA) and Factor Analysis (FA) are used to reduce the dimensions while keeping as much information about the original variables as possible [Martinez and Martinez, 2005].

This adresses the difficulty of the *volume* of the data. A first inspection of the data is done to verify that the parameters are correlated to some extent and with this motivate the dimensionality reduction. This is done by computing the sample correlation matrix of the data

$$\Sigma = \begin{pmatrix} \rho_{1,1} & \rho_{1,2} & \cdots & \rho_{1,j} \\ \rho_{2,1} & \rho_{2,2} & \cdots & \rho_{2,j} \\ \vdots & \vdots & \ddots & \vdots \\ \rho_{i,1} & \rho_{i,2} & \cdots & \rho_{i,j} \end{pmatrix}. \tag{7.1}$$

Two parameters x_i and x_j are correlated to some extent if $\rho_{ij} \neq 0$ [De Veaux et al., 2010].

Both PCA and FA are compared regarding how well they can describe new properties functioning as a projection of the population on a lower dimensional space. These properties span a reduced population working as the input to the clustering described in Section 7.2.

7.1.1 Principal Component Analysis

PCA is a method used to reduce the dimension of a dataset from d to p dimensions, where $p < d$, while at the same time capturing as much of the variations in the original data as possible [Martinez and Martinez, 2005]. Furthermore, PCA can be defined as the orthogonal projection of the data onto a lower subspace, such that the explained variance of the projection is maximized [Bishop, 2006]. The subspace of the projected data then represents a new orthogonal set of variables [James et al., 2013], *principal components*, which are linear combinations of the old variables [Martinez and Martinez, 2005]. In our case, these principal components can be seen as properties defining truck usage, which later on will be used to model fuel consumption.

There are two approaches that can be used when performing PCA; using the *sample covariance matrix* of the data or using the *sample correlation matrix*. In this thesis the sample covariance matrix is used on centered data, a subtraction of the mean of each sample. This is actually close to the sample correlation matrix since the latter finds the covariance of standardized samples on both the mean and standard deviation.

Consider a dataset

$$\mathbf{X} = \begin{pmatrix} x_{1,1} & x_{1,2} & \cdots & x_{1,d} \\ x_{2,1} & x_{2,2} & \cdots & x_{2,d} \\ \vdots & \vdots & \ddots & \vdots \\ x_{N,1} & x_{N,2} & \cdots & x_{N,d} \end{pmatrix}, \tag{7.2}$$

Algorithm 1 PCA

1. Center the data around the mean by subtracting the sample mean
$$\bar{x}_i = \frac{1}{N}\sum_{n=1}^{N} x_{n,i} \tag{7.3}$$
for each variable x_i.

2. Compute the data covariance matrix
$$C = \frac{1}{N} X_c^T X_c \tag{7.4}$$
where X_c is the centered data.

3. Find the eigenvalues of C and choose the p largest eigenvalues using a scree plot and cumulative percentage of variance explained described in the following section. The corresponding eigenvectors, E_p, are the principal components, the new variables, of the reduced dataset.

where x_{ij} denotes parameter j for truck i. The procedure used for PCA, applied to the dataset $\mathbf{X}$ in (7.2), is summarized in Algorithm 1, based on Martinez and Martinez [2005], Marsland [2009] and Bishop [2006].

Evaluation of the number of principal components to keep

There are several approaches available for deciding how many principal components to keep. Two of them used in this thesis are the graphical method using a scree plot and the cumulative percentage of variance explained.

In a *scree plot* the eigenvalues of the principal components are plotted as a function of their indices, see Figure 7.1. To decide how many components to retain, one looks for an "elbow" in the graph, i.e. the point after which the curve levels off [Martinez and Martinez, 2005].

When using the *cumulative percentage of variance explained* the idea is to select the p principal components contributing to a percentage of the total variation in the data, according to
$$\sigma^2_{explained} = 100\frac{\sum_{i=1}^{p}\lambda_i}{\sum_{j=1}^{d}\lambda_j}, \tag{7.5}$$
where λ_i denotes the eigenvalue of the i:th principal component. The percentage chosen is typically between 70% and 90% [Jolliffe, 1986].

In the scree plot of the principal components, see Figure 7.1, one eigenvalue is significantly larger than the remaining and at least three components seem to have eigenvalues deviating from the rest. By investigating the explained variance of the components, see Table 7.1, an additional component, giving four components in total, seems appropriate as this would lead to a cumulative percentage of variance explained of 78.2%. However, a cumulative percentage of variance

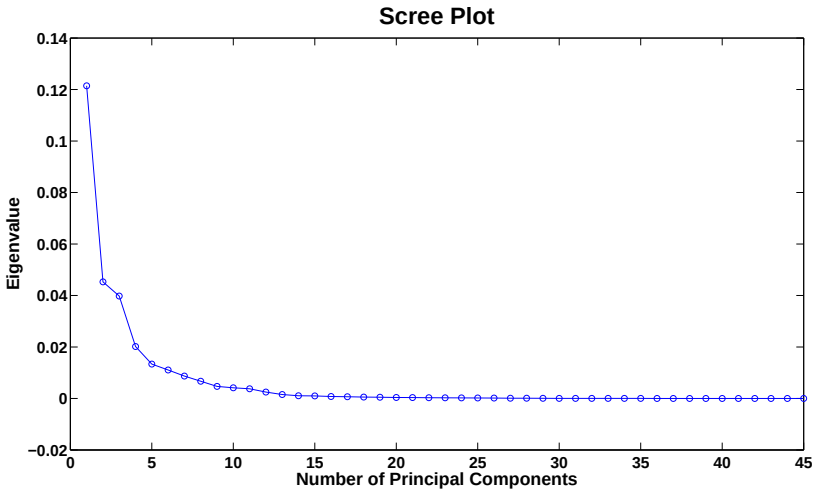

Figure 7.1: *Scree plot showing the size of the eigenvalue for each principal component, which can be used to decide on how many components to include. One generally looks for an "elbow" in the plot, which here can be seen around 4 and 10 components.*

explained of up to 90% is generally a good rule of thumb but as this would mean seven components would need to be included, four components seem more appropriate since the explained variance is still quite high. One generally seeks to be able to describe the population with as few components as possible. In the end we want a sparse model so that it is easy to distinguish what role the different parameters have in each cluster.

An orthogonal projection of the data onto the principal components now gives the low-dimensional subspace of the original data [Murphy, 2012]. This transformed data is now to be used in the clustering, see Section 7.2.

Interpretation of the principal components

To be able to interpret the meaning of the components, a summary of which parameters that affect each component the most can be seen in Table 7.2. Only parameters having loadings greater than 0.3 are included in the table, to facilitate the interpretation. From the size of the parameter loadings in each component, an interpretation is made, which is of course subjective.

The first component, PC_1, only contains parameters related to the use of pedal, i.e. when the trucks have been driven with the accelerator, and the use of cruise control. The parameters related to the usage of the pedal are also of opposite sign relative the parameters connected to the usage of cruise control and the parameter d_{cruise} has the largest loading. An interpretation is therefore that this component represents how the truck is normally driven, i.e. how much cruise control is used. Cruise control mode and the usage of the accelerator are natural

Table 7.1: *In the first column the variance contribution in percent of the total variance for each principal component is presented while the second column presents the cumulative percentage of variance explained for each additional principal component.*

	Variance (%)	**Cumulative variance (%)**
PC_1	41.9	41.9
PC_2	15.6	57.5
PC_3	13.7	71.2
PC_4	7.0	78.2
PC_5	4.6	82.8
PC_6	3.8	86.6
PC_7	3.0	89.6
PC_8	2.3	91.9
PC_9	1.6	93.5
PC_{10}	1.4	94.9

Table 7.2: *A summary of the parameters being most important in each principal component. The loading of the parameter in the principal component decides how much it affects the component. Parameters with loadings larger than 0.3 are included in the table.*

PC_1	**Parameter**	**Loading**	**PC_2**	**Parameter**	**Loading**
	d_{pedal}	-0.3722		f_{top1}	0.3168
	d_{cruise}	0.5028		t_{top1}	0.3509
	f_{cruise}	0.4731		GCW_4	0.3907
	t_{cruise}	0.3549		GCW_5	0.3008
	f_{pedal}	-0.3348			
PC_3	**Parameter**	**Loading**	**PC_4**	**Parameter**	**Loading**
	GCW_4	-0.5592		GCW_4	0.3983
	GCW_6	0.6189		GCW_5	-0.8034
PC_5	**Parameter**	**Loading**	**PC_6**	**Parameter**	**Loading**
	t_{drive}	-0.315		GCW_6	-0.4852
	d_{top1}	0.3119		GCW_7	0.5251
	top_6	0.4073		top_6	-0.3013
PC_7	**Parameter**	**Loading**	**PC_8**	**Parameter**	**Loading**
	GCW_3	0.6635		d_{top1}	0.3699
	GCW_6	-0.3434		GCW_3	0.4879
				top_6	-0.4321
PC_9	**Parameter**	**Loading**	**PC_{10}**	**Parameter**	**Loading**
	d_{pedal}	-0.3112		d_{econ}	0.344
	d_{cruise}	-0.3975		f_{top1}	-0.3382
	f_{pedal}	-0.3087		f_{total}	0.3417
				GCW_7	-0.4095

opposites, as the sign of the loadings indicates.

With four loadings with the same sign and of approximately the same size, the second component, PC_2, is not as straightforward to interpret as the first one. It contains two parameters related to the usage of top gear and the remaining two are weight classes representing trucks having a gross combination weight of 28 to 40 tonnes. An interpretation is that the top gear is mainly used for these weight classes.

The third and fourth components, PC_3 and PC_4, are of similar character, both having two weight class parameters as the most important ones. This is probably a result of high variance in these parameters, making them important to describe the total variance of the population.

The rest of the components are not given any interpretation as the previous argumentation regarding the scree plot and explained variance indicates they should not be included. An overall observation, however, is that the weight class parameter GCW_4 appears in several components (PC_2, PC_3 and PC_4) which indicates that it is important. The total distance spent in this weight class seem to vary a lot between the trucks. The weight class also seem to have a relationship with some other parameters, as seen in PC_2.

7.1.2 Factor Analysis

FA aims to explain the data consisting of d variables with p number of uncorrelated *factors*, where $p < d$, originating from some underlying structure that is not directly known [Marsland, 2009]. The original variables can then be expressed as a linear combination of the underlying factors f_i [Martinez and Martinez, 2005] according to

$$\underbrace{\begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_d \end{pmatrix}}_{=X} = \underbrace{\begin{pmatrix} a_{1,1} & a_{1,2} & \cdots & a_{1,p} \\ a_{2,1} & a_{2,2} & \cdots & a_{2,p} \\ \vdots & \vdots & \ddots & \vdots \\ a_{d,1} & a_{d,2} & \cdots & a_{d,p} \end{pmatrix}}_{=A} \underbrace{\begin{pmatrix} f_1 \\ f_2 \\ \vdots \\ f_p \end{pmatrix}}_{=F} + \underbrace{\begin{pmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_d \end{pmatrix}}_{=\epsilon} \tag{7.6}$$

where $a_{1,1}, \ldots, a_{d,p}$ are called the *factor loadings* and ϵ_i are error terms [Martinez and Martinez, 2005, Hastie et al., 2008]. The error terms ϵ_i are assumed to be uncorrelated with each other and have zero mean [Martinez and Martinez, 2005]. The factor loadings are used to interpret the factors [Hastie et al., 2008]. In our case the factors are used to translate the parameter variations into truck usage. The idea is that there are several truck usages hidden in the data and that truck usage can be defined by more than one parameter. By identifying these patterns it can be easier to understand how different truck usages affect fuel consumption.

The factor loadings for a dataset as in (7.2) can be found in a similar way to the procedure used to find the principal components. The procedure is summarized

in Algorithm 2, based on Marsland [2009] and Murphy [2012].

Algorithm 2 FA

1. Center the data around the mean by subtracting the sample mean according to (7.3) for each variable x_i.

2. Assume the model
$$\mathbf{X_c} = \mathbf{AF} + \epsilon \tag{7.7}$$
with
$$\mathbf{A} = \begin{pmatrix} a_{1,1} & a_{1,2} & \cdots & a_{1,d} \\ a_{2,1} & a_{2,2} & \cdots & a_{2,d} \\ \vdots & \vdots & \ddots & \vdots \\ a_{p,1} & a_{p,2} & \cdots & a_{p,d} \end{pmatrix}$$
containing the factor loadings $a_{i,j}$, $\mathbf{F}$ the factors f_i and $\mathbf{X_c}$ the centered data. The factors and error terms are assumed to be independent.

3. The data covariance matrix is then on the form
$$\mathbf{C} = \mathbf{AA}^T + \mathbf{\Psi} \tag{7.8}$$
where $\mathbf{\Psi}$ is a diagonal matrix with the variance of ϵ in its diagonal elements.

4. Estimate $\mathbf{A}$ and $\mathbf{\Psi}$ with Maximum Likelihood using an Expectation Maximization (EM) algorithm. For further reading, see e.g. Murphy [2012].

To make the factors more interpretable, $\mathbf{A}$ is often rotated using some rotational method [Murphy, 2012]. One popular method is *varimax* [Murphy, 2012], which optimizes a criterion based on the variance of the loadings and gives orthogonal factors [Martinez and Martinez, 2005]. This is the rotational method used in this thesis.

After having found the factors using Algorithm 2 and varimax rotation, some kind of evaluation criteria have to be used to decide on how many factors to keep. Since four principal components seems to be a good choice above, this is used as an initial guess. The factors are then evaluated based on how interpretable they are, i.e. what kind of parameters they included and how large the loadings from each parameter are. The size of the loadings reflects how important the parameters are for each factor. A summary of how these four factors turned out can be seen in Table 7.3. When interpreting the factors, a common rule of thumb is to use a cutoff value of 0.3 to 0.4 to decide if a parameter has significant effect on each factor [Schmitt and Sass, 2011]. Only the factor loadings with a cutoff value larger than 0.3 are therefore shown.

Up to ten factors are considered, but since the factors are difficult to interpret and the interpretation process is very time consuming, four factors are chosen to be sufficient. When allowing more factors, the factors appearing when using a small number of factors appear again, but their loadings are slightly changed. FA does not yield any information regarding how well the factors represent the

population in terms of explained variance.

Table 7.3: *A summary of the most important parameters in each factor. The loading of the parameter in the factor decides how much it affects the factor.*

$Factor_1$	**Parameter**	**Loading**	$Factor_2$	**Parameter**	**Loading**
	top_5	-0.9862		t_{econ}	0.9683
	top_6	0.9786		t_{drive}	0.9441
	top_7	0.9128		f_{top1}	0.8199
	top_8	0.8322			
$Factor_3$	**Parameter**	**Loading**	$Factor_4$	**Parameter**	**Loading**
	d_{pedal}	0.9069		d_{top2}	0.5094
	d_{cruise}	-0.8992		f_{top1}	-0.4872
	f_{pedal}	0.8776		d_{top1}	-0.4300
	f_{cruise}	-0.8747		t_{coast}	0.4270
	t_{pedal}	0.8733			
	t_{cruise}	-0.8377			

Interpretation of the factors

An attempt to an interpretation of the factors in Table 7.3 is made, which is, as the interpretation of the principal components, subjective, but can be used to better understand the meaning of the factors.

A first interesting observation is that the third factor, Factor_3, is similar to the first principal component, PC_1, explaining most of the variance in PCA. Parameters related to the use of the accelerator and cruise control are included and the parameters related to cruise control are of opposite sign relative the ones connected to the use of the accelerator.

The first factor, Factor_1, only contains topography parameters logging distance in the interval –1% to 10% road slope.

The second factor, Factor_2, contains the parameter t_{econ} related to the usage of economy mode on the gearbox, which means that the automated gearbox prioritizes fuel economy, and another parameter related to the usage of the top gear. The last parameter included has logged when the truck has been in drive mode, i.e. not idling. An interpretation of this factor could be that economy mode and the highest gear are used frequently when driving.

Finally, the last factor, Factor_4, contains parameters connected to the use of the two top gears and if the truck has coasted. The sign of the usage of the second highest gear and the parameter indicating coasting are of opposite sign relative the ones connected to the usage of the highest gear.

7.1.3 Differences between PCA and FA

The major difference between PCA and FA is that the latter is based on a model, which is not the case for PCA. This leads to the two techniques also being computed in two very different ways. For PCA, the loadings are a result of computations whereas for FA, the loadings need to be estimated and afterwards rotated to give interpretable results [Trendafilov et al., 2013]. This leads to FA, in contrast to PCA, not having a unique solution [Martinez and Martinez, 2005].

The two methods both try to represent the data from the covariance (or correlation) matrix, but they focus on different parts of it [Jolliffe, 1986]. PCA focuses on the diagonal elements of the matrix whereas FA focuses on the off-diagonal elements. One can describe it as PCA is variance orientated whereas FA is covariance orientated [Lawley and Maxwell, 1963]. This can also give the two methods different results when it comes to their capability of reducing the dimensions of the dataset. If the original variables are nearly independent there will be a principal component corresponding to each one of the variables. In FA, a factor must contribute to at least two variables and thus a factor for each variable cannot be found as for PCA. However, since the aim is to use the components and factors to explain new properties from truck usage parameters, it is not interesting here to have only one variable per property.

Another difference is that when changing the dimensionality of the model in FA the solution, i.e. the loadings, does not need to stay the same [Martinez and Martinez, 2005, Jolliffe, 1986]. When an additional component is added in PCA the original components are still unaffected. In FA some will typically still be similar if the model is appropriate, but they do not need to stay the same [Jolliffe, 1986]. This phenomenon was seen when some factors seemed to move around as the dimensionality was chosen differently. However some factors such as Factor_3 always seemed to appear.

Since the loadings found using FA were significantly larger than the ones found for PCA, the interpretation of the factors of the former were more straightforward. However, the choice of how many factors to retain for FA is more complicated due to the fact that the loadings and found factors differ depending on the chosen dimensionality.

Furthermore, it is interesting that Factor_3 and PC_1 are very similar in terms of the most important parameters included in them. This underlines the fact that this factor best describes the variation in the data.

7.2 Clustering

Clustering is a collection of techniques that have the objective of grouping data into several homogeneous groups, *clusters*, in applications where partitioning of data is needed or so that a studied phenomenon is easier to understand and interpret [Bouveyron and Brunet-Saumard, 2012]. Clustering is performed to find out if there are certain trucks being used in a similar way. The idea is that trucks used

in the same way based on the properties describing truck usage, found from the dimensionality reduction described in Section 7.1, then will belong to the same cluster. Having grouped the trucks in this way, finding out what affects fuel consumption and how it varies depending on how the trucks have been used can be found.

Furthermore, clustering is an unsupervised learning method which means that, in contrast to supervised learning methods, the clusters can be formed from the data without any knowledge about the correct output [Marsland, 2009]. Various algorithms for clustering are available and which one to choose depends on what kind of data the clustering is to be performed on. In this thesis, the difficulty of the *variety* of the data is addressed through clustering by using the two algorithms *k-means Clustering* and *Hierarhical Clustering*. For these two algorithms the idea is to form clusters based on some dissimilarity measure.

7.2.1 Dissimilarity measures

To form clusters one has to decide on a dissimilarity measure, i.e. how the observations should be compared. Two common measures are the *squared Euclidean distance*, $\Delta_j^{(1)}(x_{ij}, x_{i'j})$, and the *city block distance*, $\Delta_j^{(2)}(x_{ij}, x_{i'j})$, given by

$$\Delta_j^{(1)}(x_{ij}, x_{i'j}) = (x_{ij} - x_{i'j})^2 \tag{7.9}$$

$$\Delta_j^{(2)}(x_{ij}, x_{i'j}) = |x_{ij} - x_{i'j}| \tag{7.10}$$

where Δ_j is the dissimilarity between trucks i and i' for parameter j from which a dissimilarity matrix

$$\mathbf{D} = \begin{pmatrix} d_{1,1} & d_{1,2} & \cdots & d_{1,N} \\ d_{2,1} & d_{2,2} & \cdots & d_{2,N} \\ \vdots & \vdots & \ddots & \vdots \\ d_{N,1} & d_{N,2} & \cdots & d_{N,N} \end{pmatrix} \tag{7.11}$$

can be formed [Murphy, 2012]. $d_{i,i'}$ are calculated from the dissimilarity measures as

$$d_{i,i'} = \sum_{j=1}^{p} \Delta_j(x_{ij}, x_{i'j}), \tag{7.12}$$

leading to the diagonal elements, $d_{i,i}$, being zero and $d_{i,i'} \geq 0$

According to Marsland [2009] the most commonly used dissimilarity measure for the k-means algorithm is the squared Euclidean distance, $\Delta_j^{(1)}$, and this is therefore the choice of dissimilarity measure in this thesis. To be able to compare the results from k-means and Hierarchical Clustering, the same dissimilarity measure is used for Hierarchical Clustering.

7.2.2 *k*-means Clustering

The k-means algorithm determines the clusters by considering how close a data point is to the center of a cluster, based on some dissimilarity measure. Therefore, this method is best suited when the variables are of the quantitative type [Hastie et al., 2008] and do not contain outliers as this would make the algorithm non-robust [Bishop, 2006]. These prerequisites are fulfilled with the used truck data as all of the parameters are of the quantitative type and in the pre-processing of the data, described in Chapter 6, outliers were removed.

The number of clusters, k, needs to be specified beforehand together with initial starting values for the center of each cluster. A data point belongs to the cluster it is closest to, with respect to the center of the cluster. The algorithm is an iterative optimization procedure where the center of each cluster is changing depending on the minimal squared Euclidean distance between a point in the cluster and its center. [Bishop, 2006]

A summary of the used k-means procedure can be seen in Algorithm 3, based on Murphy [2012].

Algorithm 3 k-means Clustering

1. Choose the number of clusters k.
2. Initialize the cluster centroids, the mean of the clusters, to $\mu_k = \mu_{k,init}$.
3. Assign each data point to its closest cluster centroid so that the Euclidean distance (7.9) between the data point x_i and the mean μ_k is minimized according to
$$z_i = \arg\min_k |x_i - \mu_k|^2 \tag{7.13}$$
where z_i is the cluster x_i belongs to.
4. Update the cluster centroids by computing the mean of all points assigned to it
$$\mu_k = \frac{1}{N_k} \sum_{i}^{N_k} x_i \tag{7.14}$$
where $k = z_i$ and N_k the number of data points in cluster k.
5. Repeat 3-4 until converged.

As input to the clustering is both the PCA reduced population and the FA reduced population. Since the two populations themselves do not have any clear classification that can indicate the number of clusters to use, different numbers of clusters are considered and evaluated. A graphical method available for cluster evaluation is parallel coordinate plots, see Figure 7.2. The parallel coordinate plots should look different with regard to the value spread on the vertical axis for each factor or principal component, for a specific number of clusters. If not, this indicates that a lower number of clusters might be a better choice. By studying the parallel coordinate plots for all clusters, having initialized the clustering with a certain k, the dissimilarity between the clusters can be compared. If distinctions between the clusters are not obvious from the parallel coordinate plots, some of the clusters

are considered redundant and can therefore be removed. How well the clusters are separated by the components or factors is also analyzed from scatter plot matrices of the components or factors plotted against each other. If obvious clusters are seen in these plots, the clustering is successful.

PCA reduced population

Six clusters are stated to be the best choice for the PCA reduced population as the clusters are reasonably different, see Figure 7.2. A scatter plot of the principal components where each cluster has an individual color, can be seen in Figure 7.3. Here the first component seems to separate the clusters well from each other.

Moreover, the parallel coordinate plot in Figure 7.2 also hints about the variance explained by each component. The first principal component has a wider span of values for all trucks than the other components, which underlines the fact that it explains 41.9% of the variance of the data, as seen in Table 7.1.

From the parallel coordinate plots and the scatter plot matrices interpretations of the clusters are made. This is done both when four principal components and six clusters are used and when four factors and four clusters are used. By comparing the sign of the principal component or factor for each cluster seen in the parallel coordinate plot, with the sign of the loadings for these principal components or factors in Table 7.3 and Table 7.2, an interpretation of what defines the clusters is made. Both alternatives have clusters mainly distinguishable from the use of pedal and cruise control, i.e. Factor_3 and PC_1.

The signs of the principal components for each cluster in Figure 7.2 are summarized in Table 7.4. The comparison gives the interpretation given in Table 7.5. If the sign of a principal component and the sign of a parameter loading is positive when multiplied, this parameter is interpreted as describing that cluster. If the sign of the principal component is both positive and negative in the parallel coordinate plot, no conclusions can be drawn about that component and its influencing parameters. PC_2 is difficult to interpret since its loadings are different from the other principal components. Its loadings are all rather small and of the same sign, which makes them difficult to interpret when the principal component has a negative sign in the clustering. Therefore the interpretation of this principal component is tentative.

FA reduced population

When having implemented Algorithm 3 on the population reduced with k-means Clustering using four factors the parallel coordinate plots, see Figure 7.4, do not directly hint about the clusters being distinguishable from each other. Therefore, both initializing the algorithm with fewer clusters and changing the number of components are tested. Figure 7.4 and Figure 7.5 show the result with four clusters and four factors, and four clusters and ten factors respectively. Including more factors does not seem to make the clusters easier to interpret, but Figures 7.6 and 7.7 suggest that four clusters is a better choice than six. With four clusters both Factor_3 and Factor_4 seem to separate the clusters. By looking at the scatter

Table 7.4: *The sign of each principal component in the six clusters when using k-means. If the cluster contained trucks taking both positive and negative values for this component this was indicated with +/–.*

	PC_1	PC_2	PC_3	PC_4
Cluster 1	–	+	+/–	+/–
Cluster 2	–	+/–	+	+/–
Cluster 3	+	–	+	+/–
Cluster 4	+	+/–	+/–	+/–
Cluster 5	+/–	+/–	–	+/–
Cluster 6	–	–	+/–	+/–

plot matrix in Figure 7.7 the clusters seem separable at least for Factor_3.

A similar interpretation is made using the signs of the parallel coordinate plot in Figure 7.4 together with the signs of the loadings in Table 7.3. The result of the interpretation can be seen in Table 7.6. No conclusions can be drawn from the first two factors since they look the same for all clusters.

7.2.3 Hierarchical Clustering

In Hierarchical Clustering a nested tree of clusters is created [Murphy, 2012]. At each level of the tree clusters are merged to create clusters on the next level, resulting in one observation in each cluster on the lowest level and one cluster for all data on the highest level [Hastie et al., 2008].

There are two ways of performing Hierarchical Clustering: bottom-up, *agglomerative clustering* and top-down, *divisive clustering*. The agglomerative approach starts at the bottom and groups the observations so that a higher level contains one less cluster in contrast to the divisive approach which instead starts at the top and splits the clusters at lower levels so that lower levels contain less clusters. [Murphy, 2012, Hastie et al., 2008]

As input to both agglomerative and divisive clustering is a dissimilarity matrix based on some dissimilarity measure [Murphy, 2012]. Agglomerative clustering is the approach which has been studied the most [Hastie et al., 2008] and is therefore the choice for this thesis.

The clusters can be visualized in a binary tree, a *dendrogram*, see Figure 7.8. The nodes of the tree represent clusters and the root node the complete dataset. Each level in the tree is therefore a possible choice of as many clusters as there are vertical lines on that level. The horizontal axis represents the number of observations in each cluster in the leaf nodes and the vertical axis represents the dissimilarity between the clusters on a given level. [Murphy, 2012, Hastie et al., 2008]

For agglomerative clustering, see Algorithm 4, merging of clusters can be done in three different ways: using *single linkage*, *complete linkage* or *average linkage*. These are all measures for dissimilarity between the two merged clusters. Single

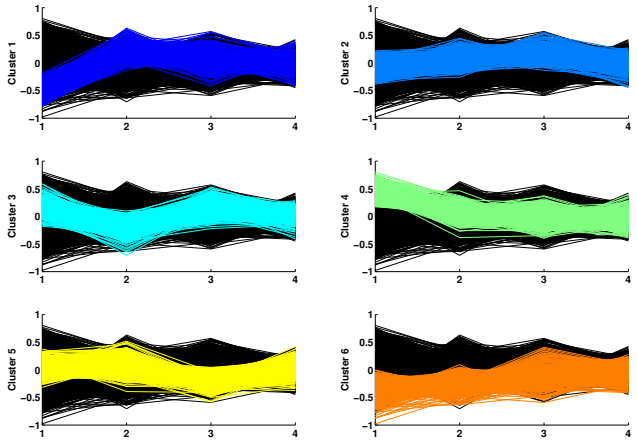

Figure 7.2: *Parallel coordinate plot of the population reduced with four principal components and clustered with the k-means algorithm. Each line in the plot represents a truck and the horizontal axis holds the four principal components, so that the vertical axis indicates how much of each principal component is affecting each truck.*

Table 7.5: *Interpretation of the six clusters found using k-means on the population reduced with four principal components.*

	Interpretation
Cluster 1	For these trucks the pedal has been used a lot and the trucks seem to have been driven in the top gear and heavily loaded.
Cluster 2	For these trucks the pedal has been used a lot and the trucks have been very heavily loaded.
Cluster 3	For these trucks the cruise control has been used a lot, but no conclusions can be drawn about the weight.
Cluster 4	For these trucks the cruise control has been used a lot, but no conclusions can be drawn about the weight.
Cluster 5	No conclusions can be drawn about the usage of pedal or cruise control, but the trucks seem to be less heavy than in e.g. Cluster 2.
Cluster 6	For these trucks the pedal has been used a lot, but no conclusions can be drawn about the weight.

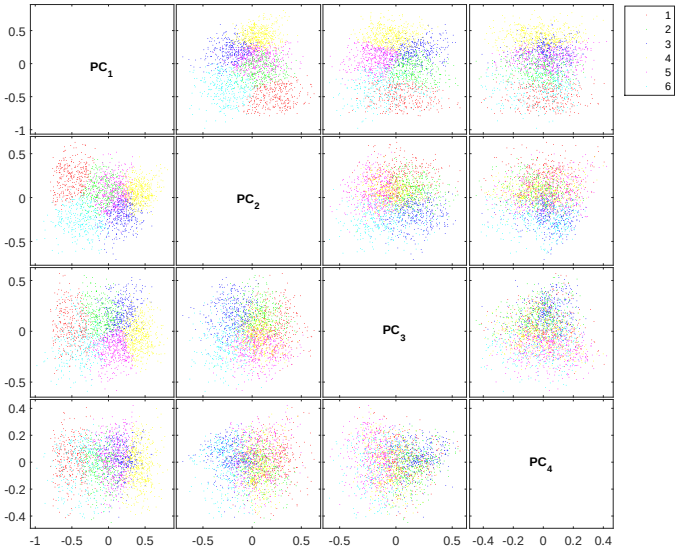

Figure 7.3: *Scatter matrix of the population reduced with four principal components and clustered with the k-means algorithm using 6 clusters. Each dot represents an observation and the color separates the clusters from each other. Each component is plotted against all other components. On the first row and column separable clusters can be seen.*

Table 7.6: *Interpretation of the four clusters found using k-means on the population reduced with four factors.*

	Interpretation
Cluster 1	No conclusions can be drawn regarding use of pedal and cruise control but it seems as the trucks have been driven on the top gear a lot.
Cluster 2	Here cruise control has been used more than pedal driving and nothing can be said about the usage of gears.
Cluster 3	Here the pedal has been used more than cruise control and the trucks have been driven on the top gear a lot.
Cluster 4	No conclusions can be drawn regarding use of pedal and cruise control and nothing can be said about the usage of gears.

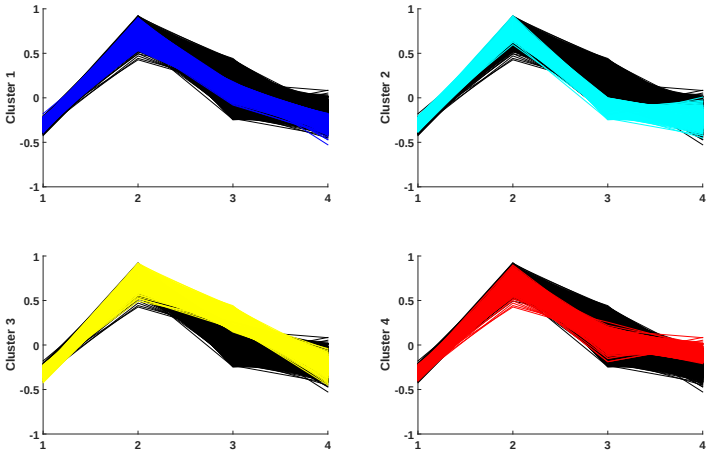

Figure 7.4: *Parallel coordinate plot of the population reduced with four factors and clustered with the k-means algorithm. The lines in the plot represent trucks and the horizontal axis holds the four factors, so that the vertical axis indicates how much of each factor is affecting each truck.*

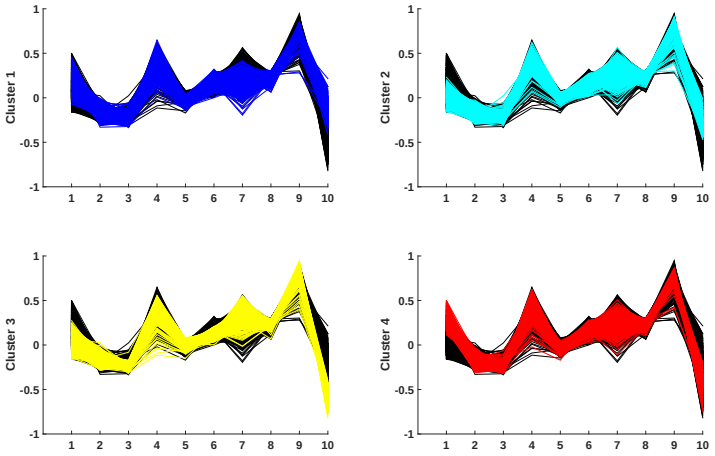

Figure 7.5: *Parallel coordinate plot of the population reduced with ten factors and clustered with the k-means algorithm. The lines in the plot represent trucks and the horizontal axis holds the four factors, so that the vertical axis indicates how much of each factor is affecting each truck.*

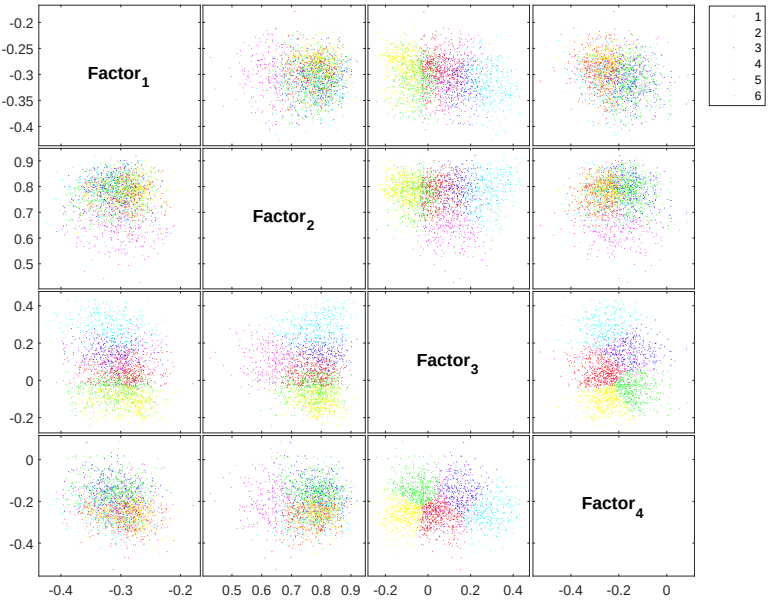

Figure 7.6: *Scatter matrix of the population reduced with four factors and clustered with the k-means algorithm.*

linkage measures the distance between the two closest members of the clusters whereas complete linkage measures the distance between the two members being the furthest away from each other. Average linkage measures the average distance between all pairs in the clusters. [Murphy, 2012]

Algorithm 4 Agglomerative Hierarchical Clustering

1. Initialize as many clusters $C_i = \{i\}$ as there are data points $i = 1 \dots n$.
2. Pick the two most similar clusters
$$(j, k) = \arg\min_{j,k \in S} d_{j,k} \tag{7.15}$$
3. Create new cluster
$$C_l = C_j \cup C_k \tag{7.16}$$
and mark j and k as unavailable for merging.
4. Update dissimilarity matrix $d(i, l)$
5. Repeat 2 - 4 until no more clusters are available for merging.

Agglomerative clustering with complete and average linkage is tested, using squared Euclidean distance as dissimilarity measure. The number of clusters used are evaluated using a dendrogram plot, comparing up to thirty clusters.

In Figure 7.8 the dendrogram formed using average linkage on the population reduced with PCA is shown. The horizontal axis represents 30 clusters, numbered from 1 to 30. The vertical line connecting leaves 26 and 27 is the longest in the tree and indicates a strong dissimilarity between the two clusters joined. By counting the lines aggregating into the nodes closest to the level where this longest line ends, four clusters is suggested to be a good choice.

The same reasoning as above can for the the dendrogram formed using complete linkage on the population reduced with PCA, see Figure 7.9, lead to the conclusion that either four or two clusters is a good choice. Since there are several lines of equal length, a definite choice cannot be made.

For the population reduced with FA using average linkage, see Figure 7.10, a conclusion about the number of clusters to use is again difficult to draw as there are several long vertical lines. However, when using complete linkage, see Figure 7.11, the two vertical lines to the left, ranging from 10 to 18 on the vertical axis can easily be stated as the longest. Four clusters therefore seems to be an adequate choice.

The results for the choice of four clusters using average linkage on the PCA reduced population and four clusters using complete linkage on the FA reduced population is now explored further using scatter plot matrices of the components and factors respectively. The population reduced with PCA, see Figure 7.12 shows an interesting result were one cluster only contains one truck.

Since the choice of clusters is insure when using Hierarchical Clustering, further interpretation of these clusters is not made.

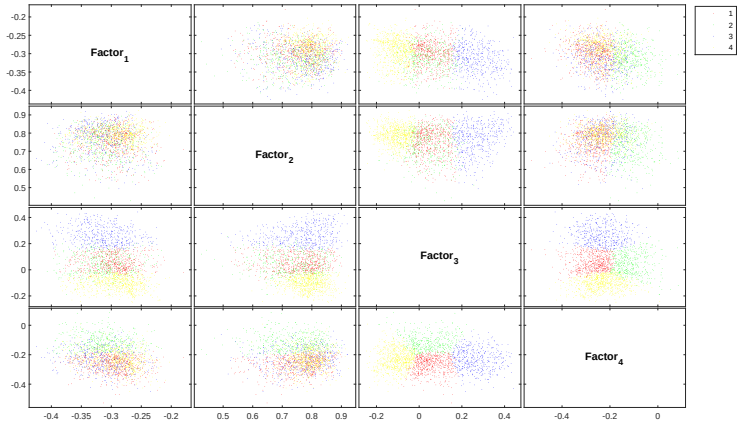

Figure 7.7: *Scatter matrix of the population reduced with four factors and clustered with the k-means algorithm. Each point represents one observation and the color separates the clusters from each other. Each component is plotted against all other components. On the first row and column separable clusters can be seen.*

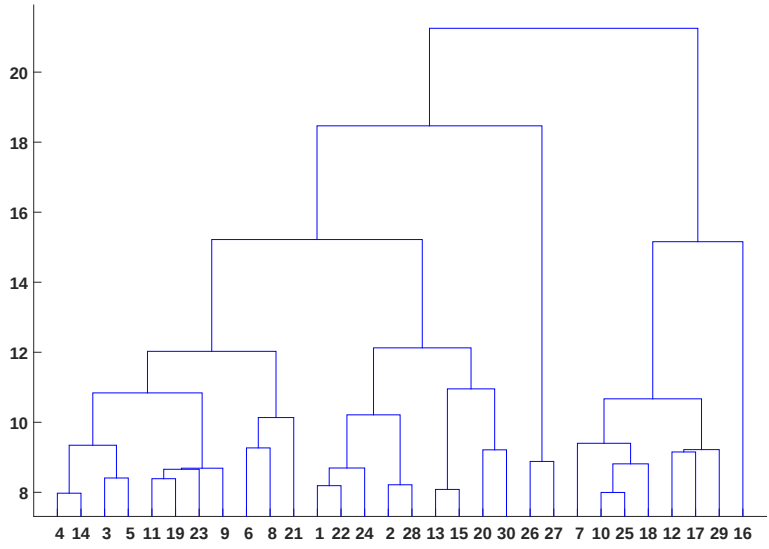

Figure 7.8: *Dendrogram for agglomerative clustering using average linkage on the population reduced with PCA.*

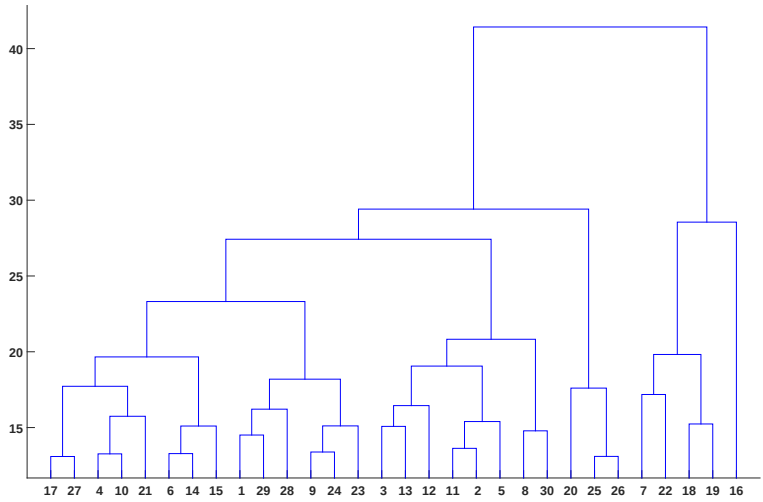

Figure 7.9: *Dendrogram for agglomerative clustering using complete linkage on the population reduced with PCA.*

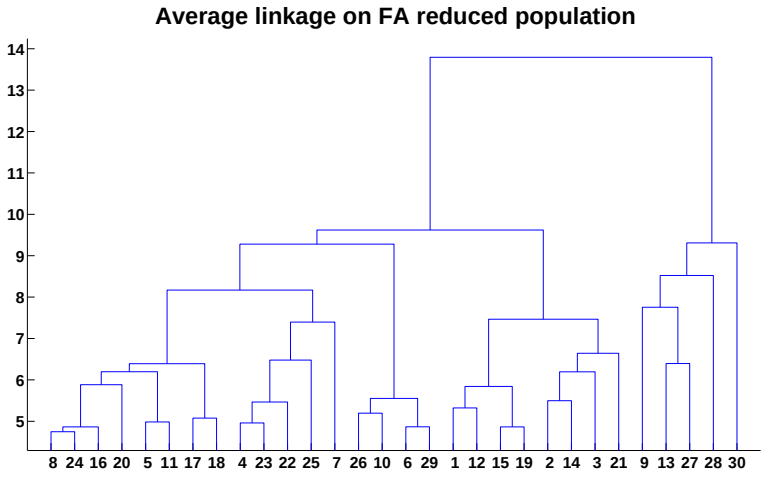

Figure 7.10: *Dendrogram for agglomerative clustering using average linkage on the population reduced with FA.*

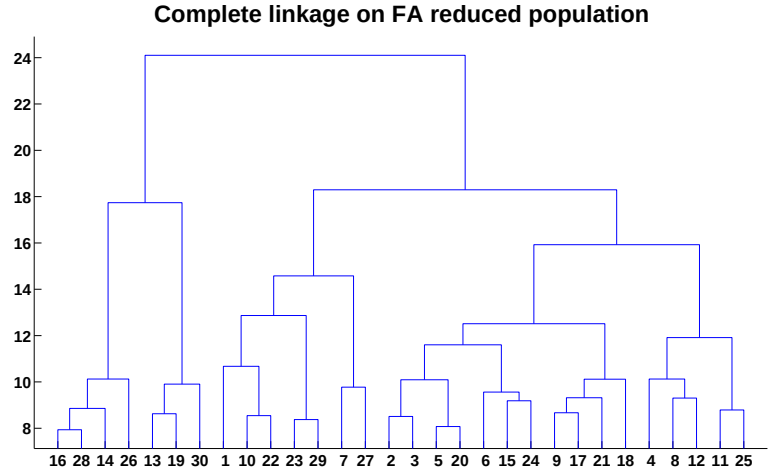

Figure 7.11: *Dendrogram for agglomerative clustering using complete linkage on the population reduced with FA.*

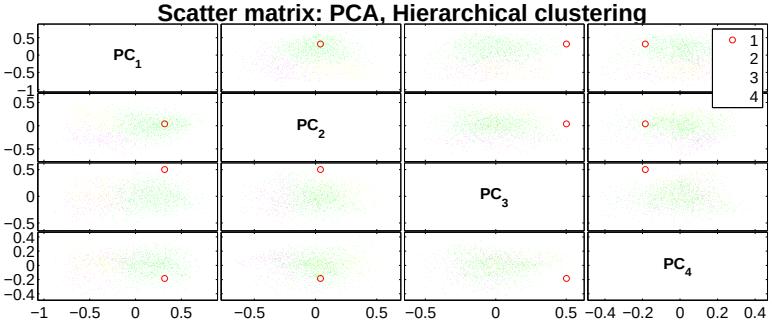

Figure 7.12: *Scatter plot of the population reduced with four principal components using Hierarchical Clustering. Each component is plotted against all other components. The different clusters are shown with different colors of the points representing trucks. One of the clusters, marked with a circle, only contains one truck. This could indicate that this truck is an outlier.*

7.2.4 Comparison of *k*-means and Hierarchical Clustering

One of the largest differences between *k*-means and Hierarchical Clustering is that *k*-means requires an initialization of the cluster centroids. The algorithm used in MATLAB did a first randomized guess of the cluster centers which therefore led to a slightly changing result between different runs, producing somewhat unstable results.

A comparison of the clusters found with *k*-means and MATLAB can be seen in Table 7.7. Here the percentage of the trucks in the clusters found with Hierarchical Clustering which also exist in the clusters found with Hierarchical Clustering are compared. A great part of the trucks in cluster 2, 3, 4 and 5 found with *k*-means seems to exist also in the second cluster found with Hierarchical Clustering, whereas cluster 6 found with *k*-means is very similar to the third cluster found with Hierarchical Clustering. A great part of the trucks in the first cluster found with *k*-means also exists in the fourth cluster found with Hierarchical Clustering. Altogether this indicates that the clusters found with *k*-means and Hierarchical Clustering are quite similar. Since fewer clusters were used for the hierarchical implementation, some of the *k*-means clusters are actually grouped together.

Table 7.7: *The percentages of the k-means clusters that also exist in hierarchical clusters.*

(%)	**Hier 1**	**Hier 2**	**Hier 3**	**Hier 4**
***k*-means 1**	0	4.28	7.24	**88.5**
***k*-means 2**	0	**73.4**	2.63	24.0
***k*-means 3**	0.03	**86.8**	12.9	0
***k*-means 4**	0	**100**	0	0
***k*-means 5**	0	**97.6**	0	2.39
***k*-means 6**	0	19.1	**71.3**	9.69

8

Model estimation

In this chapter the clusters are explored further to find out more about what affects fuel consumption for the trucks of interest. This is initially done by using a simple Least Squares method. The result from Least Squares is then compared to the outcome from using shrinkage methods such as the Lasso and Elastic Net.

The goal is to find out what affects fuel consumption for the clusters of similar trucks found in Chapter 7. Due to time restrictions, only the the six clusters found with k-means Clustering reduced to four dimensions with PCA are used, see Section 7.2.2. The results from the Hierarchical Clustering and the k-means Clustering reduced with FA are not evaluated further.

This analysis is done by estimating a regression model [Hastie et al., 2008, James et al., 2013] of fuel consumption, $\mathbf{y}$, on the form

$$\mathbf{y} = \mathbf{X}\beta + \epsilon, \tag{8.1}$$

where $\mathbf{X}$ is an $N \times p$ matrix containing an intercept and the regressors believed to influence fuel consumption, $\mathbf{y}$ an N-dimensional vector of the fuel consumption and β a p-dimensional vector of the coefficients. Here, p is the number of regressors believed to explain fuel consumption and N the number of trucks. The coefficients β can give information about how much and if a regressor influences the fuel consumption. The error term ϵ is a $N \times 1$-vector independent of $\mathbf{X}$. The elements of ϵ are assumed to be uncorrelated and identically distributed.

Common model estimation methods are the least squares method, LS, principal component regression, PCR, and shrinkage methods. PCR is simply LS with principal components as regressors instead of the original variables [Montgomery et al.,

2006]. Shrinkage, also known as regularization, shrinks the estimated coefficients towards zero which has the effect of reducing variance [James et al., 2013]. Common shrinkage methods are the Lasso, Ridge Regression and Elastic Net. The *Lasso* does a kind of variable selection, similarly to PCR, by estimating some of the coefficients to exactly zero. *Variable selection* can thus remove some of the irrelevant variables from our regression model. *Ridge Regression* shrinks all of the the coefficient estimates. *Elastic Net* is a combination of both Lasso and Ridge Regression and is therefore able to produce both smaller coefficients and variable selection. Since for this thesis it is interesting to find a parsimonious model, Ridge Regression is left out as it does not lead to a sparse solution. This property will however be covered by Elastic Net as it can include Ridge Regression depending on the choice of tuning parameters.

Initially one of the simplest of the earlier mentioned estimation methods, PCR, is used on all six clusters. Here a model of fuel consumption, $\mathbf{y}$, is found for each of the clusters using the principal components, representing truck usage, as input. More about this in Section 8.2. To see how the shrinkage methods handle the problem, a comparison of Lasso and Elastic Net is done for one of the clusters. Here the non-reduced population containing the original 43 variables, together with trucks from this particular cluster is used. The idea is that the two shrinkage methods will perform the variable selection PCA previously did, see Section 8.3. The remaining variables indicate what affects fuel consumption. At the end, a comparison of the performance of all methods is made.

8.1 Possible problems

A problem that might occur is *multicollinearity*, which means that there is a near-linear dependence among the regressors. This can result in too large LS coefficient estimates, $\hat{\beta}_i$, in absolute value and a large variance between the regressors [Montgomery et al., 2006]. Furthermore, multicollinearity makes the individual roles of the regressors in the model harder to interpret. Since some of the parameters included are basically the same parameter but logged in different units, such as how much the pedal has been used in both time and distance, it is likely that they are also linearly dependent. However, a benefit when using PCR is that the regressors, the principal components, then are orthogonal, thus linearly independent, and multicollinearity can be avoided [Montgomery et al., 2006].

When using the non-orthogonal original variables as input, Ridge Regression is probably a better choice than ordinary LS since it handles multicollinearity. In Ridge Regression some bias is introduced in the estimators, with the shrinkage parameter, which can lead to a smaller variance. If there is multicollinearity in the variables Elastic Net will most likely have a larger part of Ridge Regression (α closer to one, see (8.16)) to compensate for this.

8.2 Principal Component Regression

PCR is very similar to LS, the main difference is that the principal components (found using PCA, see Section 7.1.1) are used as regressors in the model. The coefficients β are selected such that the *Residual Sum of Squares* (RSS),

$$\mathrm{RSS}(\beta) = (\mathbf{y} - \mathbf{X}\beta)^T(\mathbf{y} - \mathbf{X}\beta) \tag{8.2}$$

is minimized. Here $\mathbf{X}$ is a $N \times p$ matrix of the p principal components.

In our case there are four principal components and around 200 to 300 trucks in each cluster. Since the number of trucks N is larger than p, the minimal solution to (8.2) can then be found by differentiating with respect to β, giving the unique solution

$$\hat{\beta} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y} \tag{8.3}$$

where $\hat{\beta}$ denotes the estimated coefficients [Hastie et al., 2008].

The result of the modeling using PCR can be seen in Appendix A. Before starting to interpret the models it is important to take a look at the residuals [Hair, Jr. et al., 2006] to assure the previously stated model assumptions hold. Since we want to use statistical tests such as the F-test to evaluate our model we also need normally distributed error terms. In total three assumptions need to be fulfilled:

1. Constant variance of the error terms
2. Independence of the error terms
3. Normally distributed error terms

8.2.1 Validation of model assumptions

Constant variance or *homoscedasticity* can be identified by plotting the residuals against the predicted output, see Figure 8.1, and examining the pattern shown. If the pattern of the plot is triangular shaped (increasing or decreasing variance) the error terms are *heteroscedastic*. If the plot looks like an evenly distributed cloud the error terms can be concluded to have a constant variance. This seems to be the case for all six clusters.

By studying the lag plots of each cluster in Figure 8.2 conclusions can be drawn about independence. In the lag plots each error term ϵ_i is plotted against the previous error term ϵ_{i-1} and if any patterns are seen this indicates that the error terms are correlated. The lag plots seems to show randomly distributed clouds for all clusters and the independence assumption therefore holds.

Figure 8.3 shows Quantile-Quantile plots (Q-Q-plots) of the six clusters. A Q-Q-plot compares two probability distributions by plotting their quantiles against each other. The Q-Q-plots in Figure 8.3 have the residual sample quantiles on the vertical axis and a standard normal distribution on the horizontal axis. The data is normally distributed if it follows the dashed line. Cluster 1, 2 and 6 have obvious normally distributed residuals. Clusters 3, 4 and 5 have heavy tails indi-

cating they are t-distributed. However, according to the Central Limit Theorem (CLT) these residuals will be approximately normal since the error terms are independent and uniformly distributed.

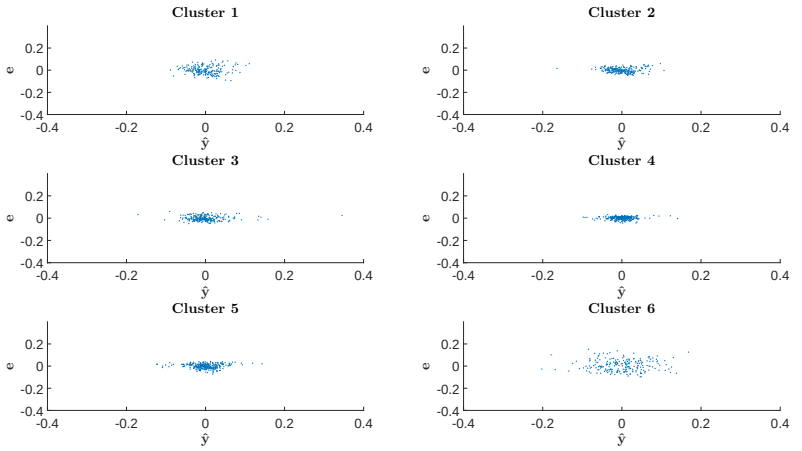

Figure 8.1: *This plot shows the residuals against predicted values using PCR for the six clusters.*

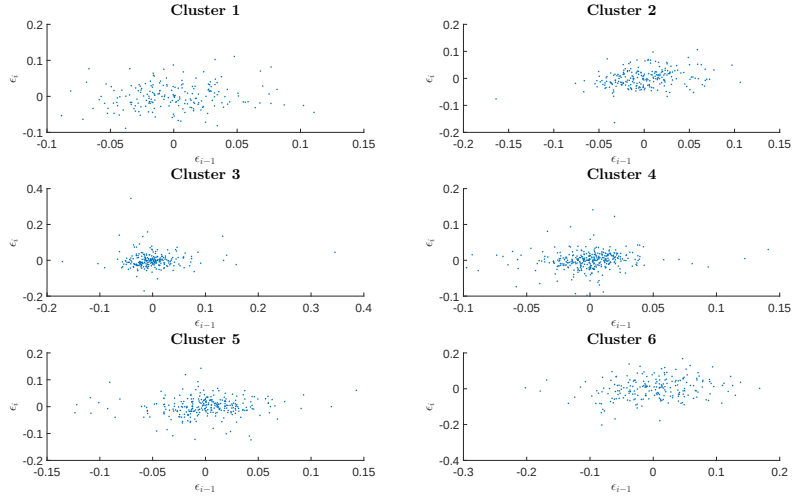

Figure 8.2: *In the lag plots each error term ϵ_i is plotted against the previous error term ϵ_{i-1} to indicate correlation between error terms.*

Overall the three model assumptions (constant variance, independence and normal distribution of error terms) can be stated to hold for all six clusters. This

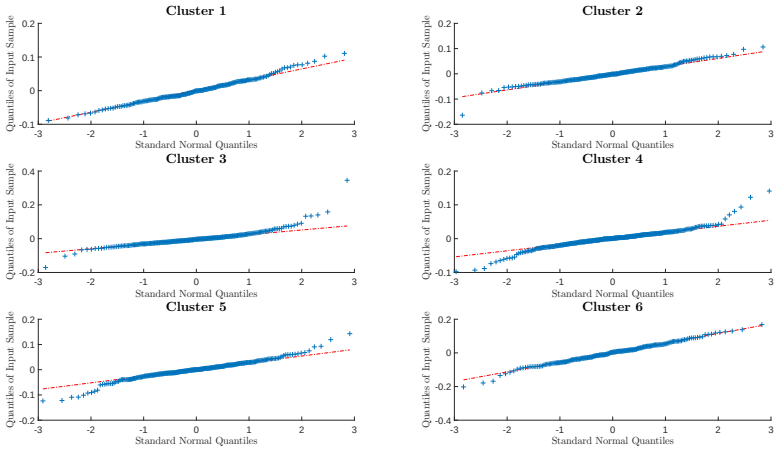

Figure 8.3: *Q-Q plots of the residuals for the six clusters. Cluster 1, 2 and 6 have normal distributed residuals. Cluster 3, 4 and 5 have heavy tails.*

means that the clusters can be evaluated further using statistical quantities.

8.2.2 Inference

The first step is to decide whether there is a relationship between the response variable fuel consumption and the regressors, i.e. the principal components. This is done using a hypothesis test

$$\begin{aligned} &\mathrm{H}_0 : \text{There is no relationship between } \mathbf{y} \text{ and } \mathbf{X},\ \beta_1 = \beta_2 = \cdots = \beta_k = 0, \\ &\mathrm{H}_A : \text{At least one of the input variables in } \mathbf{X} \text{ has a relationship with } \mathbf{y},\ \beta_i \neq 0, \end{aligned} \tag{8.4}$$

where H_0 is the null hypothesis. The test is done using the F-statistic [Montgomery et al., 2006, James et al., 2013] defined as

$$F = \frac{SS_R/p}{SS_{Res}/(n-p-1)} \tag{8.5}$$

where SS_R denotes the sum of squares due to regression, according to

$$SS_R = \hat{\beta}^T \mathbf{X}^T \mathbf{y} - \frac{\left(\sum_{i=1}^{n} y_i\right)^2}{n} \tag{8.6}$$

and SS_{Res} the residual sum of squares, according to

$$SS_{Res} = \mathbf{y}^T \mathbf{y} - \hat{\beta}^T \mathbf{X}^T \mathbf{y}. \tag{8.7}$$

The numerator of the F-statistic, see (8.5), represents the variance explained by the regression model and the denominator the unexplained variance. If the ratio of the explained variance to the unexplained variance is high one can conclude

that there is some relationship between the fuel consumption and the principal components. On the other hand, a value of the F-statistic that is close to 1 means that the size of the unexplained variance is very close to the size of the explained variance. The null hypothesis can then not be rejected, hence a relationship between fuel consumption and the principal components cannot be concluded.

The values of the F-statistic for the six clusters are summarized in Table 8.1. The size of the F-statistic indicates all six clusters have at least one principal component related to fuel consumption. It is now of interest to investigate whether

Table 8.1: *Table showing the value of the F-statistics for the regression models using PCR for the six clusters.*

Cluster	**1**	**2**	**3**	**4**	**5**	**6**
F-value	57.9	29.3	18.6	32.2	42.9	51.4

the regressors are statistically significant, i.e. which principal components that can be used to explain the fuel consumption in each cluster. This is once again done using a hypothesis test, but now one test is stated for each regressor and the *t-statistics* is used to decide whether to reject the null hypothesis or not.

The hypothesis test is now formed as follows

$$\begin{aligned} \mathrm{H}_0 &: \text{The estimated coefficient is zero, } \hat{\beta}_i = 0, \\ \mathrm{H}_A &: \text{The estimated coefficient is nonzero, } \hat{\beta}_i \neq 0, \end{aligned} \tag{8.8}$$

where H_0 is the null hypothesis. If the estimated coefficient is zero, the regressor is not statistically significant.

To be able to reject the null hypothesis, one needs to assure $\hat{\beta}_i$ is sufficiently far from zero. This is done by first deciding on at what level of significance we want to reject the null hypothesis. A value of 0.05 (5%), which is a typical value, is chosen. The t-statistics is calculated as follows

$$t_{\hat{\beta}_i} = \frac{\hat{\beta}_i}{\mathrm{SE}(\hat{\beta}_i)} \tag{8.9}$$

where the standard error for the coefficient $\hat{\beta}_i$ is defined as

$$\mathrm{SE}(\hat{\beta}_i) = \sqrt{\hat{\sigma}^2 C_{ii}} \tag{8.10}$$

where C_{ii} is the the diagonal element of $(\mathbf{X}^T\mathbf{X})^{-1}$ corresponding to $\hat{\beta}_i$.

The associated p-value is the probability that we can observe a result as extreme or more extreme by chance, given that the null hypothesis is true. The larger the p-value is the more likely it is that what we observe is given by chance. The p-value is typically given from tables but can also be calculated with e.g. MATLAB as in this case. To be able to reject the null hypothesis, this p-value now has to be smaller or equal to 0.05. A small p-value is typically associated with a large t-statistics. Thus, a small standard error compared to the coefficient estimate gives

a larger t-statistics and also a small p-value.

The regressors with a large t-statistics and a p-value smaller than 0.05 are marked in bold in Table A.1 and Table A.2. A summary of the estimated coefficients that are statistically significant can be seen in Table 8.2.

Table 8.2: *Table showing estimated coeffcients that are statistically significant for the six clusters.*

	$\hat{\beta}_0$	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_3$	$\hat{\beta}_4$
Cluster 1		X	X		X
Cluster 2		X	X		X
Cluster 3		X	X	X	X
Cluster 4	X	X	X	X	
Cluster 5	X	X	X		
Cluster 6	X	X	X	X	

Three of the clusters have one estimated coefficient that is not statistically significant and the remaining have two estimated coefficients that are not statistically significant. This means that it is not possible to interpret these coefficients since it cannot be concluded that these coefficients are neither zero nor not zero. This is of course problematic when the objective is to analyze the influence on fuel consumption from the principal components.

To further assess the performance of the models, the measures R^2 and adjusted R^2 are of interest. For linear models, they both indicate how well the data fits the model, i.e. the proportion of variance the model explains. They are defined according to

$$R^2 = 1 - \frac{\sum_i (\hat{y}_i - y_i)^2}{\sum_i (y_i - \mu_i)^2} \tag{8.11}$$

$$\text{adjusted } R^2 = 1 - \frac{\sum_i (\hat{y}_i - y_i)^2 / \text{df}_e}{\sum_i (y_i - \mu_i)^2 / \text{df}_t} \tag{8.12}$$

where $\text{df}_t = n - 1$ are the degrees of freedom of the estimate of the population variance of the dependent variable, $\text{df}_e = n - p - 1$ the degrees of freedom of the estimate of the underlying population error variance and μ_i is the output mean. A larger value of the two coefficients is preferable. The R^2 and the adjustedR^2 for the clusters are all in the range of 0.166 and 0.436 which indicates that the explained variance of the models is low.

Overall the size of the statistically significant coefficients does not vary a lot in each cluster. Taking Cluster 2 as an example from Table A.1, $\hat{\beta}_2$ is the largest coefficient and it is approximately twice as large as the two other statistically significant coefficients $\hat{\beta}_1$ and $\hat{\beta}_4$. This would then indicate that for this particular cluster PC_2 is most important for fuel consumption. However, due to the fact that there are so many statistically insignificant coefficients, the overall performance of this model is hard to evaluate. An interesting approach is therefore to

use shrinkage methods to solve the same problem. More about this in the next section.

8.3 Shrinkage methods

Used as input to this section are the trucks included in Cluster 1, found with k-means clustering on the PCA reduced population. However, the non-reduced data is used for these trucks. The idea is to evaluate the performance of shrinkage methods compared to PCR and it is therefore considered sufficient to only look at one of the clusters.

In contrast to PCR, the non-orthogonal original variables are used as regressors for the shrinkage methods. This results in a method replacing both the previously used dimensionality reduction and the modeling used to extract information about what affects fuel consumption for our clusters of trucks. The shrinkage methods Ridge Regression, the Lasso and Elastic Net have an additional penalty in the RSS compared to the RSS used for both LS and PCR (compare (8.13), (8.15) and (8.16) to (8.2)). There are however some differences between the three mentioned shrinkage methods.

Ridge Regression penalizes the size of the coefficients β with an L^2 penalty. The coefficients are estimated, in analogy with PCR, by minimizing

$$\mathrm{RSS}_{\mathrm{ridge}}(\beta) = (\mathbf{y} - \mathbf{X}\beta)^T(\mathbf{y} - \mathbf{X}\beta) + \lambda \sum_{j=1}^{p} \beta_j^2 \tag{8.13}$$

Lasso where $\lambda \geq 0$ is a tuning parameter. This gives the solution

$$\hat{\beta}_{\mathrm{ridge}} = (\mathbf{X}^T\mathbf{X} + \lambda\mathbf{I})^{-1}\mathbf{X}^T\mathbf{y} \tag{8.14}$$

The magnitude of the coefficients estimates $\hat{\beta}$ depends on λ. A larger value of λ results in a greater shrinkage and smaller coefficient estimates.

For the Lasso a L^1 penalty is instead added to the RSS of least squares according to

$$\mathrm{RSS}_{\mathrm{Lasso}}(\beta) = (\mathbf{y} - \mathbf{X}\beta)^T(\mathbf{y} - \mathbf{X}\beta) + \lambda \sum_{j=1}^{p} |\beta_j|. \tag{8.15}$$

Here a large λ forces some of the coefficient estimates to be equal to zero.

Elastic Net combines the L^1 and L^2 penalties according to

$$\mathrm{RSS}_{\mathrm{Elastic\ Net}}(\beta) = (\mathbf{y} - \mathbf{X}\beta)^T(\mathbf{y} - \mathbf{X}\beta) + \lambda\Big((1-\alpha)\sum_{j=1}^{p} |\beta_j| + \alpha \sum_{j=1}^{p} \beta_j^2\Big), \tag{8.16}$$

where $\alpha \in [0, 1]$ is a second tuning parameter [Zou and Hastie, 2005]. A solution equivalent of the Lasso is given by $\alpha = 0$ whereas $\alpha = 1$ gives a solution similar to Ridge Regression. The L^2-part of the penalty will encourage the correlated

variables to be averaged, handling multicollinearity as discussed in Section 8.1, while the L^1-part of the penalty will result in the desired sparse solution.

8.3.1 Choice of tuning parameters

The tuning parameters α and λ can be chosen using for example one of the two cross-validation (CV) methods *leave-p-out* and *k-fold* CV. A grid of tuning parameters is chosen and then the best value is found using CV. For Lasso and Ridge Regression this grid is a one dimensional vector whereas for Elastic Net a 2D-grid is needed to find the best value of both tuning parameters.

k-fold CV divides the data into k parts of nearly equal size, were one of the k parts is kept as a validation set and the rest used for estimation [James et al., 2013]. This is then repeated k times, using a different part as validation set each time. For each iteration the mean squared error is computed and the CV error

$$\mathrm{CV}_k = \frac{1}{k}\sum_{i=1}^{k} \mathrm{MSE}_i \tag{8.17}$$

is calculated as the sum of the Mean Squared Error, MSE, for all k. By repeating this process for each λ_i, an appropriate value of λ can be chosen as the one giving the smallest CV error. The final model is then calculated with the chosen λ. A typical value for k is 5 or 10 and the choice depends of the computational cost. The ideal case can seem to be to set k equal to the number of data points n minus one, that is to validate each data point against all others. However, as discussed by James et al. [2013], this method often gives less accurate MSE estimates due to a high variance. The variance comes from the the different validation parts being highly correlated since they contain almost the same data. With $k < n - 1$, bias is instead introduced. The choice of k therefore needs to be a trade-off between bias and variance and empirically shown, k equal to 5 or 10 gives neither high variance nor bias.

Leave-p-out CV is very similar to k-fold CV, but instead of dividing the data into k equal parts, p data points are typically randomly chosen for validation. The process is then repeated a fixed number of times. How many times the process should be repeated depends on the size of the data set and p. The goal is to repeat this as many times as it takes to cut the data in all ways possible with p data points excluded. In this case this leads to a noisy result as the process is not repeated sufficiently many times. With k-fold CV a better result is achieved with fewer computations, which leads to choosing the latter method.

8.3.2 Implementation

The technique described above can be used when finding the tuning parameters for both the Lasso and Elastic Net. Two algorithms that can be used to solve the Lasso and Elastic Net problem respectively are LARS and LARS-EN, the latter proposed in Zou and Hastie [2005]. In this thesis a simpler approach is instead used. A model estimation for both problem formulations is implemented using the MATLAB based modeling system **CVX** [Grant and Boyd, 2014, 2008] which

solves convex optimization problems. The Lasso is implemented according to Algorithm 5 with k equal to 10. The grid of tuning parameters is chosen as 100

Algorithm 5 Lasso using k-fold CV

1. Choose a 1d-grid of values for the tuning parameters: $\lambda = \lambda_1, \lambda_2, .., \lambda_n$.
2. Choose the number of folds k.
3. Iterate over the n values of λ and k folds and calculate the MSE for each k and then the CV error for each λ.
4. Evaluate the CV errors and choose the λ_{opt} minimizing the CV error.
5. Estimate model according to 8.15 with chosen λ_{opt}.

Table 8.3: *A table showing the four parameters correspondig to the regressors that had to be removed from the shrinkage methods.*

Parameter	Notation
PTO distance percentage	d_{pto}
Road gradient -10% to -8% percentage	top_2
Road gradient -7% to -5% percentage	top_3
Road gradient 11% to 20% percentage	top_9

Algorithm 6 Elastic Net using k-fold CV

1. Choose a 2d-grid of values for the tuning parameters: $\lambda = \lambda_1, \lambda_2, .., \lambda_p$ and $\alpha = \alpha_1, \alpha_2, .., \alpha_j$.
2. Choose the number of folds k
3. Iterate over a subset of the p λ-values, the j α-values and k folds and calculate the MSE for each k and then the CV error for each α.
4. Evaluate the CV errors and choose the α_{opt} minimizing the CV error.
5. Iterate over the p λ-values and k folds and calculate the MSE for each k and then the CV error for each λ.
6. Estimate model according to 8.16 with chosen λ_{opt}

logarithmically distributed values $\lambda \in [0.000001, 1]$. The model is thus estimated 10 times for each one of the 100 tuning parameters. How the size of the estimated coefficients vary for different values of the tuning parameter can be seen in Figure 8.4 as a trace plot. One can conclude that a larger value of λ gives a more sparse solution as some of the estimated coefficients approach zero. The optimal size of the tuning parameter is given by evaluating the CV error for each tuning parameter. At first the original 43 parameters were used as regressors in the model, this however resulted in four estimated coefficients being up to a thousand times larger than the rest of the estimated coefficients. When examining these four regressors, seen in Table 8.3, a bit closer they contained data up to a hundred times smaller than the data of the remaining parameters and also many zeros. When these four regressors where removed from the model, the much more stable result seen in Figure 8.4 was found.

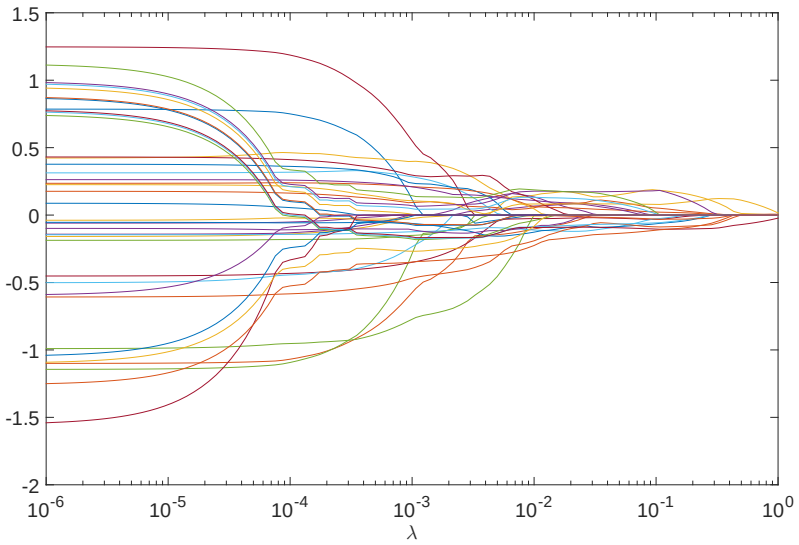

Figure 8.4: *This figure shows the trace plot for the Lasso. On the vertical axis is the size of the estimated coefficient and on the horizontal axis the tuning parameter. The different colors represent the different estimated coefficients. As the tuning parameter λ increases, more and more estimated coefficients approach zero.*

For Elastic Net, 10-fold CV is instead repeated twice, once for each tuning parameter, according to Algorithm 6. To keep down the computational cost, the α_{opt} is first found by iterating over a smaller λ grid, containing only some values of interest in a given interval. By choosing 10 logarithmically distributed values of $\alpha \in [0.01, 1]$ and 5 logarithmically distributed values $\lambda \in [0.00001, 1]$ we get a result according to Figure 8.5. The largest value, $\alpha_{10} = 1$ seems to give the best result, leading to smaller CV errors. As this would mean pure Ridge Regression and we want to keep some of the variable selective property of the Lasso, we choose the second best value which is $\alpha_9 = 0.5995$.

Having chosen $\alpha_{opt} = \alpha_9$, 10-fold cross-validation is done for 100 logarithmically distributed values $\lambda \in [0.000001, 1]$, the same grid as used in the Lasso case. The result can be seen in Figure 8.6.

Comparing this trace plot with the trace plot of the Lasso, it seems as if a more sparse solution is given for smaller tuning parameter values than in the Elastic Net case. This is reasonable as the Elastic Net, besides the variable selective property also found in the Lasso, also has an averaging property from the L^2 penalty.

The CV error plotted against the tuning parameter λ for both Elastic Net and the Lasso can be seen in Figure 8.7. The value of the tuning parameter corresponding to the smallest CV errors is now used to estimate the two models. The best suit-

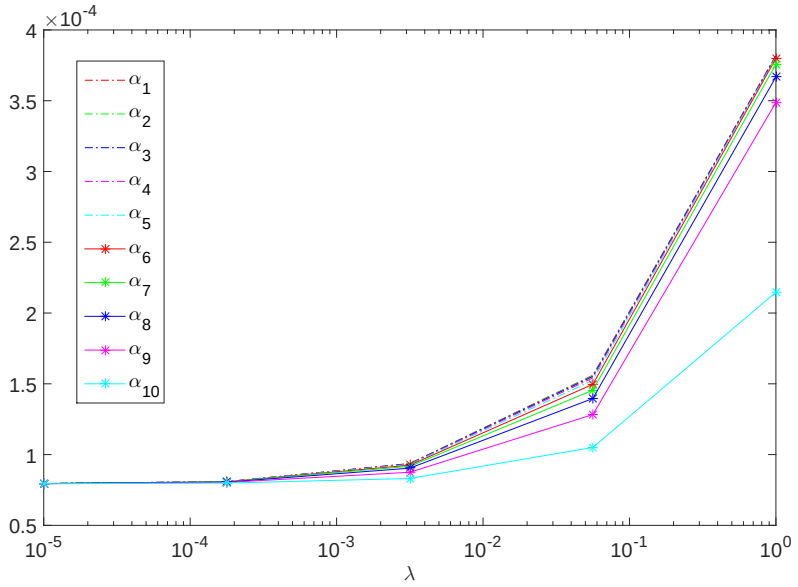

Figure 8.5: *This figure shows the CV error towards λ. The curves with various styling indicate how the CV error curve varies for different α. The goal is to choose an α that minimizes the CV error.*

able λ for the two model estimation techniques can be seen in Table 8.4 together with the corresponding MSE. The two values are very close to each other. The R^2, indicating how well the data fits the model also gives a similar result for the two techniques. If compared to the same cluster where PCR instead is used, see Table A, the R^2 is as low as 0.436. Both Lasso and Elastic Net performs better than PCR.

Table 8.4: *Table comparing the MSE and the R^2 for Lasso and Elastic Net.*

	Lasso	Elastic Net
MSE_{min}	$3.4549 * 10^{-5}$	$3.3774 * 10^{-5}$
R^2	0.7219	0.7282

A summary of the corresponding parameters to the estimated coefficients that remains in Lasso and Elastic Net can be seen in Table 8.5. As stated earlier, the parameters in Table 8.3 are not included in the estimation and no estimated coefficient is therefore shown for them. Furthermore, estimated coefficient values smaller than 0.08 are considered as zero coefficients. These coefficients are also shown as blanks in Table 8.5.

A more detailed analysis of the parameters leading to large estimated coefficients for the Lasso reveals f_{idle}, t_{top1} and d_{top1} as the most important parameters for fuel consumption. In the Elastic Net case the most important parameters are in-

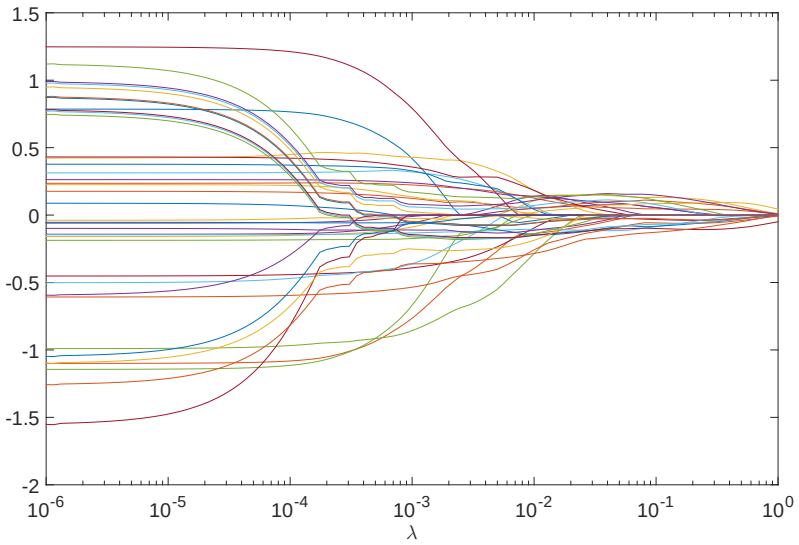

Figure 8.6: *This figure shows the trace plot for the Elastic Net. On the vertical axis is the size of the estimated coefficient and on the horizontal axis the tuning parameter. The different colors represent the different estimated coefficients. As the tuning parameter λ increases, more and more estimated coefficients approach zero.*

stead t_{coast}, d_{coast}, t_{brake}, f_{pto} and f_{idle}. If this result is compared to what is seen for Cluster 1 in the PCR case, no obvious resemblances can be seen. In the PCR case, PC_1 is the most important component being a property mainly described by f_{top1}, t_{top1}, GCW_4 and GCW_5. Event though some of the parameters appear in all methods, the similarity of the three methods is vague. No reliable conclusions about what affects fuel consumption in Cluster 1 are therefore reasonable to draw.

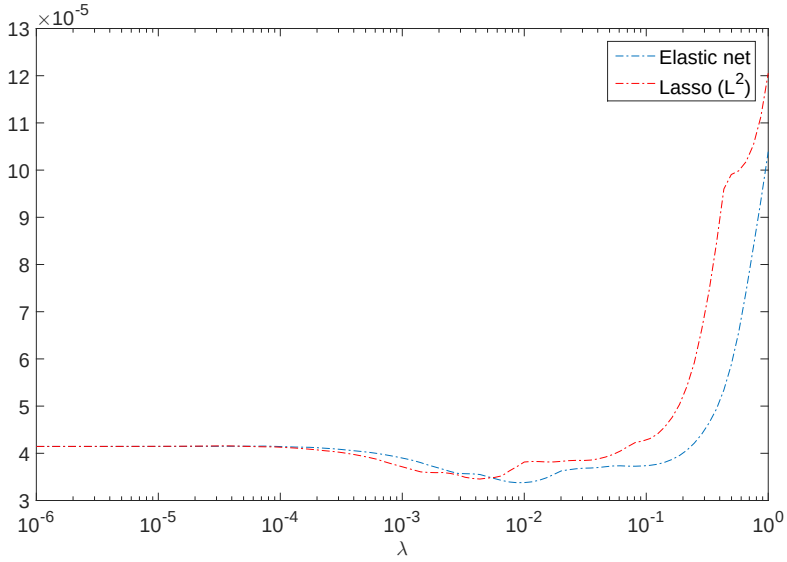

Figure 8.7: *Seen here is the CV error versus the tuning parameter λ for the two shrinkage methods Elastic Net and Lasso. The smallest CV error indicates the best tuning parameter to use in the estimated models. For both model estimation techniques this value lies between 0.001 and 0.01.*

Table 8.5: *Table showing the estimated coefficients for the included parameter regressors included in the two model estimation techniques Lasso and Elastic Net.*

	Lasso	Elastic Net
d_{brake}		0.5213
d_{coast}	−0.1493	−0.9745
d_{econ}		
d_{pedal}		
d_{pto}		
d_{cruise}		0.2846
d_{drive}		0.2299
d_{top1}	0.2876	−0.1469
d_{top2}		
t_{brake}		−0.7752
t_{coast}		1.1138
t_{econ}	0.1313	0.3318
t_{pedal}		0.2268
t_{pto}	0.1389	0.1506
t_{cruise}		0.1495
t_{drive}	0.1274	−0.1156
t_{idle}	0.1603	0.3318
t_{top1}	−0.3459	0.2740
f_{coast}		
f_{econ}		0.2804
f_{pedal}	−0.0931	−0.2453
f_{pto}		−0.8926
f_{cruise}	−0.0853	−0.4638
f_{drive}	0.2017	0.3910
f_{idle}	−0.5133	−0.8926
f_{top1}	−0.1543	
GCW_1	−0.1157	−0.1558
GCW_2	−0.1386	−0.1581
GCW_3	−0.1525	−0.1627
GCW_4	−0.1507	−0.1672
GCW_5		
GCW_6		
GCW_7		
GCW_8	0.1021	
GCW_9	0.1461	0.1252
top_1		
top_2		
top_3		
top_4		−0.4742
top_5		
top_6	−0.3002	−0.3541
top_7	−0.2140	−0.2505
top_8		0.1533
top_9		

9 Analysis

In this chapter, the findings from the data analysis described in Chapters 6, 7 and 8 are analyzed based on the theoretical framework in Chapter 3 and the research in Chapter 4. It is discussed how these findings can highlight how a company needs to handle the data in order to be able to use value adding information to increase customer satisfaction of their products and services. The structure of this chapter will follow the structure of the big-data plan described in Section 4.1.4.

Figure 9.1 shows how the different parts of the theory and the data analysis of this thesis are connected as a whole. The following sections will describe how each part of the big-data plan is connected to the cornerstones of Total Quality Management and the data analysis. This is used to analyze the findings of the case study to understand how a company needs to handle the data in order to use it to increase customer satisfaction of their products and services.

9.1 Data

It has been established in Chapter 3 that in order for a company to create business value, it must create value for its customers and consequently make the customers satisfied. Therefore *focus on customers* is the center cornerstone of Total Quality Management. A company can create value for its customers by systematically trying to fulfill their needs when developing and manufacturing a product or a service [Bergman and Klefsjö, 2010]. In order to do this, the company needs to find out what the customers want and need, which involves finding out how they experience the product provided by the company [Bergman and Klefsjö, 2010, Deming, 1986]. To be able to focus on customers, systematic information

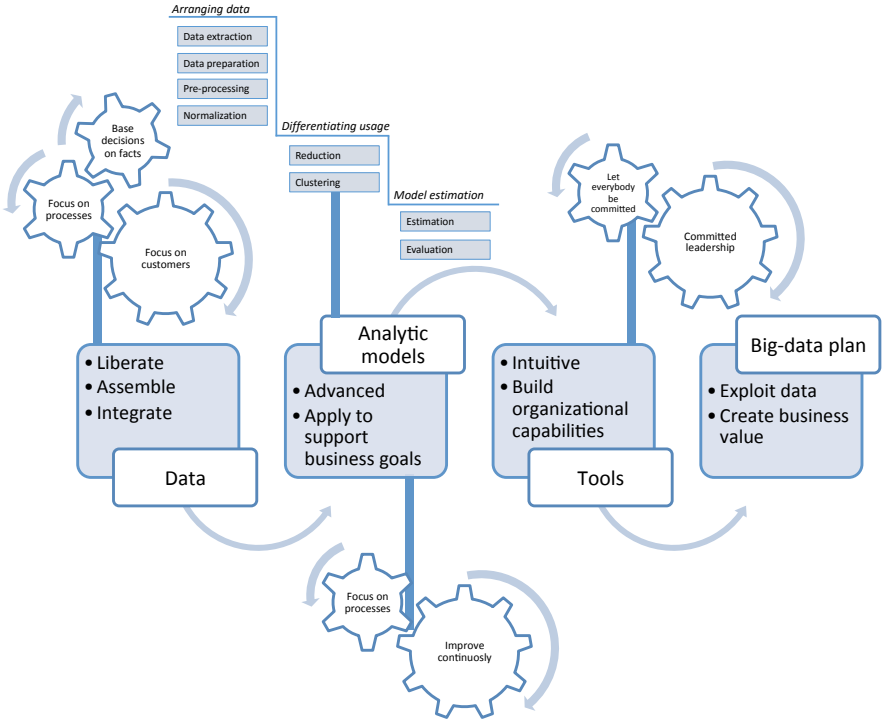

Figure 9.1: *Data, analytic models and tools are the three parts of a big-data plan and the different cornerstones of Total Quality Management are connected to different parts of the plan. The data analysis method of this thesis is naturally connected to the analytic models part.*

about their needs, requirements, reactions and opinions is required [Bergman and Klefsjö, 2010]. This information can then be fed back into the processes of the company and into the design of the products in order to improve them [Bergman and Klefsjö, 2010, Deming, 1986]. It is also important to fulfill the future needs of the customers by innovation [Deming, 1986].

By the advancements in technology it is now possible for organizations to acquire large amounts of information about their customers and how they use their products and services, so called *big data*. Using this data, improvements can be made in product quality and better meeting customer needs through more precisely targeted products and effective promotion [Manyika et al., 2011]. Important insights from the customer data can be used, not only for enhancing existing products, but also to help develop specifications for new models and variants to systematically design new products as well as enhance service offerings in order to fulfill the future needs of the customers [Manyika et al., 2011].

GTT has access to a large amount of data in the LVD database containing logged vehicle data described in Section 2.3. There is tremendous potential in this data. The data is so called real-world data, meaning that it is logged from vehicles used by customers and therefore shows actual customer usage. In addition, it contains data from all vehicles of the brands sold by Volvo Group including other business areas like Volvo Buses and Volvo Construction Equipment. However, this data is not used to its full potential. There are examples where sensor data describing actual customer usage have been used to create business value in the manufacturing industry [Manyika et al., 2011] and in the automotive industry [Bartram, 2013], so therefore it is considered as possible also for Volvo Group. For GTT particularly the data could for example be capitalized in product development in solving quality problems with customer complaints and in product improvements such as enhancing fuel consumption.

As addressed in Chapter 4, there are however a number of technical as well as organizational challenges associated with big data. To overcome these obstacles and make use of the big data, a company needs to create a big-data plan as described in Section 4.1.4 [Biesdorf et al., 2013]. The first part of this plan, *data*, is about setting up a strategy for assembling and integrating data, which is essential [Biesdorf et al., 2013]. Moreover, a system for data-gathering, recording and presentation is essential [Oakland, 2003]. Factual data of both numerical and verbal character is needed and an organization must gather and structure different kinds of information in order to *base decisions on facts* [Bergman and Klefsjö, 2010]. In order for a process to produce an output that meets the customer requirements it is necessary to define, monitor and control the inputs to the process [Oakland, 2003], i.e. it is also important to *focus on processes* [Bergman and Klefsjö, 2010].

In this case, one input to the product development process, and more particularly the analysis process, is the large amounts of logged vehicle data. In the data analysis presented in this thesis, 65 percent of the trucks had to be removed from the chosen population due to missing parameter values. Moreover, an additional 33 percent were classified as outliers and were therefore also removed. Finally, due to missing values and uninterpretable parameters, some of the targeted parameters to be included in the data analysis had to be excluded. In total, 77 percent of the original population was excluded. These findings indicate that there resides incompleteness, inconsistencies and unreliability in the data. It has been established that handling these issues of *veracity* is challenging [Jagadish et al., 2014]. In order for the incompleteness, inconsistencies and unreliabilities not to propagate through the data analysis and make it useless these measures of data removal were taken.

The reason for the immense incompleteness in the data could be the routines regarding the logging and the downloading of the data and the transfer into the database. As described in Section 2.3, the data is downloaded while the technician in the workshop has connected the TechTool to the truck. This means however that when the technician is finished with the service, the TechTool is disconnected from the truck regardless if the downloading was finished or not, which

explains missing values for some parameters. Moreover, considering the large amounts of outliers in the data, this may indicate some inaccuracy in the measurements, which causes unreliability. There were also indications of errors in the structure of the data such as incorrect axes for some vector parameters. Some unreliability may also originate from the fact that the data is on aggregated form with very low granularity. This is a way to handle the *volume* of the data. The *velocity* of the data is handled by only downloading the data when the truck is in the workshop for service. However, for closer follow up of the vehicles this downloading rate might be too slow in order to maximally utilize the commercial value of the data. For this kind of usage of the data telematics is a good alternative, for example as how it is used in the service Remote Diagnostics mentioned in Section 2.1.

One of the most time consuming steps when exploring the data in this thesis was the choice of population; both the choice of trucks and parameters to include. When extracting logged truck data from the database a set of search criteria had to be stated to assure that the desired truck population was extracted. Since patterns about the fuel consumption for various kinds of trucks was the goal, with the hypothesis that the data would be able to reveal such differences, a population heterogeneous when it comes to application of usage was what the thesis authors initially sought. Different applications of usage, such as long haul trucks and trucks used in city traffic, require different truck configurations when it comes to e.g. size and engine. However, a population of this kind requires information about the appropriateness of search criteria regarding their configuration, to be able to extract this data from the database. The information was difficult to access and therefore a more constraint population had to be chosen, in line with the knowledge and needs of the team where the thesis was conducted. If this limitation could have been overcome, this could have led to some obvious pattern which especially the clustering could have revealed. Since the data is on aggregated form over long periods of time, it is reasonable to assume that strong patterns are needed if the method is going to be able to reveal them.

Regarding the choice of parameters used to find patterns in the data, a drawback was that all desired parameters could not be included, according to the explanation given in Chapter 6. All parameters were not stored for all trucks (65 percent of the trucks had to be removed from the population due to this fact) in the database and many of them were difficult to interpret when it came to both physical quantity and their meaning, especially the parameters on vector format. The information concerning the meaning of the parameters had to be derived from a second system, which again complicated the selection for the thesis authors. An important aspect is that the stored parameters were not meant to be used for an analysis of this kind and to be able to take advantage of their full potential, knowledge about the functioning of the trucks is needed to better understand how the parameters interact. The selected parameters therefore mainly included parameters representing usage of different modes on the gearbox and the usage of gears, excluding engine related parameters and vehicle speed. To include information about how the engine had been used together with for example topography could

have given more interesting information connected to both truck usage and fuel consumption. Despite the special treatment of the parameters on vector format, they were a useful contribution as they gave more information than the single value parameters about each truck, which is also supported by the fact that they appear in both the principal components and factors, see Chapter 7 .

A big-data plan would enable for GTT to control the inputs to the data analysis process, i.e. to manage the encountered issues with the logged vehicle data, by stating a strategy for assembling the data. Having a plan can help setting up the right systems for data-gathering, recording and analysis in line with Oakland [2003] stating that the key to quality is for everyone in the organization to have well-defined customers, since this facilitates fulfilling the needs of the next link of the value creating customer-supplier chain that runs throughout the organization. This means that if it is clear what the systems are there to support and what goal this function has, this will facilitate setting up a system that will fulfill the needs of the next link in the process.

Also, the big-data plan should include a strategy for integrating the data residing already in different existing IT systems at GTT. In order to access the information needed to be able to use the logged vehicle data, a number of databases and tools were needed. In line with Manyika et al. [2011], to fully capitalize the data, GTT needs to have a free interchange of data among different functions and business units such as sales and marketing and product development. This big-data plan would therefore highlight a need for a reorganization of data architectures over time, among other things implementing data-governance standards that systematically maintain accuracy, in line with Biesdorf et al. [2013].

Some have stated that all processes should be measured and that all measurements should be recorded since if data is recorded and not used it will be abused [Oakland, 2003]. However, since one of the most pressing challenges of big data is its *volume* because of current storage problems, there are others who suggest that an important principle related to the analytical value of the data should be developed to decide which data shall be stored and which data shall be discarded [Chen et al., 2014].

9.2 Analytic models

In order to create customer value and consequently to generate business value, a company must work with its processes so that the outputs satisfy its customers while using as little resources as possible, i.e. *focus on processes*. The purpose of the people and their relationships, resources and tools of an organization is to support these processes [Bergman and Klefsjö, 2010]. Each process can be analyzed by examining its inputs and outputs [Oakland, 2003]. With statistical tools and models conclusions can be drawn from the process history about its future results, and to recover necessary information to improve the process [Bergman and Klefsjö, 2010]. To *improve continuously* is a way to improve quality, which means delivery of those features of the product that respond better to customer needs

[Juran, 2010]. Designing and continuously improving the quality of an organization's products and services creates high stakeholder and employee satisfaction and both will have an effect of the revenues of the company [Juran, 2010]. To improve continuously will also have an effect of the costs of the company since improved quality also means a lower cost of poor quality [Juran, 2010]. Moreover, in continuous improvement, the task is to improve upon the standardized best practices of the organization by continuously using tools for determining the root cause of inefficiencies [Liker and Meier, 2006].

All of this can be done with analytic models. Advanced analytic models are needed to enable data-driven optimization or predictions [Biesdorf et al., 2013]. However, it is important not to include too many variables in the models since this will create complexity while making the models harder to apply and maintain [Biesdorf et al., 2013] and therefore, data analysis should be carried out by means of some basic systematic tools [Oakland, 2003]. To handle the balance between using advanced data analysis in order to extract value adding information from the big data with many technical issues without having to much complexity, the approach for the data analysis in this case study was to use the different techniques to simplify as much as possible. First of all the data was arranged and processed in order to handle the *volume*, *variety* and the *veracity* of the data. The *volume* and the *variety* of the data were also handled in the dimensionality reduction. The dimensionality reduction was used to find underlying structure in the data while at the same time reduce dimensions. This was done to reduce the complexity of the model in the model estimation without losing any of the information in the data by selecting only a few parameters. It is desirable to be able to describe the population with as few components as possible to get a sparse model where it is easy to distinguish the impact of the different components.

Moreover, using analytics and analytic models builds a basis for improving decisions over time [Davenport et al., 2010]. A big-data plan must identify where models will create additional business value, who will need to use them and how to avoid inconsistencies and unnecessary proliferation as models are scaled up across the organization [Biesdorf et al., 2013].

This case study included a direct observation of the data as the authors wished to use it in a data analysis for revealing patterns about what affects fuel consumption for various kinds of trucks. Since this case study was conducted at GTT the data analysis was intended primarily to support the product development process by revealing information about the actual fuel consumption of different trucks since this was considered as a large factor in what affects the customer satisfaction, and therefore creating business value this way. Furthermore, Volvo Group Trucks has several customers concerning this issue, not only the haulage contractors buying the trucks, but also the drivers. There are also other external customers that are affected by the usage of the trucks developed by GTT, which can be extended to society at large since everyone is affected by the effects of exhaust emissions on the environment. Therefore, this kind of analysis was of great interest for GTT to be able to perform.

Since the data analysis was intended to support product development, the product development engineers especially working with fuel consumption, but also other product development engineers at GTT were the target group of users of the analysis. Product development is a very wide area and includes both the "hardware" of the truck such as the engine and powertrain as well as the the chassis and the cabin which can be improved with respect to fuel consumption by reducing the weight of the truck with the usage of lighter materials and by making the cabin more aerodynamic. The output from the data analysis could also be extremely useful in the development of applications and services sold with the trucks with the purpose of helping the drivers and the haulage companies to decrease their fuel consumption, such as the on-board application I-See and the off-board service Fuel Advice described in Section 2.1. Thus, there are several application areas where this kind of data analysis method can be used at GTT.

Furthermore, this kind of analysis could also be useful for the marketing and sales departments. In knowing how the customers use the products it can be verified and made sure that they sell the right product for the actual needs of the customer, as well as using the information for enhancing their customer offerings.

9.3 Tools and organizational capabilities

The third part of the big-data plan is about planning for having the right tools and organizational capabilities [Biesdorf et al., 2013]. In order for the analytic models to contribute to creating value, the managers and frontline employees must be able to understand and use them [Biesdorf et al., 2013]. This kind of problem was observed in this case study. The tool (LAT) used for accessing the logged vehicle data, described in Section 2.3, was rather complicated and not sufficiently intuitive to use, since not many used it even after having gone through training in this tool. It was not either the lacking knowledge and experience about trucks that made this tool difficult to use since many of the product development engineers did not really know how or did not feel comfortable to use it. It took a lot of experience of working in the tool to be able to use it properly. This signifies that the tool was not intuitive enough which led to that only a few people used the tool and therefore only a few people used the logged vehicle data residing in the database. The tool was not developed ideally for the purpose of analyzing the data with the *process view* and recognition of patterns in large populations, but often to analyze one vehicle at a time as isolated occurrences, such as when the trucks were serviced in the workshops. A few people were skilled at using this tool, but it was not completely user friendly for them either since it was very time consuming to use and not error free, which could be connected to the *volume* of the big data and the issue of cloud computing described in Section 4.1.3. In order to fully capitalize the data, GTT should consider making the connected tools intuitive and so that they integrate data into day-to-day processes and translate modeling outputs into tangible business actions as stated by Biesdorf et al. [2013]. Their big-data plan should include how the outputs from the data analysis should be transformed into useful information for for example product development and

marketing and sales so that it can be used where value is created.

Nevertheless, it is not enough to have a big-data plan and follow it if the organization lacks capabilities and the right people [Biesdorf et al., 2013]. In many companies there is a shortage of people with the right experience and deep analytical expertise for handling this level of complexity [Manyika et al., 2011]. Even Volvo Group has shortage of data scientist competence since there is a very high demand of this competence in many companies in all kinds of industries and the companies are fighting hard to attract these resources. However, there are resources with this kind of competence in the company but they are scarce and there could be a problem in bringing the right people together to share knowledge when the organization is large and wide.

In this case study, there was a problem in accessing knowledge about the configurations of trucks in order to extract the right population from the database. This difficulty could have been mitigated if this knowledge could somehow be captured and shared in the organization. This is a form of *standardization* as described by Liker and Meier [2006], since by standardizing today's best practices, the learning up to this point is captured. By sharing best practices in the organization this way, standards provide a launching point for true and lasting innovation [Liker and Meier, 2006]. In accordance with Liker and Meier [2006], this kind of knowledge-sharing of both trucks and the information gained by the data analysis would make a groundwork and enable both continuous improvement of existing products and services as well as innovation.

For the quality work to be successful, it is essential to create conditions for participation in the work with continuous improvement [Bergman and Klefsjö, 2010]. There are three important components that are key to *let everybody be committed* and to participate, which are communication, delegation and training [Bergman and Klefsjö, 2010]. Training is important for a person to take responsibility, and the employee must have a chance to feel commitment, professional and personal pride, as well as responsibility, to be able to do a good job [Bergman and Klefsjö, 2010]. Even though the tools should be intuitive and easy to use, it is important to include training in the big-data plan, and as mentioned above, GTT trains their employees in different tools already. Once employees are trained authority can be delegated to them together with the freedom to do their job [Ishikawa, 1985]. For GTT training could enable delegating decisions which have to be made based on the model outputs provided by the tools.

Communication is also important since information is needed for a person to take responsibility and to understand the importance of his or her task for the goals of the whole organization [Bergman and Klefsjö, 2010]. Therefore an important strategic function of a big-data plan for GTT would also be to communicate to the organization the importance of this data, its potential and its role in order for the organization to reach its strategic goals described in Section 2.2. It can be argued that it is important to communicate the purpose of working with the big data so that for example the task of maintaining the database with correct data also becomes an essential step in the value chain and the process of reaching the

strategic goal of improving fuel efficiency through vehicle optimization, diesel efficiency and electromobility. Marketing and sales could also use this data to improve and create new business opportunities by using the data to build new services and tomorrow's business models.

In order to succeed in communication and letting everybody be committed so as to engage the organization in using the tools, strong leadership is needed [Biesdorf et al., 2013]. *Committed leadership* is essential to create a culture for successful and sustainable quality improvements [Bergman and Klefsjö, 2010]. Strong leadership is important to handle the challenge of bringing business players and data scientists to work together to find the best solutions [Biesdorf et al., 2013] along with a cultural shift to establish the mind-sets and behaviors to breech today's silos [Manyika et al., 2011]. In order for GTT to fully exploit the data, people with different knowledge and experience need to be brought together since there are both organizational and technical challenges as well as challenges connected to the specific application of the data analysis when dealing with big data. Strong and committed leadership can enable this, and will be important to consider in for example projects concerning big data involving Volvo IT, GTT and the marketing and sales organizations.

As mentioned above, a big-data plan may highlight a need for a massive reorganization of data architecture over time [Biesdorf et al., 2013]. The rising volume of data from new sources requires a new level of storage and computing power [Manyika et al., 2011]. Investments to develop interfaces and protocols to share data effectively across extended enterprises are needed [Manyika et al., 2011]. This is also going to be needed at GTT in order to create intuitive tools and to share the big data both cost- and storage-effectively across different parts of the organization along the value chain.

10 Conclusions

In this chapter, conclusions are drawn in order to answer the purpose and research questions of this thesis. It also outlines suggestions for future work in this reserach area.

10.1 Conclusions

Properties describing truck usage could be found in the logged vehicle data using the two techniques FA and PCA. One specific property appeared in the results from both techniques, which underlines its importance for the chosen truck population. This property described how much cruise control and the accelerator was used. PCA gave a more stable result compared to FA where the factors and their loadings changed as the number of factors was changed. There are also more approaches available for deciding on how many components to keep in PCA than factors in FA. Clustering of the data with the two techniques k-means Clustering and Hierarchical Clustering using these properties revealed groups of trucks where the importance of the truck usage properties varied. This result could however not be linked to truck applications as the analysis of the configuration of the trucks was too complex. This was due to the population containing only the application long haul trucks with complex configurations and not a heterogeneous truck population containing more easily differentiable trucks. Moreover, for Hierarchical Clustering it was difficult to decide on the number of clusters using the dendrogram. This was a lot more straightforward for k-means Clustering.

The model estimation techniques PCR, Lasso and Elastic Net were used to explore the importance of fuel consumption parameters for the found clusters of trucks.

PCR revealed models with a low explained variance and some insignificant coefficients which led to getting models difficult to evaluate. Both the Lasso and Elastic Net outperformed PCR but had a similar performance. The explained variance was a lot higher for these models. Overall, the parameters important for fuel consumption according to the model estimations varied a lot between the techniques. No conclusions about which parameters that are important for fuel consumption can therefore be drawn.

The used logged vehicle data revealed issues related to how the data has been logged, stored and structured. 65 percent of the data had to be removed due to missing parameter values and some of the initially wanted parameters could not be included. This led to a more complicated pre-processing step and made the data less suitable for this kind of analysis. In this case study all the so called Vs of big data *volume*, *variety*, *velocity* and *veracity* were observed. *Velocity* was already somewhat handled in the downloading strategy of the data, but the other issues had to be considered in the data analysis performed.

For GTT to be able to use this data for more advanced analysis and exploit its full potential, it can be concluded that a big-data plan needs to be created at a strategic level in the organization. This plan should include a strategy for assembling and integrating data, which is connected to the Total Quality Management cornerstones *focus on customers*, *base decisions on facts* and *focus on processes*. The gathered data should then be used in an advanced analytic model, for example including some of the data analysis techniques evaluated in this thesis. The analytic model should be designed so that quality can be improved in accordance with the cornerstones *focus on processes* and *improve continuously*. It should also be designed to provide the right people with the right information at the right place in order to put the data into use where it creates the most value, both for the customers and for the organization. User friendly and intuitive tools, knowledge sharing and strong leadership is needed in order for this to be obtained, which is also connected to the cornerstones *let everybody be committed* and *committed leadership*.

10.2 Suggestions for future work

There are several possibilities for improvement of the choice of data, such as reconsidering the parameter choice to include engine related parameters, speed and acceleration. Integrating geographical data could also be interesting and useful for finding patterns in the data related to fuel consumption.

One of the ideas which did not fit into the scope of this thesis was correlating the aggregated data to time-series data by calculating aggregated parameters with the goal to validate the results from the aggregated data. Using time-series data also has potential of revealing more precise conclusions. Data downloaded more frequently with telematics could be used and maybe give clearer results.

There is also room for improvements in interpreting the patterns into applica-

tions of the trucks. Here the trucks were grouped together and differentiated based on usage. Another approach is to leave out the clustering step and sort the trucks into different already known applications or different fleets if applicable. Some customers only have one truck, but some have fleets of trucks of the same product. In this case the model estimation could be done on the different fleets. This could be advantageous since it could be that the trucks in the same fleet drive in similar conditions.

The result from using Hierarchical Clustering showed a potential of using this technique to find outliers in the population. This could be used as a diagnosis tool to find trucks deviating from the standard.

In this case study the data analysis was applied on trucks to discover what affects the fuel consumption. Another strategic technology area for Volvo Group is electromobility. This technology of electrifying the driveline has a high positive impact on energy efficiency of the vehicles and is applied in for example hybrid buses and now also electric buses. Since this is new technology the logged vehicle data and this type of data analysis could be used to verify how this complex technology works in real customer operation in relation to simulation and test environments. It would also be interesting to explore what usage and operational factors affect both the energy consumption and the performance of the electric propulsion system. To analyze what usage and operational factors affect the life time of the energy storage system, which is the battery storing the electric energy used to propel the bus, would also be interesting since this is a very sensitive and costly system.

Appendix

A

Tables from the Principal Component Regression

Table A.1: *The result and statistics from the PCR of the first three clusters estimated by k-means when modelling t_{fuel} with four principal components.*

Cluster 1				
	Est.	t-stat.	p-val.	CI
(intercept)	-0.014	-1.67	0.096	[-0.030, -0.003]
β_1	**-0.087**	**-6.28**	$\mathbf{1.17 \cdot 10^{-9}}$	**[-0.115, -0.060]**
β_2	**-0.122**	**-9.69**	$\mathbf{1.70 \cdot 10^{-19}}$	**[-0.147, -0.098]**
β_3	-0.011	-1.02	0.309	[-0.033, -0.010]
β_4	**-0.090**	**-6.68**	$\mathbf{1.20 \cdot 10^{-10}}$	**[-0.116, -0.063]**
Adjusted R^2	0.429	R^2	0.436	
F-value	57.9			
Cluster 2				
	Est.	t-stat.	p-val.	CI
(intercept)	0.006	1.89	0.060	[-0.0002, 0.012]
β_1	**-0.045**	**-3.75**	**0.0002**	**[-0.068, 0.021]**
β_2	**-0.113**	**-9.16**	$\mathbf{5.08 \cdot 10^{-18}}$	**[-0.138, -0.089]**
β_3	-0.012	-0.92	0.358	[-0.037, 0.014]
β_4	**-0.060**	**-4.45**	$\mathbf{1.17 \cdot 10^{-5}}$	**[-0.087, -0.034]**
Adjusted R^2	0.249	R^2	0.258	
F-value	29.3			
Cluster 3				
	Est.	t-stat.	p-val.	CI
(intercept)	-0.001	-0.133	0.895	[-0.012, -0.010]
β_1	**-0.044**	**-2.82**	**0.005**	**[-0.075, -0.013]**
β_2	**-0.090**	**-5.61**	$\mathbf{4.13 \cdot 10^{-8}}$	**[-0.122, -0.059]**
β_3	**-0.043**	**-2.94**	**0.003**	**[-0.073, -0.014]**
β_4	**-0.058**	**-3.03**	**0.003**	**[-0.096, -0.020]**
Adjusted R^2	0.166	R^2	0.175	
F-value	18.6			

Table A.2: *The result and statistics from the PCR of the last three clusters estimated by k-means when modelling t_{fuel} with four principal components.*

Cluster 4				
	Est.	t-stat.	p-val.	CI
(intercept)	**0.019**	**5.17**	$\mathbf{3.42 \cdot 10^{-7}}$	**[0.012, 0.027]**
β_1	**-0.033**	**-3.86**	**0.0001**	**[-0.050, -0.016]**
β_2	**-0.071**	**-8.13**	$\mathbf{3.43 \cdot 10^{-15}}$	**[-0.089, -0.054]**
β_3	**0.036**	**4.65**	$\mathbf{4.24 \cdot 10^{-6}}$	**[0.021, 0.051]**
β_4	-0.014	-1.94	0.053	[-0.028, 0.0002]
Adjusted R^2	0.2	R^2	0.207	
F-value	32.2			
Cluster 5				
	Est.	t-stat.	p-val.	CI
(intercept)	**0.015**	**5.08**	$\mathbf{5.64 \cdot 10^{-7}}$	**[0.009, 0.021]**
β_1	**-0.067**	**-5.93**	$\mathbf{6.28 \cdot 10^{-9}}$	**[-0.089, -0.045]**
β_2	**-0.111**	**-11.23**	$\mathbf{9.30 \cdot 10^{-26}}$	**[-0.131, -0.092]**
β_3	0.023	1.77	0.078	[-0.003, 0.049]
β_4	-0.005	-0.480	0.632	[-0.026, 0.016]
Adjusted R^2	0.286	R^2	0.293	
F-value	42.9			
Cluster 6				
	Est.	t-stat.	p-val.	CI
(intercept)	**-0.089**	**-10.04**	$\mathbf{9.07 \cdot 10^{-21}}$	**[-0.106, -0.071]**
β_1	**-0.108**	**-6.87**	$\mathbf{3.36 \cdot 10^{-11}}$	**[-0.139, -0.077]**
β_2	**-0.143**	**-6.97**	$\mathbf{1.84 \cdot 10^{-11}}$	**[-0.183, -0.103]**
β_3	**-0.061**	**-3.72**	**0.0002**	**[-0.093, -0.029]**
β_4	-0.179	-9.34	$1.784 \cdot 10^{-18}$	[-0.217, -0.141]
Adjusted R^2	0.3287	R^2	0.395	
F-value	51.4			

Bibliography

Accenture. Most U.S. Companies Say Business Analytics Still Future Goal, Not Present Reality. Press Release, December 11 2008. http://newsroom.accenture.com/article_display.cfm?article_id=4777, Accessed on [2014-06-16]. Cited on pages 24 and 38.

A. Alessandrini, F. Filippi, F. Ortenzi, and F. Orecchini. A new method for collecting vehicle behaviour in daily use for energy and environmental analysis. In *Proceedings of the Institution of Mechanical Engineers, Part D (Journal of Automobile Engineering)*, volume 220, pages 1527–1537, 2006. Cited on page 4.

Andreassons Åkeri. Kvalité. Webpage, 2014. http://www.andreassonsakeri.se/kvalite.html, Accessed on [2014-08-09]. Cited on page 10.

P. Bartram. The value of data. *Financial Management*, 42(2):26–31, March 2013. Cited on pages 33 and 95.

B. Bergman and B. Klefsjö. *Quality from Customer Needs to Customer Satisfaction*. Studentlitteratur, Lund, 3:6 edition, 2010. Cited on pages xii, 1, 12, 18, 19, 20, 21, 22, 23, 24, 25, 26, 27, 93, 94, 95, 97, 100, and 101.

S. Biesdorf, D. Court, and P. Willmott. Big data: What's your plan? *McKinsey Quarterly*, (2):40–51, June 2013. Cited on pages 36, 37, 38, 39, 95, 97, 98, 99, 100, and 101.

C. M. Bishop. *Pattern Recognition and Machine Learning*. Springer, 2006. Cited on pages 53, 56, 57, and 65.

C. Bouveyron and C. Brunet-Saumard. Model-based clustering of high-dimensional data: A review. *Computational Statistics and Data Analysis*, December 2012. Cited on page 63.

I. Carpatorea, S. Nowacryk, T. Rögnvaldsson, and M. Elmer. Towards Data Driven Method for Quantifying Performance of Truck Drivers. Ongoing Ph.D. thesis at Halmstad University, 2014. Cited on pages 4 and 7.

M. Chen, S. Mao, Y. Zhang, and V. C. M. Leung. *Big data : related technologies, challenges and future prospects*. Springer, 2014. Cited on pages 29, 30, 34, 35, and 97.

H.-J. Cho, R.-J-Li, H. Lee, and J.Y.-J. Wu. Vehicle classification using support vector machines and k-means clustering. In *AIP Conference Proceedings*, volume 1148, pages 449–452, Hersonissos, Crete (Greece), August 2009. Cited on page 5.

T. H. Davenport, J. G. Harris, and R. Morison. *Analytics at Work: Smarter Decisions, Better Results*. Harvard Business Press, 2010. Cited on pages 24, 38, 39, and 98.

R. D. De Veaux, P. F. Velleman, and D. E. Bock. *Stats: Data and Models*. Pearson Education, Inc., 3rd edition, 2010. Cited on page 56.

J. W. Dean and D. E. Bowen. Management Theory and Total Quality: Improving Research and Practice through Theory Development. *The Academy of Management Review*, 19(3):392–418, July 1994. Cited on pages 17 and 18.

W. E. Deming. *Out of the crisis*. Cambridge University Press, Cambridge, Massachusetts, 1986. Cited on pages xii, 1, 19, 20, 23, 24, 25, 27, 93, and 94.

M. Denscombe. *The Good Research Guide for small-scale social research projects*. McGraw-Hill Open University Press, Maidenhead, England, 3rd edition, 2007. Cited on pages 43 and 45.

E. Ericsson. Independent Driving Pattern Factors and Their Influence on Fuel-Use and Exhaust Emission Factors. *Transportation Research: Part D: Transport and Environment*, 6(5):325–345, September 2001. Cited on pages 5 and 7.

Facts Hunt. Total number of websites & size of the internet as of 2013. Webpage, 2014. `http://www.factshunt.com/2014/01/total-number-of-websites-size-of.html`, Accessed on [2014-07-24]. Cited on page 29.

Foria. VD har ordet. Webpage. `http://www.foria.se/vd-har-ordet.html`, Accessed on [2014-08-09]. Cited on page 10.

J. Gantz and D. Reinsel. Extracting Value from Chaos. *IDC iView*, June 2011. Available at: `http://www.emc.com/collateral/analyst-reports/idc-extracting-value-from-chaos-ar.pdf`, [Accessed on 2014-07-24]. Cited on page 29.

Michael Grant and Stephen Boyd. Graph implementations for nonsmooth convex programs. In V. Blondel, S. Boyd, and H. Kimura, editors, *Recent Advances in Learning and Control*, Lecture Notes in Control and Information Sciences, pages 95–110. Springer-Verlag Limited, 2008. `http://stanford.edu/~boyd/graph_dcp.html`. Cited on page 85.

Michael Grant and Stephen Boyd. CVX: Matlab software for disciplined convex

programming, version 2.1. http://cvxr.com/cvx, March 2014. Cited on page 85.

T. Grubinger. Knowledge Extraction from Logged Truck Data using Unsupervised Learning Methods. Master's thesis, Halmstad University, 2008. Cited on pages 4 and 7.

T. Grubinger, N. Wickström, A. Björklund, and M. Hellring. Knowledge Extraction from Real-World Logged Data. *SAE International Journal of Commercial Vehicles*, 2(1):64–74, 2009. Cited on pages 4 and 7.

J. F. Hair, Jr., W. C. Black, R. E. Anderson, and R. L. Tatham. *Multivariate Data Analysis*. Pearson Prentice Hall, 6th edition, 2006. Cited on page 79.

T. Hastie, R. Tibshirani, and J. Friedman. *The Elements of Statistical Learning*. Springer, 2nd edition, 2008. Cited on pages 60, 65, 67, 77, and 79.

J. Hetterich, A. Georgiadis, S. Bonnemeier, and M. Pritzke. Ecological sustainability – a customer requirement? Evidence from the automotive industry. *Journal of Environmental Planning & Management*, 55(9):1111–1133, November 2012. Cited on page 6.

C.-H. Hsu, T.-Y. Lee, and H.-M. Kuo. Mining the body features to develop sizing systems to improve business logistics and marketing using fuzzy clustering data mining. *WSEAS Transactions on Computers*, 8(7):1215–1224, July 2009. Cited on page 5.

S. W. Hunt, A. M. C. Odhams, R. L. Roebuck, and D. Cebon. Parameter measurement for heavy-vehicle fuel consumption modeling. In *Proceedings of the Institution of Mechanical Engineers, Part D (Journal of Automobile Engineering)*, volume 225, pages 567–589, May 2011. Cited on page 4.

IBM. Raising the game – the ibm business tech trend study. Technical report, IBM Center for Applied Insights, 2014. Cited on page 3.

M. Imai. *KAIZEN : att med kontinuerliga, stegvisa förbättringar höja produktiviteten och öka konkurrenskraften*. Konsultförlaget AB; with KAIZEN Institute of Europe, Uppsala, 1986. Translated from English by K. G. Fredriksson, 1992. Cited on pages 26 and 27.

K. Ishikawa. *What is total quality control? The Japanese way*. Prentice Hall, Englewood Cliffs, N.J., 1985. Translated from Japanese by D. J. Lu. Cited on pages 26 and 100.

H. V. Jagadish, J. Gehrke, A. Labrinidis, Y. Papakonstantinou, J. M. Patel, R. Ramakrishnan, and C. Shahabi. Big Data and Its Technical Challenges. *Communications of the ACM*, 57(7):86–94, July 2014. Cited on pages 2, 29, 30, 34, 35, 36, and 95.

G. James, D. Witten, T. Hastie, and R. Tibshirani. *An Introduction to Statistical Learning*. Springer, 2013. Cited on pages 55, 56, 77, 78, 81, and 85.

I. T. Jolliffe. *Principal Component Analysis*. Springer-Verlag New York Inc., 1986. Cited on pages 57 and 63.

J. M. Juran. Attaining Superior Results through Quality. In J.M. Juran and J. A. De Feo, editors, *Juran's Quality Handbook : The Complete Guide to Performance Excellence*, chapter 1. McGraw-Hill, 6th edition, 2010. Cited on pages 19, 20, 24, 25, and 98.

G. Köksal, I. Batmaz, and M. C. Testik. A review of data mining applications for quality improvement in manufacturing industry. *Expert Systems with Applications*, 38:13448–13467, 2011. Cited on page 5.

KPMG International. Competing in the Global Truck Industry – Emerging Markets Spotlight. Technical report, Institut für Automobilwirtschaft, September 2011. Cited on pages 2 and 12.

D. Laney. 3D Data Management: Controlling Data Volume, Velocity, and Varitey. *META Group Research Note*, February 2001. Cited on pages 30 and 34.

D. N. Lawley and A.E. Maxwell. *Factor Analysis as a Statistical Method*. Butterworths, 1963. Cited on page 63.

J. K. Liker and D. Meier. *The Toyota Way Fieldbook – A Practical Guide for Implementing Toyota's 4Ps*. McGraw-Hill, 2006. Cited on pages 23, 26, 98, and 100.

J. Manyika, M. Chui, B. Brown, J. Bughin, R. Dobbs, C. Roxburgh, and A. Hung Byers. Big data: The next frontier for innovation, competition and productivity. Technical report, McKinsey Global Institute, May 2011. Cited on pages 30, 31, 32, 33, 37, 38, 94, 95, 97, 100, and 101.

S. Marsland. *Machine Learning: An Algorithmic Perspective*. Chapman & Hall/CRC, Boca Raton, FL, 2009. Cited on pages 57, 60, 61, and 64.

W. L. Martinez and A. R. Martinez. *Exploratory Data Analysis with MATLAB*. Chapman & Hall/CRC, 2005. Cited on pages 48, 55, 56, 57, 60, 61, and 63.

MATLAB. *version 8.4.0.150421 (R2014b)*. The MathWorks Inc., Natick, Massachusetts, 2014. Cited on page 43.

A. McGordon, J. E. W. Poxon, C. Cheng, R. P. Jones, and P. A. Jennings. Development of a driver model to study the effects of real-world driver behaviour on the fuel consumption. In *Proceedings Of The Institution of Mechanical Engineers, Part D (Journal Of Automobile Engineering)*, volume 225, pages 1518–1530, November 2011. Cited on pages 4 and 5.

M. Montazeri-Gh, A. Fotouhi, and A. Naderpour. Driving patterns clustering based on driving feature analysis. In *Proceedings of the Institution of Mechanical Engineers, Part C (Journal of Mechanical Engineering Science)*, volume 225, pages 1301–1317. Published for the Institution of Mechanical Engineers by Professional Engineering Publishing Ltd, UK, June 2011. Cited on pages 5 and 7.

D. C. Montgomery, E. A. Peck, and G. G. Vinning. *Introduction to Linear Regression Analysis*. Wiley, 4th edition, 2006. Cited on pages 77, 78, and 81.

Moore's Law. Moore's Law or how overall processing power for computers will double every two years. Webpage. `http://www.mooreslaw.org/`, Accessed on [2014-07-30]. Cited on page 34.

K. P. Murphy. *Machine Learning: A Probabilistic Perspecitve*. The MIT Press, Cambridge, Massachusetts, 2012. Cited on pages 58, 61, 64, 65, 67, and 72.

G. Subramanya Nayak, D. Ottolina, and C. Puttamadappa. Classification of Bio Optical signals using K-Means Clustering for Detection of Skin Pathology. *International Journal of Computer Applications*, 1(2):92–96, 2010. Cited on page 48.

A. Neely, C. Adams, and M. Kennerley. *The Performance Prism: The Scorecard for Measuring and Managing Business Success*. Prentice Hall Financial Times, London, 2002. Cited on page 11.

J. S. Oakland. *Total Quality Management : text with cases*. Elsevier Butterwort-Heineman, Oxford, 3rd edition, 2003. Cited on pages xii, 17, 18, 20, 21, 22, 23, 24, 27, 95, 97, and 98.

R. Prytz, S. Nowaczyk, T. Rögnvaldsson, and S. Byttner. Analysis of Truck Compressor Failures Based on Logged Vehicle Data. In Hamid Reza Arabnia, editor, *CREA Press*, 2013. Cited on pages 4 and 12.

V. Ribeiro, J. Rodrigues, and A. Aguiar. Mining geographic data for fuel consumption estimation. In *Proceedings of the International IEEE Conference on Intelligent Transportation Systems (ITSC)*, pages 124–129, The Hague, Netherlands, 2013. Cited on pages 4 and 7.

T. A. Schmitt and D. A. Sass. Rotation Criteria and Hypothesis Testing for Exploratory Factor Analysis: Implications for Factor Pattern Loadings and Interfactor Correlations. *Educational and Psychological Measurement*, 71(1):95–113, January 2011. Cited on page 61.

Skoogs Åkeri och Logistik. Om företaget: Detta är Skoogs Åkeri och Logistik AB. Webpage, 2014. `http://www.skoogsakeri.se/`, Accessed on [2014-08-09]. Cited on page 10.

M. Söderman. Driver coaching. Interview, 2014. Interviewed by S. Felldin and S. Johnsen at Volvo Group Trucks Technology, Advanced Technology and Research, Gothenburg, 2014-06-09. Cited on pages xii, 10, and 11.

R. Sousa and C. A. Voss. Quality management re-visited: a reflective review and agenda for future research. *Journal of Operations Management*, 20(1):91–109, February 2002. Cited on page 17.

N. Tare, S. Kabirdas, P. Chandrasekar, and S. Singh. Strategic Outlook of Commercial Vehicles Telematics in 2014. Technical report, Frost & Sullivan, March 2014. Cited on page 33.

N. T. Trendafilov, S. Unkel, and W. Krzanowski. Exploratory factor and principal component analyses: some new aspects. *Statistics and Computing*, 23(2):209–220, March 2013. Cited on page 63.

Volvo Group. The Volvo Group's Environmental Policy. Inhouse document, 2012a. http://www.volvogroup.com/SiteCollectionDocuments/VGHQ/Volvo%20Group/Volvo%20Group/Our%20values/Environment/policy_environment_volvo_eng.pdf, Accessed on [2014-09-13]. Cited on page 10.

Volvo Group. The Volvo Group's Quality Policy. Inhouse document, 2012b. http://www.volvogroup.com/SiteCollectionDocuments/VGHQ/Volvo%20Group/Volvo%20Group/Our%20values/Quality/policy_quality.pdf, Accessed on [2014-09-13]. Cited on pages 10 and 18.

Volvo Group. The Volvo Group Sustainability Report 2013: Our Progress Towards Sustainable Transport Solutions. Available at: http://www3.volvo.com/investors/finrep/sr13/sr_2013_eng.pdf, [Accessed on 2014-06-17], 2013. Cited on pages 5, 9, 10, and 13.

Volvo Group. Volvo Group report on the fourth quarter 2014. Available at: http://www3.volvo.com/investors/finrep/interim/2014/q4/q4_2014_eng.pdf, [Accessed on 2016-02-03], 2014. Cited on pages 2 and 9.

Volvo Group Trucks. Delivering our full potential: Group trucks strategy 2013–2015, key focus areas and strategic objectives. Inhouse document, 2013. Cited on page 13.

Volvo Group Trucks Technology. Volvo Group Trucks Technology. Webpage, 2016. http://www.volvogroup.com/group/global/en-gb/volvo%20group/our%20companies/GTtechnology/Pages/GTT2.aspx, Accessed on [2016-02-06]. Cited on page 9.

Volvo IT. About us. Webpage, 2014. http://www.volvoit.com/volvoit/global/en-gb/about_us/Pages/Aboutus.aspx, Accessed on [2014-06-18]. Cited on page 15.

Volvo Trucks. Every Drop Counts. Webpage, 2014a. http://www.volvotrucks.com/trucks/uk-market/en-gb/aboutus/every-drop-counts/Pages/landing.aspx, Accessed on [2014-06-17]. Cited on page 12.

Volvo Trucks. I-See comes with a bank of fuel savings. Webpage, 2014b. http://www.volvotrucks.com/trucks/uk-market/en-gb/aboutus/every-drop-counts/Pages/i-see.aspx, Accessed on [2014-06-17]. Cited on page 12.

Volvo Trucks Support Services. Introducing Remote Diagnostics. Brochure, 2012. http://www.volvotrucks.com/SiteCollectionDocuments/VTNA_Tree/ILF/Flash/Remote%20Diagnostics%20Brochure2.pdf, Accessed on [2016-02-09]. Cited on page 12.

C. Voss, N. Tsikriktsis, and M. Frohlich. Case research in operations management. *International Journal of Operations & Production Management*, 22(2):195–219, 2002. Cited on pages 42, 43, and 44.

VSAT. In 2013 the amount of data generated worldwide will reach four zettabytes. VSAT Voice, Blog, 21 June, 2013. `http://vsatglobalseriesblog.wordpress.com/2013/06/21/in-2013-the-amount-of-data-generated-worldwide-will-reach-four-zettabytes/`, Accessed on [2014-07-24]. Cited on page 29.

X. Wang, K. Smith, and R. Hyndman. Characteristic-based clustering for time series data. *Data Mining and Knowledge Discovery*, 13(3):335–364, November 2006. Cited on page 48.

L. Witell. Kundorienterat förbättringsarbete – att lyssna och agera på kundens röst. In M. Elg, V. Gauthereau, and L. Witell, editors, *Att lyckas med förbättringsarbete – förbättra, förändra, förnya*, chapter 4. Studentlitteratur, Lund, 2007. Cited on pages 19, 20, and 26.

I. H. Witten. *Data mining: practical machine learning tools and techniques*. Morgan Kaufmann Publishers, Burlington, Massachusetts, 3rd edition, 2011. Cited on page 3.

A. G. Woodside. *Case Study Research: Theory, Methods, Practice*. Emerald, 2010. Cited on page 43.

World Meteorological Organization. A summary of current climate change findings and figures. A WMO information note, March 2013. Available at: `http://unfccc.int/files/cc_inet/application/x-httpd-php/ccinet_getfile.php?file=147`, Accessed on [2014-07-09]. Cited on page 5.

H. Zou and T. Hastie. Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society Series B.*, 67(2):301–320, 2005. Cited on pages 84 and 85.