Fall 12-20-2015

Predictive Modeling of Fuel Efficiency of Trucks

Srivatsa Bindingnolle Narasimha

PREDICTIVE MODELING OF FUEL EFFICIENCY OF TRUCKS

BY

SRIVATSA BINDINGNOLLE NARASIMHA

BS, National Institute of Technology, Surat, 2011

THESIS

Submitted in partial fulfillment of the requirements for
the degree of Master of Science in Industrial and Systems Engineering
in the Graduate School of
Binghamton University
State University of New York
2015

Accepted in partial fulfillment of the requirements for
the degree of Master of Science in Industrial and Systems Engineering
in the Graduate School of
Binghamton University
State University of New York
2015

December 8, 2015

Dr. Nagendra N Nagarur, Faculty Advisor
Department of Systems Science and Industrial Engineering, Binghamton University

Dr. Chun-An Chou, Committee Member
Department of Systems Science and Industrial Engineering, Binghamton University

Sung Hoon Chung, Committee Member
Department of Systems Science and Industrial Engineering, Binghamton University

ABSTRACT

This research studied the behavior of several controllable variables that affect the fuel efficiency of trucks. Re-routing is the process of modifying the parameters of the routes for a set of trips to optimize fuel consumption and also to increase customer satisfaction through efficient deliveries. This is an important process undertaken by a food distribution company to modify the trips to adapt to the immediate necessities. A predictive model was developed to calculate the change in Miles per Gallon (MPG) whenever a re-route is performed on a region of a particular distribution area. The data that was used, was from the Dallas center which is one of the distribution centers owned by the company. A consistent model that could provide relatively accurate predictions across five distribution centers had to be developed. It was found that the model built using the data from the Corporate center was the most consistent one. The timeline of the data used to build the model was from May 2013 through December 2013. The predictive model provided predictions of which about 88% of the data that was used, was within the 0-10% error group. This was an improvement on the lesser 43% obtained for the linear regression and K-means clustering models. The model was also validated on the data for January 2014 through the first two weeks of March 2014 and it provided predictions of which about 81% of the data was within the 0-10 % error group. The average overall error was around 10%, which was the least for the approaches explored in this research. Weight, stop count and stop time were identified as the most significant factors which influence the fuel efficiency of the trucks. Further, neural network architecture was built

to improve the predictions of the MPG. The model can be used to predict the average change in MPG for a set of trips whenever a re-route is performed. Since the aim of re-routing is to reduce the miles and trips; extra load will be added to the remaining trips. Although, the MPG would decrease because of this extra load, it would be offset by the savings due to the drop in miles and trips. The net savings in the fuel can now be translated into the amount of money saved.

ACKNOWLEDGEMENT

It is a great pleasure to thank my committee members for their time and support. First, I would like to express my sincere gratitude to my advisor, Dr. Sarah S. Lam for her constant encouragement and guidance throughout my study at Binghamton University. Her wisdom and attitude inspired me greatly to achieve the required goals during my research work. In addition, I extend my appreciation for committee member, Thomas Kneebis from Maines Inc. Paper and Food Service, Inc. who provided guidance on the fuel efficiency project that has greatly deepened my understanding and interests in this area. His valuable insights about the food transportation business went a long way in helping me to identify ways to tackle difficult obstacles during the project. I also thank Bryan Mozeleski for his timely insights on routing. I would like to thank the faculty and staff of the SSIE department that have given me great support during my time at Binghamton University. Particularly, I want to acknowledge the help provided by Dr.Peter Borgesen who inspired and encouraged me during the initial days of my stay at this university. I also express my gratitude to Dr. Krishnaswami Srihari for giving me a chance to join the WISE group and work with many great people. His work ethic and discipline is something which I want to follow throughout my career. Most importantly, I have to express my deepest appreciation to my parents for their persistent encouragement, support and instilling confidence in me. I appreciate everything they have done for me and I would not be where I am without them. It is their great love towards me that have enlightened every moment of my life. One other person who was

also a source of encouragement was my dear friend, Dasharathy Ramesh, an alumni of Binghamton University who guided me when I first entered this country, advised me on the very small things that were important and also helped me with my coursework. He also guided me to the right people and hence, I will always be grateful to him.

Finally, thanks to all the people that have helped me and this thesis is dedicated to them.

Table of Contents

List of Tables

List of Figures

Chapter 1 - Introduction

1.1 Introduction

The transportation of commodities is a very important part of any manufacturing business. The transition of goods to the customers located in different regions brings in revenue and also increases customer satisfaction. This generates profit and adds value to the business. The goods transportation sector can be considered as a major source of revenue for the US economy by providing timely delivery of quality goods at lower costs. The U.S. transportation network serves more than 300 million people and 7.5 million business establishments across 3.8 million square miles of land. The transportation of goods takes place through a vast network of rails, roads, waterways, and airways. Fruits, vegetables, and equipment are being transported across the length and breadth of the country to thousands of customers. The goods transportation sector accounts to 6.2% of the GDP of the entire country. Consequently, it has become a huge sector for investment and for obtaining large profits. There are a lot of logistic companies that ply their trade in this sector. The competition has increased many fold, especially in recent years. According to a Bureau of Transportation Statistics (BTS) report in 2001, on an average day, about 43 million tons of goods valued at about $29 billion moved nearly 12 billion ton-miles on the nation's interconnected transportation network. This represents an increase from about 37 million tons, valued at $20 billion in 1993. Given the increase in the demand for quality goods, the transportation sector has become an area of importance in terms of revenue and customer satisfaction.

1.2 Problems in Transportation and Distribution

The number of customers banking on reliable and just in time deliveries of products has grown in recent years. Hence, it is necessary to develop state-of-the-art equipment to provide efficient service. This will help to build a large customer base for the distribution industry. Distribution companies must tend to the demand in locations very far away from the warehouse. There are numerous factors that affect the transportation of the finished products, such as weather, terrain, distance, and type of transport. The transport vehicles, irrespective of their type, must be sufficient enough to handle all the above factors when distributing the goods to the desired location. A distribution business cannot lose a potential customer just because the location is far away or they do not have the vehicles to transport the products. Sufficient research and development must be done by the distribution companies in this area. Larger distances also mean that the quality of the products should be maintained all through the journey from the warehouse to the destination. The appropriate technology must be used to prevent damage to the products. The key point is to reach the desired level of infrastructure to meet all kinds of demands. As trucks are involved in the transportation of goods, fuel plays a major role in contributing to the expenditure of a distribution business. Since fuel is an expensive commodity, the trips for the vehicles must be planned out efficiently. Also, the products must be transported shortly after their production to remove the problem of excess inventory due to which the company can incur holding and shrink loss. There is a certain amount of uncertainty involved in all the processes that make up a distribution system. Since, they cannot be prevented, steps must be taken to minimize them so as to keep the customers happy and continue the inward flow of revenue.

1.3 Food Distribution

As discussed earlier, goods can be transported to different customer locations through rail, roads, airways, and water ways. For the purpose of this research, we will be focusing on the transportation of food products through trucks. Food transportation is very different from the transportation of hardware and equipment. Food should be handled with care and proper safety measures must be taken to maintain its quality. There are types of food that spoil very fast. Hence, while transporting food over long distances, built-in equipment on the trucks are a necessity to prevent it from perishing. Different kinds of food products require different kinds of packaging. The trucks must have the proper technology to maintain the quality of the food products. The majority of the transportation of products in the U.S. takes place through trucks. According to the BTS report of 2002, 67 percent of shipments were delivered through trucks all over the country. In 2012, truck transport services accounted for about $228 billion. The trucks may be designed for long distance or short distance trips, depending upon the customer location. It is essential that these trucks are of the highest quality to ensure the efficient delivery of food products. The drivers who are driving the trucks must be well trained to cope with the challenges that come with long distance trips. One of the main problems with truck transportation is that these trucks are large and difficult to maneuver. The next problem that truck transportation faces is driver shortage. A deficiency of experienced drivers might result in customer dissatisfaction, which leads to a decrease in the revenues and profits of the company. It has been reported in 2012 that there was a shortage of 25,000 drivers for every 750,000 trucks. Trips must therefore be planned properly and assigned to the appropriate drivers. The distribution company must also ensure that the drivers meet the drivers meet the federal and state safety regulations for the trucks for the trucks. Drivers are required to undergo extensive health checkups before they embark on their respective trips. The aim is to maximize safety and minimize incidents.

The performance of trucks is affected by many factors, both preventable and unpreventable. Weather and topography cannot be controlled and the trucks must be modified to adapt to these changes. For example, snow tires might be suitable for the winters when the roads are very slippery. The topography also plays a very important role in the performance of the trucks. The altitude changes and the slopes of the roads improve or decrease the engine efficiency. The specifications of the trucks play a major role in the trucks' performance. The load carried by the trucks and the tire resistance are another set of important factors.

1.4 Problem Statement and Objective

Fuel is an expensive commodity and is a major part of the expenditure of any distribution company, as all trucks need fuel. According to the BTS report of 2012, truck transport services consume about 32% of imported fuel. Fuel demand is going to increase about 50% by the year 2050. This is a very costly issue and hence an aim should be to maximize the output of trucks by improving their performance and minimizing their fuel consumption. Petroleum is a limited source of fuel and its usage must be minimized. The problem statement of this research is to identify the factors that affect the fuel efficiency of the trucks and their sensitivity towards changes in route plans. As explained earlier, there are a lot of factors that affect engine and truck performance. Many of them directly or indirectly affect the fuel consumption. While most distribution companies tend to look at changing the routes and trips of the trucks to increase the performance, they do not look at minimizing fuel consumption. It has been estimated through studies by the distribution company that, at the present cost of fuel, 1/10th of an MPG can save around $400,000 annually. This amount is significant and is certainly an interesting area of research to be examined. Improper routing is another common problem in distribution companies. This can lead to a decrease in truck performance, driver dissatisfaction, and thereby result in inefficient service to the customers. There is extensive data mining research being done in order to save fuel by increasing fuel economy and optimizing routing of trucks. However, research has not been done to predict the impact on MPG due to re-routing by identifying the most significant factors that affect fuel efficiency and building a predictive model based on these factors. This thesis currently addresses the above problems and also develops a tool to visualize these parameters on which decisions can be made to save fuel. The main objectives of this research is to develop a predictive model to measure the changes in fuel efficiency based on the identified factors and consequently, measure the sensitivity of the fuel efficiency towards changes in the trip plans for the trucks. If a working system is developed to predict the impact of re-routing on MPG and to calculate the net savings in fuel and dollars, it helps in developing an optimized routing system which is cost efficient and at the same time, develops a delivery system which provides quality service to the customers

1.5 Thesis Overview

Chapter 1 provides a brief overview of the common problems in the field of transportation and distribution. A part of the chapter is devoted to food distribution. It also provides a summary of the problem that is addressed through this research and the objectives that are going to be fulfilled by the end of it.

Chapter 2 provides information about the various kinds of data mining techniques that are used to solve various problems generally and also in the field of transportation and distribution. Various research papers are reviewed as a basis for this particular research. It also explains what has not been done and how this thesis is unique compared to the reviewed articles.

Chapter 3 explains the methodology used to meet the objectives, in detail. It provides the steps taken to develop the predictive model to estimate fuel efficiency, related numbers and the validation of the models. This also provides a detailed study of the sensitivity analysis of the fuel efficiency towards the re-routing of trucks.

Chapter 4 provides insights based on the observations made from the output obtained from the model. It also suggests the measures that can be taken to improve the fuel efficiency based on these insights.

Finally, Chapter 5 concludes with a brief explanation of the future work wherever there are areas of improvement in the current thesis.

Chapter 2 - Literature Review

2.1 Problems in Transportation and Distribution

The transportation of goods involves many uncertainties, which affect the performance of the various participants in the processes. A lot of these uncertainties cannot be prevented, but they can be controlled. The transportation of goods is delayed by weather, terrain, traffic, driver performance, fuel performance, and technical glitches. There are also other factors related to logistics which play a major role in transportation and distribution. Historically, the transportation of goods has been very problematic. During the emergency of the Second World War, when American forces landed in Normandy to usurp the Nazi regime, they had to travel through dangerous terrains, which made the transportation of arms and ammunition a major hurdle. Several techniques were developed to make this process easier. Temporary bridges and all terrain trolleys were used to move the goods faster (Cohen, 1986). Moving forward in the timeline, during the Gulf War, many efficient ways were developed to move war machinery through treacherous routes of the Middle Eastern desert (Pisters, 2010). In more recent times, many logistics companies like DHL and FedEx have developed efficient ways of transporting goods to distant places in a quick and risk-free manner (Ulrich, 2011). With the competition in the logistics industry becoming more intense than ever, the companies need to find cheaper and quicker methods of transportation to be successful. Much research has been conducted to achieve these goals. The routing of trucks involved in distribution is one of the most basic problems that have been addressed in recent years. Although the routing problem was extensively discussed during the 1950s (Dantzig and Ramser, 1959), it has only been recently implemented on a large scale. The vehicle routing problem involves the planning of a route for the truck to transport goods from the warehouse to the various customer centers. These customer centers might be located at varied distances from the warehouse. Therefore, an efficient route must be planned in a way that the truck delivers the goods to all the customers in the shortest time possible. The objective is to minimize the number of trucks involved and also to minimize the delivery time (Fan et al., 2009). In doing so, the trucks must not be overloaded and they must return to the warehouse after delivery. Simulation models were also developed to create efficient route plans using queuing networks and perturbation analysis (Solberg, 1977). An optimization solution for the transportation of soft drinks was developed with various factors like number of trucks, truck capacity and routing costs (Privé, 2006). Most of the research is focused on decreasing the number of trips to the destinations as much as possible and maximizing the utilization of the trucks. Electronic devices have been developed to record various parameters to analyze the performance of the trucks based on the route plan (Duin and Kneyber, 2004). There has also been research done on the integration of advanced technology into the routing of the trucks like the establishment of wireless communication networks from the trucks to a relay center where the performance is analyzed. The aim is to reduce the number of instances when the trucks returns empty because backhaul is introduced (Hayashi and Yano, 2003). The effect of traffic on route plans is a major player in the efficiency of the transport processes. Deliveries to customers located in metropolitan areas have resulted in longer times compared to customers located on the outskirts. However, longer distances also become a major player (Kovács, 2010). Therefore a balance must be struck between both. Another important problem in transportation and distribution is fuel economy. As explained earlier, fuel is a costly commodity and must be utilized efficiently. Since the days of the Gulf War, prices of fuel have increased exponentially with increasing demand. To add to this, studies have shown that the energy to output ratio has decreased by 55 percent since 1978 (Greene et al., 1999). This is a very alarming trend. Encouragement of the usage of renewable sources of energy, reduction of oil extraction and restrictions on percentage of carbon emission from fuels are some of the steps undertaken to increase the energy to output ratio of fuels. The development of technology has had a significant impact on the efficient usage of fuels. Efficient fuel consumption leads to a significant amount of savings. This could increase the profits of a company many fold. Trucks should have proper emission systems to increase fuel efficiency. An added benefit it that this also helps minimize fuel's effect on the environment. The quality of the fuel should be maintained at a very level. The extra cost of investment in high quality fuel could go a long way in saving millions of dollars due to the increase in fuel efficiency (Kydes, 1999). There are many factors that affect fuel efficiency. The Corporate Average Fuel Economy Standard sets a particular threshold for fuel economy. Trucks are allowed to carry more weight compared to passenger vehicles and the threshold is lower compared to that of cars (Godek, 1997). In spite of this, there are many trucks that breach the threshold. This has been a major problem, which is costing millions of dollars. The engine performance plays a major role in fuel emission and economy. The effect of hydrocarbons, carbon monoxide, and nitrous oxide in the fuel exhaust on the fuel economy is very significant because of the deposition of precipitates on the inner walls of the exhaust pipes. This hinders the complete combustion of fuel which in turn decreases the efficiency. The engines need to be designed in such a way that all the above mentioned chemical particles need to be filtered before exhaust. Studies have shown that by doing this, the efficiency goes up by 10%. This is a significant increase and can save fuel and money (He et al., 2011).

There are numerous factors that affect the fuel efficiency of the trucks, which range from fuel quality to driver performance. There has to be a systematic analysis to study the effect of each of these factors on the fuel economy. A data mining methodology to develop relationships between the fuel efficiency and these factors will shed some light on it.

2.2 Introduction to Data Mining

Data mining methodologies have already been used in the field of transportation and distribution of commodities widely. Data mining techniques have been used to identify the factors influencing the fuel economy of cars (Gushue and Wu, 2004). But, the research was mainly on comparing the variables related to the car models and their sales, to the variables related to the specifications of the cars. Data mining is the process of identifying hidden patterns to study the behavior of certain practices followed by an institution. Data is investigated to obtain useful information based on which important decisions can be made. Invariably, the output of data mining can be used to build models that help in foreseeing future behavior, identify behaviors that could not be observed earlier, develop relationships between two or more participants in the process, or adjust an existing principle in order to improve the process. There are various practical applications where data mining can be applied to improve processes. It is applicable in fields like health care, finance, space technology, and retail (Fayyad et al., 1996). There are a lot of data mining methods like neural networks, clustering, decision and association trees, which are widely used (Barai, 2003). Neural networks have seen a recent bump in their applications to various real life problems. They are very powerful tools to detect complex nonlinear patterns to fit the data in an accurate manner. Researchers have broadly classified data mining into two stages, the first stage being the generation of trends from the data that is acquired and the second stage being the translation of these trends to business decisions (Kohavi, Sommerfield and Dougherty, 1997). Based on the outcome, we can proceed to develop models that can be used to produce these trends with the incoming data. Regression tree techniques have been used on various applications like root cause analysis and fault detection (Loh, 2011). Pattern detection systems employ this technique in criminal analysis, forensic studies, and finger print analysis. This method involves the division of data into sub groups, upon which separate models are built.

Naturally, the data that is used for analysis and model building becomes a vital part of data mining. Without proper data, it is not possible to develop insights as to what decision can be taken. These expensive decisions cannot be based on improper data. This can be harmful to a business. The first step is to verify the origin of the data. Data can be collected in real time or from a device that has already recorded it. Both of these processes involve errors that need to be removed while analyzing the data. The next step is to look into the timeline of the data, which depends upon the objectives based on which the usage of long term or short term data can be determined. Once the data is obtained, it needs to be cleansed of the missing values and abnormal values, which might create bias in the analysis. The cleansed data can then be used to investigate for patterns and develop models.

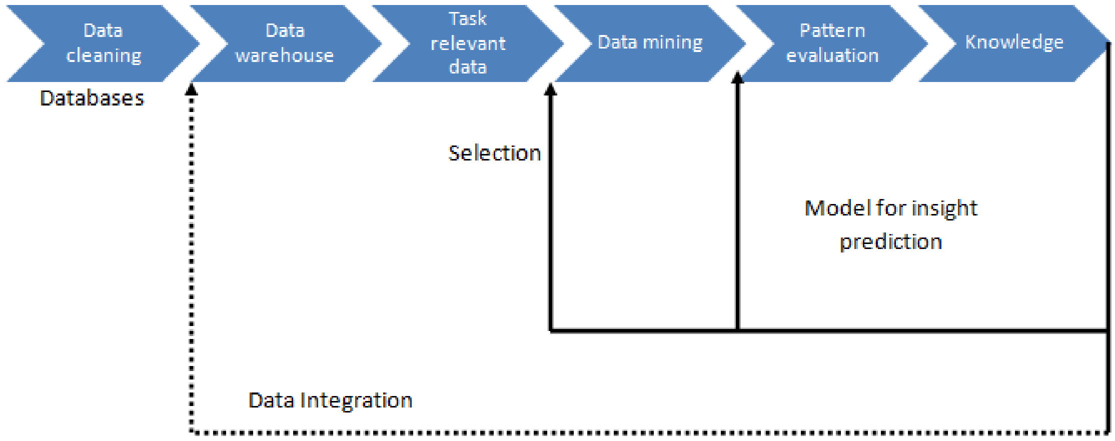


Figure 2.1 Data mining process (Barai, 2003)

As shown in the Figure 2.1, data must be mined to uncover patterns which are crucial to the development of knowledge based on which decisions can be taken. The most important part of data mining is to choose the right method to build models for the data. This will depend upon the type and behavior of the data (Fayyad et al., 1996). The developed models need to be validated by comparing the outcomes of the models to the real time results. Once the models are validated, business decisions can be established which will help improving the various operations in the company.

2.3 Applications of Data Mining

Data Mining is applicable in a wide range of fields, and within many specialties that require process improvement.

In the field of space technology, computational software has been developed to process huge amounts of image data of the universe from the Hubble space telescope and to categorize them based on certain characteristics. NASA has provided significant funding to integrate all the image observations from far away galaxies into a data bank. Its success can be measured by its ability to process three terabytes of data and compile them. A variation of the principal component analysis is used since the data is in the form of images instead of numbers (Foslien et al., 2004). The patterns obtained from mining this data is useful for measuring the velocity and mass of the galaxies and also the rate at which stars are born in these galaxies. It is also used to study the nature of formation of the galaxies. Data mining is also used in business and finance. Fayyad et al. highlighted the use of clustering to identify specific customer groups to market a particular product and forecast their behavior in the near future (Fayyad et al., 1996). Retail businesses have a lot of room for improvement in terms of increasing sales, customer satisfaction, and quality of products. Agrawal and Psaila, used clustering to identify closely related brands of the same product. The clusters can be used to place an alternate brand of the same product in a store if the original brand is out of stock (Agrawal and Psaila, 1995). This helps the retail stores to maintain the flow of revenue even though the major selling brand is not available. It is widely known that all the major investment bankers use data mining techniques like neural networks and genetic algorithms to identify potential areas of investment and the business decisions to be taken on existing ones (Hall et al., 1996). The US treasury uses data mining methods to detect any kind of malpractice happening in the financial world (Shapiro et al., 1996). Credit card frauds and risk analytics requires the application of advanced data modeling to predict the occurrence of strange activities in a customer transaction. Clustering, regression and classification techniques are employed to develop a fraud detection system (Ngai et al., 2011). Data mining techniques are also used to prevent money laundering (Liao et al., 2012). Tax evasion is a serious crime and amounts to great losses to the country's revenue. A lot of research has gone into the application of data mining techniques to detect the occurrence of tax frauds. False transactions and identities have been detected, which could have led to tax evasions in billions of dollars. Clustering algorithms are used to develop root cause models which can lead to the detection of such kind of illegal practices (Gonzalez and Velazquez, 2013). Bayesian techniques and association rules were used to detect irregularities in the ATM transactions (Li et al., 2012). In the field of manufacturing, clustering methods are used to detect and classify faults appearing in Boeing aircrafts (Kusiak, 2006). Data mining is also used to classify the type of weld formation. Machine learning models are developed to predict the type of weld formation on the joints of two metals (Barai and Reich, 2002). Studies have been done to compare the accuracy of predictions of advanced data mining techniques like decision trees, support vector machines and neural networks. The models are used to predict random occurrences where normal data mining techniques cannot be adopted. These methods are used to predict the occurrences of landslides (Pradhan, 2012). Neural networks have been used to develop steering technologies for unmanned vehicles. They also have applications in cyber security where they have been used to detect attacks on sensitive data (Ahmad et al., 2009). Neural networks have been used in fingerprint detection (Arrietta et al., 2009). Data mining is also used in the field of healthcare. They are used to identify the occurrence of any malpractices in health insurance (Koh and Tan, 2011). Prediction of surgery times and scheduling of hospital staff are done through data mining techniques. Sequential rules are used to analyze alarm sequences for a telecommunication company. This helps to obtain information from the incoming signals and translate them into a meaningful form (Mannila et al., 1995). Neural networks have found many uses in medicine. Neural networks are particularly useful in recognition and aiding in medical diagnosis for breast cancer analysis (Abbass, 2009).

2.4 Data Mining in Transportation

Transportation and distribution involves various processes, many of which are expensive and have little room for error. A small mistake can lead to significant losses. Hence, it is necessary to be careful while making decisions related to scheduling and routing of trips assigned to the trucks. It has become clear that within transportation and distribution there are opportunities for process improvement via data mining. A lot of data that is collected has either been analyzed manually, which is subject to human errors, or the data has not been used at all. More recently, there has been a sharp rise in the amount of data mining research within this field. A huge amount of data is generated by the flow of vehicles, occurrence of accidents, quality of roads, transport of goods, routing of trucks, and maintenance of vehicles. Reputed companies in goods distribution like DHL, FedEx, APL Logistics, Wilson Logisitics and other automobile industries base their decisions on the trends developed by analyzing the data. It has also been realized that this area requires a lot of collection of real time data. As a result, high-end devices have been developed to systematically store real time information and relay it back to the center for further analysis (Amado, 2000).

In predicting the behavior of drivers, data mining techniques are used to predict the sleeping patterns and temperament of the drivers in order to promote safe driving (Ji et al., 2004). Traffic is a serious problem with a lot of uncertainties involved. By clustering traffic intervals with similar conditions, patterns are developed which are used to monitor the flow of traffic without any congestions (Wong and Woon, 2008). Optimization algorithms are also used to develop traffic patterns to prevent accidents and allow the free passage of vehicles. These algorithms are used to optimize the routing of trucks in distribution companies.

Trucks that are carrying goods and are traveling long distances must have reliable equipment embedded in them. Any breakdown en route to the destination can cause huge delays in the delivery of goods and can decrease customer satisfaction. This has a huge impact on revenue and future business for the transportation company. Decision forest analysis has been used to find out the cause for fault occurrence in the trucks (Singh et al., 2012). Randomly occurring faults are hard to predict and analyze. The engine has a lot of controls with very intricate designs. It is very important to collect data on various parameters affecting the engine performance. With all this data, a decision tree was developed to find out the cause of the fault. Artificial Neural Networks (ANN) provide a broad spectrum of functions which are required in the field of engine applications (modeling, especially for controller design, onboard testing and diagnostics). Exhaust emissions laws are becoming progressively more stringent, while the pressure on fuel economy has been intensifying significantly in the last few years. For diesel engines, a large number of technologies, such as, multi-pulse injection and variable valve actuation, show significant promise to both improve fuel economy and reduce exhaust emissions (Deng et al., 2008).

2.5 Data Mining in Distribution - Routing and Fuel Economy

Distribution companies possess trucks to transport the goods from warehouses to customers. They consume a significant of fuel and are very expensive. Hence, it is essential to develop optimized trips for the trucks so they will consume as little fuel as possible without compromising the quality of the service. It is also required to identify the factors that influence the fuel consumption. Once this is done, a predictive model can be developed to measure the fuel consumption based on the values for the identified factors and measures can be taken to save fuel.

There are many data mining techniques that can be used to solve problems in routing. Association rules technique is one of them. It involves developing relationships between metrics present in the data. Rules are generated based on this so as to satisfy a set of pre-defined conditions for the problem. This technique was used to develop a routing plan for a set of vehicles so as to prevent congestion of vehicles in a particular area (Hasheem, 2011).

Principal component analysis has been used to investigate the major factors which influence the fuel consumption of a vehicle. Vernon and Meier identified that 91% of the fuel that is consumed is exposed to two major principal factors that decreases its quality, thereby causing pollution and increased consumption (Vernon and Meier, 2012). Statistical analysis can also be used to pinpoint the factors causing an increase in fuel consumption. Clustering is another method in which closely related factors associated with fuel can be grouped and insights can be generated based on the observations made (Vernon and Meier, 2012).

Since transportation involves the moving objects, it is appropriate to use spatio-temporal data to solve many problems. Spatio-temporal data holds information about the location and time of the object in question at a particular instant. There are many case studies where this kind of data is analyzed, which have found that the objects travel in a particular pattern at specific intervals. Cao et al. considered the usage of this data to study the uncertainties involved with the movement of trucks. (Cao et al., 2007).

Global positioning systems are used to collect real time data for vehicle movements. The data is collected through satellites, which are then transmitted back to data collection station. This station is equipped with servers which maintain the database. A combination of clustering and neural networks is used to mine this kind of data to obtain insights. Edelkamp and Schrödl used the data to develop search algorithms to find out the shortest path for the vehicles to move from one destination to another (Edelkamp and Schrödl, 2003).

A feature that would allow any given airline the potential of predicting its fuel consumption throughout the year, select the plane/flight that minimizes the whole fleet consumption, make more precise estimations on the cost of fuel during a certain period of time and acquire the fuel when it is at its lowest price is of the utmost value (Spencer, 2011).

2.6 Chapter Summary

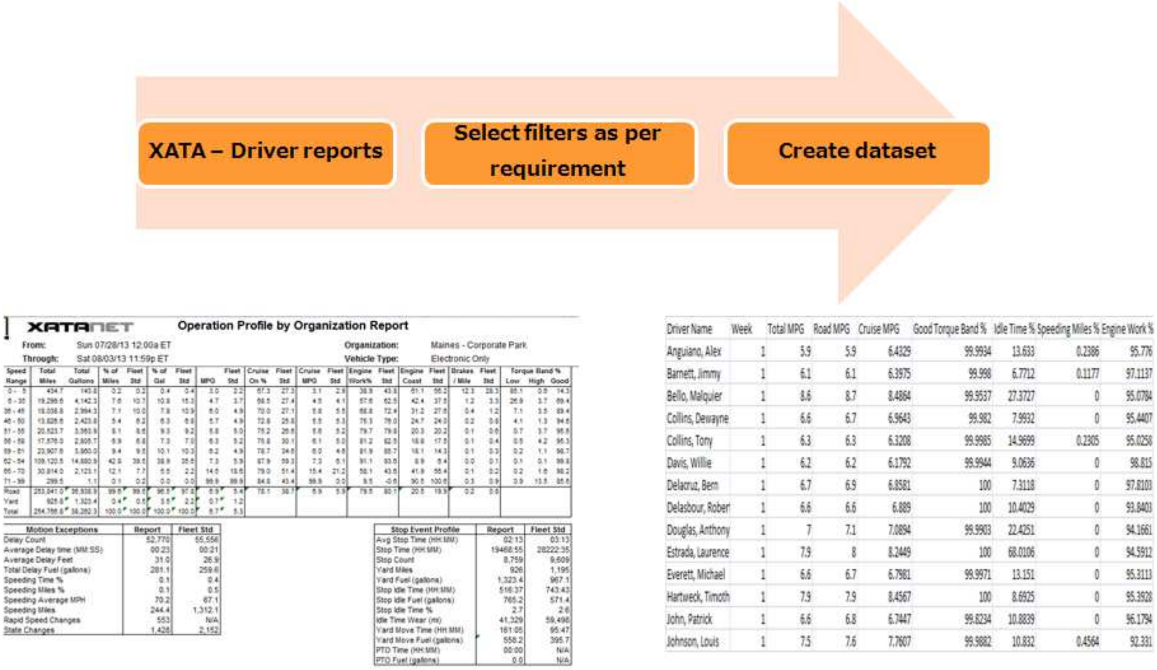
There are many problems that affect the transportation processes involved in delivery of goods to customers in a timely manner. The trucks involved must be of the highest quality to increase the performance, thereby increasing the customer satisfaction and revenue to the company. These problems need to be addressed systematically to create an efficient distribution network. Fuel efficiency and vehicle routing are two of the major factors that affect the performance. But there are numerous other factors that affect the fuel efficiency. A data mining methodology helps in understanding the behavior of these factors and their effects on the fuel efficiency. This chapter reviews the literature on research, which is done to address the problems in the transportation and distribution industry. It also throws light on the usage and advantages of data mining to solve many problems.

Chapter 3 - Methodology

In this chapter, all the methods implemented to achieve the research objective are explained. A description of how the data is extracted and cleansed is provided, and methods for data modeling and validation are discussed.

3.1 Data Description

The distribution company which is headquartered in Conklin, NY, specializes in the distribution of food products to customers located in over 38 states. Trucks are used to transport the goods from the warehouses located at the 10 distribution centers across the country to customer locations. The company delivers goods to established food chains like Red Lobster and Burger King. The drivers of the trucks are assigned trips to a set of customer locations. The trucks are equipped with Electronic On-Board Recorder (EOBR) systems to record the driver and truck data. This data is later used to manage the fuel efficiency of the trucks. The data that arrives needs to be stored in a database. There are many systems available and the company is currently using XATA, which is now called XRS Corporation. The original report from XATA is extracted by selecting filters, and then a dataset is created. The data is at the driver level. The extracted report consists of several variables, which describe the performance of the drivers in the selected time period. The primary aim is to investigate the relationships between the variables and also between fuel efficiency and the variables. It would also be helpful to identify the variables that have the most impact on the MPG. Regression is a method to investigate these aforementioned relationships (Shapiro et al., 1996).



XATANET Operation Profile by Organization Report

From: Sun 07/28/13 12:00a ET
Through: Sat 08/03/13 11:59p ET
Organization: Maines - Corporate Park
Vehicle Type: Electronic Only

Speed Range	Total Miles	Total Gallons	% of Miles	Fleet Std	% of Gal	Fleet Std	MPG	Fleet Std	Cruise On %	Fleet Std	Cruise MPG	Fleet Std	Engine Work%	Fleet Std	Engine Coast	Fleet Std	Brakes / Mile	Fleet Std	Torque Band % Low	High	Good
0 - 5	434.7	140.8	0.2	0.2	0.4	0.4	3.0	2.2	67.3	27.3	3.1	2.9	38.9	43.8	61.1	56.2	12.3	28.3	85.1	0.8	14.3
6 - 35	19,298.6	4,142.3	7.6	10.7	10.8	15.3	4.7	3.7	68.5	27.4	4.5	4.1	57.6	62.5	42.4	37.5	1.2	3.3	26.9	3.7	69.4
36 - 45	18,038.8	2,994.3	7.1	10.0	7.8	10.9	6.0	4.9	70.0	27.1	5.8	5.5	68.8	72.4	31.2	27.6	0.4	1.2	7.1	3.5	89.4
46 - 50	13,826.6	2,423.8	5.4	6.2	6.3	6.8	5.7	4.9	72.8	25.8	5.5	5.3	75.3	76.0	24.7	24.0	0.2	0.8	4.1	1.3	94.6
51 - 55	20,523.7	3,560.9	8.1	8.6	9.3	9.2	5.8	5.0	75.2	26.8	5.8	5.2	79.7	79.8	20.3	20.2	0.1	0.6	0.7	3.7	95.8
56 - 58	17,576.0	2,806.7	6.9	6.8	7.3	7.0	6.3	5.2	75.8	30.1	6.1	5.0	81.2	82.5	18.8	17.5	0.1	0.4	0.5	4.2	95.3
59 - 61	23,907.8	3,960.0	9.4	9.5	10.1	10.3	6.2	4.9	78.7	34.8	6.0	4.8	81.9	85.7	18.1	14.3	0.1	0.3	0.2	1.1	98.7
62 - 64	109,120.5	14,880.9	42.8	39.5	38.9	35.6	7.3	5.9	87.9	59.3	7.3	6.1	91.1	93.6	8.9	6.4	0.0	0.1	0.1	0.1	99.8
65 - 70	30,814.0	2,123.1	12.1	7.7	5.5	2.2	14.5	13.6	79.0	51.4	15.4	21.2	58.1	43.6	41.9	56.4	0.1	0.2	0.2	1.6	98.2
71 - 99	299.5	1.1	0.1	0.2	0.0	0.0	99.9	99.9	84.8	43.4	99.9	0.0	9.5	-0.6	90.5	100.6	0.3	0.9	3.9	13.5	85.6
Road	253,841.0	36,938.9	99.6	99.5	96.5	97.8	6.9	5.4	78.1	38.7	6.9	5.9	79.5	80.1	20.5	19.9	0.2	0.6			
Yard	925.6	1,323.4	0.4	0.5	3.5	2.2	0.7	1.2													
Total	254,766.6	38,262.3	100.0	100.0	100.0	100.0	6.7	5.3													

Motion Exceptions	Report	Fleet Std
Delay Count	52,770	55,556
Average Delay time (MM:SS)	00:23	00:21
Average Delay Feet	31.0	26.9
Total Delay Fuel (gallons)	281.1	259.6
Speeding Time %	0.1	0.4
Speeding Miles %	0.1	0.5
Speeding Average MPH	70.2	67.1
Speeding Miles	244.4	1,312.1
Rapid Speed Changes	553	N/A
State Changes	1,428	2,152

Stop Event Profile	Report	Fleet Std
Avg Stop Time (HH:MM)	02:13	03:13
Stop Time (HH:MM)	19468:55	28222:35
Stop Count	8,759	9,609
Yard Miles	926	1,195
Yard Fuel (gallons)	1,323.4	967.1
Stop Idle Time (HH:MM)	516:37	743:43
Stop Idle Fuel (gallons)	765.2	571.4
Stop Idle Time %	2.7	2.6
Idle Time Wear (mi)	41,329	59,498
Yard Move Time (HH:MM)	161:05	95:47
Yard Move Fuel (gallons)	558.2	395.7
PTO Time (HH:MM)	00:00	N/A
PTO Fuel (gallons)	0.0	N/A

Driver Name	Week	Total MPG	Road MPG	Cruise MPG	Good Torque Band %	Idle Time %	Speeding Miles %	Engine Work %
Anguiano, Alex	1	5.9	5.9	6.4329	99.9934	13.633	0.2386	95.776
Barnett, Jimmy	1	6.1	6.1	6.3975	99.998	6.7712	0.1177	97.1137
Bello, Malquier	1	8.6	8.7	8.4864	99.9537	27.3727	0	95.0784
Collins, Dewayne	1	6.6	6.7	6.9643	99.982	7.9932	0	95.4407
Collins, Tony	1	6.3	6.3	6.3208	99.9985	14.9699	0.2305	95.0258
Davis, Willie	1	6.2	6.2	6.1792	99.9944	9.0636	0	98.815
Delacruz, Bern	1	6.7	6.9	6.8581	100	7.3118	0	97.8103
Delasbour, Rober	1	6.6	6.6	6.889	100	10.4029	0	93.8403
Douglas, Anthony	1	7	7.1	7.0894	99.9903	22.4251	0	94.1661
Estrada, Laurence	1	7.9	8	8.2449	100	68.0106	0	94.5912
Everett, Michael	1	6.6	6.7	6.7981	99.9971	13.151	0	95.3113
Hartweck, Timoth	1	7.9	7.9	8.4567	100	8.6925	0	95.3928
John, Patrick	1	6.6	6.8	6.7447	99.8234	10.8839	0	96.1794
Johnson, Louis	1	7.5	7.6	7.7607	99.9882	10.832	0.4564	92.331

Figure 3.1 Data extraction and cleansing

Roadnet is a database software that is used by the company. This database stores the data for the trips undertaken by the drivers. The information stored in the database is related to weight of the trucks, number of trips, and the data for various parameters related to these trips. This software also helps in performing re-routing using a map depicting all the trips for the selected region.

The data from XATA with the driver information and the data from Roadnet related to the trip information are extracted and combined to form one single dataset containing both the driver and trip information against the driver names. The data was extracted for the Dallas region from May 2013 to December 2013. The reason for Dallas distribution center as a choice for extraction of data was because of the following reasons. Dallas is a new distribution center that was started in April 2012 and also one of the smallest ones. The weekly MPG values for the drivers from July 2013 were showing a decline compared to the corresponding values in July 2012

The following factors were present in the initial dataset.

Cruise MPG: The MPG of the trucks when the truck is in cruise mode (50 - 70 MPH). The general belief is that the longer the truck is in cruise mode, the higher the fuel efficiency of the truck is. The cruise mode of the truck depends upon the location of the destination, weather, terrain, and other topological factors. It might seem logical that the longer the distance is between the customer destination and the warehouse, the higher the chance is that the truck will be in cruise mode for a longer time.

Cruise %: The percentage of the distance for which the trucks were in cruise mode. This is an extension of the "Cruise MPG" variable. This might not be a true reflection of the MPG because even if the distance is smaller and the truck stays in the cruise mode for most of the journey, then the Cruise % is high. This does not imply the cruise MPG is high.

Speeding miles and gallons: The number of miles and the gallons consumed when the truck is traveling over the speed limit.

Idle time: The amount of time the truck is in idle mode even though the engine is running. This decreases the fuel efficiency of the trucks. The truck might idle under different circumstances ranging from stopping at a traffic signal or stopping for the delivery of the goods.

Idle fuel: The amount of fuel consumed by the truck when it is idle.

Torque band percentage: This reflects the range of operating speeds under which the engine is able to operate efficiently. The torque band range is only a center percentage of the engine speed range. Hence, higher the torque band percentage, higher is the efficiency of the engine which in turn increases the MPG.

Weight: The amount of load that the truck is carrying while completing the trip. The higher the weight of the truck, the lower the fuel efficiency.

Backhaul: The amount of weight the truck is carrying as return load. This adds to the weight of the truck and decreases the fuel efficiency.

Stop time: The total time that the truck has stopped without the engine running during the trip.

Stop count: The amount of stops by the truck during the trip to deliver the goods.

Age: The age of the drivers

Experience of drivers in food service: Numbers of years in service of the drivers in the food service business.

Experience of drivers in the company: Number of years in service of the drivers in the company from which the data is obtained.

Driver names: Name of the drivers.

These factors affect fuel efficiency in different ways. Many of them are controllable and many or not. There is a necessity to study the relationships between the controllable variables and fuel efficiency.

3.2 Data Preprocessing

The data set consists of the information related to the performance of the drivers in the Dallas region. The data contains of information related to a particular driver name called "AIMS" which is the identification name assigned to the trip when the truck is driven to the yard and back. After the completion of a trip or at the beginning of a trip, a truck is driven from the warehouse to the yard or vice versa. No particular driver is assigned to this task and hence it is assigned the name "AIMS". This task also consumes fuel and hence, it is recorded in the driver information log. It has numerous zero values for the associated variables and might cause bias in the results for any future analysis (Acuna and Rodriguez, 2004). Since it only comprises less than 2% of the data set, which has

2048 rows of data, it is feasible to omit them from analysis without any loss of information (Little and Rubin, 1989).

There were also rows of data with the MPG as zero. This occurrence is not possible and was found to be an error after a discussion with the involved persons. As a result 25 rows of data were then removed.

3.3 Regression

The dataset in hand consists of many variables. To re-iterate, the primary goal is to develop a predictive model for fuel efficiency by developing relationships between the MPG and the performance variables found in the data. One of the most basic techniques to develop relationships between independent and dependent variables is regression analysis. If a data set contains both dependent and independent variables, regression analysis can help in establishing and quantifying the effects of the independent variables on the dependent variables (Bartz-Beielstein and Markon, 2004). One other use of regression analysis is to predict the dependent variables by fitting a model into the dataset of the independent variables. The two main types of regression are linear regression and nonlinear regression.

3.3.1 Linear Regression

The simplest way to fit a model into the data with many variables is linear regression. It is the process of developing linear relationships between the independent and dependent variables. If a single dependent variable is present, it is called simple linear regression, and if there are more than one dependent variable, it is called multiple linear regression. Linear regression involves the development of a linear function, which predicts the independent variable and also quantifies the relationships between them (Bartz-Beielstein and Markon, 2004).

The linear relationship between the dependent variable Y and "n" independent variables is denoted as:

$$Y = a + b_1X_1 + b_2X_2 + \dots\dots\dots\dots + b_nX_n + e \qquad (3\text{-}1)$$

where $X_{1,}$ $X_{2,}$ X_3........X_n are the independent variables, e is the predictive error between the actual value of Y and the predicted value from the expression.

The co-efficient "b" is the slope of the regression line. The co-efficient "b" represents the value of Y when the independent variables are 0.

When the prediction model is built, it should calculate the fuel efficiency very accurately. The difference error between the predicted MPGs and the actual MPGs should be low.

Thus, the aim of linear regression is to find the line that bests fit into the data so that the errors are minimal.

This means that the line should provide values of Y which are very close to the actual values of Y for the same values of X. Thus the error term "e" in the equation 3-1 represents the difference between the actual value and predicted value of Y from the model.

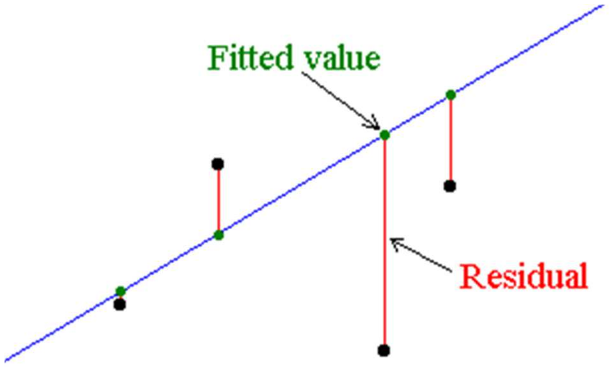


Figure 3.2 Depiction of error term and residual

The measured distance between the fit and the actual points of Y are called residuals as shown in Figure 3.2. Thus a good model would be where the residuals are close to 0. The effect of X on Y can be explained by the co-efficient "a" or the slope of the line. The higher the co-efficient value, the greater is the effect of X on Y. Since the dataset consists of numerous variables, it is not possible for all those variables to have the same effect on the fuel efficiency. A particular variable might have a greater positive effect compared to another which has a lower negative effect. So it is very important to measure the significance of each variable. The significance of the variable X can also be explained by another term called the p-value. This value determines statistically whether there is really a non-zero co-efficient for X, or in other words, whether there is a true linear relationship between X and Y. The p-value is determined once the line is fit into the data and then the error values and the co-efficients are calculated (Passing and Bablok, 1983). If the null hypothesis is assumed as that there is no relationship between X and Y, a p-value of anything below 0.05 reveals that, from the available data provides evidence against the null hypothesis. Thus it is rejected and it can be inferred that there is indeed a strong relationship between X and Y. The lower the p-value, the higher the accuracy of the model in terms of low prediction errors, which also means a high linear relationship between X and Y.

The goodness of fit for the linear model is measured by the R-squared value. It represents the amount of variation in Y that is captured by X. This measures the accuracy of the linear regression model. The R-squared value will be high if the fitted line is closer to the data points.

$$R - squared \;=\; 100 * {}^{SS(regression)}\!\big/_{SS(total)} \qquad (3\text{-}2)$$

SS (regression) in (3-2) describes the variance within the estimated values of Y, and is the sum of the squared difference between each instance of Y and the mean of Y. The squares are considered to avoid the positive and negative signs from the residual values. SS (error) accounts for the deviation from observed Y of the estimated Y. It is the sum of all the individual residual values. SS (total) describes the variation within the values of Y, and is the sum of the squared difference between each instance of Y and the mean of Y.

3.3.2 Stepwise Linear Regression

Stepwise regression is a type of regression where the best model is obtained by adding or removing variables in the model. There are basically two types:

1. Forward selection
2. Backwards elimination

This method produces a model by including or excluding variables from the model based on the alpha values.

1. Alpha to add variables, to include variables that are not present in the model at the time.
2. Alpha to remove variables, to remove variables that are already present in the model at the time.

The criteria used for model selection are:

1. Standard deviation of the residuals in the model. The model fits the data better if the standard deviation is small.
2. R-Squared is the amount of variation in the dependent variable explained by the terms in the model
3. R-Squared (adjusted) is a modified R that has been weighted to prevent overly optimistic R-squared values.

Forward elimination method starts with no independent variables in the model. The variables are added one by one based on the critical value i.e. alpha. Once they are included in the model, they are never removed.

Backward elimination method starts with all the independent variables in the model. The variables are removed one by one based on the critical value i.e. alpha. Once, they are removed from the model, they are never included.

3.3.3 Best Subset Regression

Best subset regression starts with developing models with one independent variables and selecting the model with the best R-squared (adjusted) value. Then it develops regression models with two independent variables and selects the model with the best R-squared (adjusted) value. This process continues till all the independent variables are exhausted.

3.3.4 Nonlinear Regression

The realm of the problem statement lies in the real world. As explained earlier, there are multiple uncertainties that affect the truck performance based on fuel efficiency. There is a possibility that there is no linear relationship at all between the variables and MPG. This means that a straight line cannot be fit into the dataset if there is no linear behavior. There is also a possibility that a combination of the variables might affect the MPG than the individual variables. All these possibilities lead to the necessity to consider nonlinear regression as part of the methodology.

Nonlinear regression is a type of regression where there is a nonlinear relationship between the dependent and independent variables.

$$Y_t = X_t^{\theta} + \varepsilon_t \qquad (3\text{-}3)$$

where θ is the parameter to be estimated. Similar to linear regression, a least squares method can be used to estimate θ by minimizing,

$$S(\theta) = \sum \left(Y_t - X_t^{\theta}\right)^2$$

(3-4)

The minimum of S is obtained by differentiating (3-4) with respect to θ, setting the derivative equal to zero,

$$\frac{\partial S}{\partial \theta} = -2 \sum\left(Y_t - X_t^{\theta}\right)(\log X_t)X_t^{\theta} = 0$$

(3-5)

Rearranging (3-5),

$$\sum Y_t \,(\log X_t)X_t^{\hat{\theta}} = \sum(\log X_t)\, X_t^{2\hat{\theta}}$$

(3-6)

(3-6) can yield the least square estimate only by an iterative procedure starting at some assumed initial value.

In most situations a linear line does not fit well. Instead, a nonlinear line with higher order terms fits very well into the data. A large sum of real world data behaves nonlinearly with many uncertain factors going into the model. A nonlinear model might help in accurately predicting the MPG and developing relationships between the variables, if it found that they behave nonlinearly.

Now that the two leading methods of regression have been discussed, there is one more technique of classifying the data based on similarity by which the multiple regression models can be built on multiple sets of data.

3.4 Cluster Analysis

Another method of developing prediction models is to break up the data into different clusters and then analyze each cluster separately. Since the data is at the driver level, there might be some minute variations that might not be captured when the whole dataset is used for regression. Thus, cluster analysis might help in studying the minute variations.

3.4.1 K-means Clustering

Since there is no prior knowledge on the rules for the division of data into clusters for the dataset in hand, it is preferable to use a technique that does not need any prior inputs to cluster the data. The K-means method is numerical, unsupervised, non-deterministic, and iterative (Kanungo et al., 2002).

The main idea is to define k centroids, one for each cluster. The next step is to measure the distance between a point and each centroid and assign the class of the centroid which is closest to the point. This process goes on until all the points are exhausted. (Kanungo et al., 2002). Now that all the points are assigned classes, new centroids of these classes must be calculated. After these k new centroids are obtained, the first step is repeated and the new classes are assigned which is followed by new K centroids and so on. The process stops when there is no significant difference between the old set of centroids and the new ones.. In other words, centroids do not move any more (Hartigan and Wong, 1979). This algorithm aims at minimizing the squared error cost function. The cost function

$$J = \sum_{j=1}^{k} \sum_{i=1}^{n} ||x_i^{(j)} - c_j||^2 \tag{3-7}$$

where $\left\| x_i^{(j)} - c_j \right\|^2$ is the squared distance between a data point $x_i^{(j)}$ and the cluster centre c_j, measures the distances between the data points and the cluster centers. Given a set of observations (x_1, x_2, …, x_n), where each observation is a d-dimensional real vector, k-means clustering divides the observations into k sets (k ≤ n) S = {S_1, S_2, …, S_k} so as to minimize the cost function (Hartigan and Wong, 1979) shown in (3-8)

$$\text{argmin} \quad \sum_{j=1}^{k} \sum_{i=1}^{n} \left\| x_i^{\{j\}} - c_j \right\| \tag{3-8}$$

The drawback of k means clustering is the assumption that the clusters must be of similar so that the nearest class assignment of the data points is performed without any errors. (Bradley and Manchek, 1998). The number of clusters k is an input parameter and hence, an inappropriate choice of k may yield poor results. That is why when performing k-means, it is important to run diagnostic checks for determining the number of clusters in the data set.

3.5 Sensitivity Analysis

Once a model is established along with the relationships between the variables and the MPG, it is necessary to validate the model for its robustness in extreme scenarios. For example, the MPG might decrease drastically when the weight carried by the truck becomes too high. It is also necessary to study the interactions between the variables themselves.

Sensitivity analysis involves the testing of the model built after the data analysis. It consists of techniques that evaluate the model by modifying the parameters in it. It is also used to study the behavior of the model in extreme conditions or boundary conditions. This helps in understanding the advantage and the limitations of the model in hand. When data modeling techniques are applied in real life situations, there exists a significant amount of uncertainty. These uncertainties contribute towards decreasing the accuracy of the model if they are not accounted for. Sensitivity analysis techniques help in analyzing the robustness of the model when encountered with these uncertainties It also helps to validate the relationships between the dependent and independent variables inserting extreme values to the independent variables and see how the dependent variables behave. Certain models have many inputs, which do not have a significant relationship with the output. These techniques help to remove unrelated inputs, thus making the model less complex. The accuracy of the models can then be increased by minimizing these errors.

Sensitivity analysis depends upon the time taken to run the model for a particular scenario (Helton et al., 2006). There are some techniques, which are normally used for linear regression models that cannot be used for nonlinear ones. By varying the inputs based on different scenarios, the behavior of the output can be studied. This variation in the output is measured by the sensitivity index (Saltelli and Annonni, 2010). The simplest and easiest method is to vary the value of one input at a time and studying the behavior of the output. The other variables are kept at their base-scenario values (Hamby, 1995). The sensitivity is measured based on the output using partial differentiation. This method is used when the input values are independent of each other. By changing only one variable at a time, an understanding of the relationships between the output and each of the inputs can be developed. However, it does not work on models where there are relationships between the inputs themselves, which might to lead to wrong observations about the prediction performance of the model (Czitrom, 1999). An alternate method is to vary one input from its maximum to minimum value and see how the output changes. The ratio of the difference in the output for the maximum and minimum values of the input gives us the sensitivity index. This helps in testing the robustness of the model in extreme conditions (Gray et al., 2005).

3.6 Chapter Summary

A description of how the data is extracted and cleansed is provided, and methods for data modeling and validation are discussed in this chapter. It reveals the methodology involved in analyzing the available data. It also explains the different techniques of regression, cluster analysis and neural network required for data modeling. Finally, it explains the steps involved in sensitivity analysis, which is required for model validation.

The data is extracted from the XATA database at the driver level. The data is cleansed by removing the missing values and treating the data that might cause bias in the analysis.

The cleansed data is used to develop a regression model to predict the MPG based on the relationship between the MPG and the variables and also between the variables themselves. It also helps in identifying the most significant variables that affect the MPG. The last step is to perform sensitivity analysis on the model to test for its accuracy in extreme scenarios. The relationships between the variables and the MPG are also quantified.

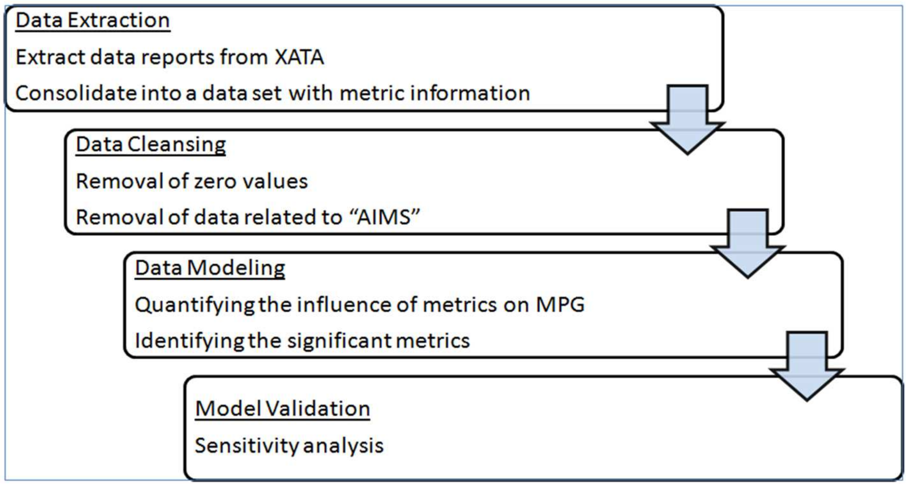

Data Extraction
Extract data reports from XATA
Consolidate into a data set with metric information
Data Cleansing
Removal of zero values
Removal of data related to "AIMS"
Data Modeling
Quantifying the influence of metrics on MPG
Identifying the significant metrics
Model Validation
Sensitivity analysis

Figure 3.3 Summary of methodology

Chapter 4 – Artificial Neural Networks

4.1 Neural Networks

An Artificial Neural Network is a data mining algorithm that is mimics the structure and operations of a human nervous system to sense and interact with objects around and process information. It consists of numerous interconnected layers consisting of processing units that function together to identify complex relationships in various problems. the human brain learns by experience and various iterations of trial and error. The artificial neural network behaves the same way. It is setup to develop relationships between the independent variables and the dependent variables. Neural Networks help in identifying complex relationships between the inputs and the output. They are called Artificial Neural Networks because of the fact they are used to study human computer interactions in artificial intelligence. In the most basic sense, they are mathematical models used to approximate a function $f : X \rightarrow Y$ or a distribution over X or both X and Y.

Neural networks are incorporated by systems to develop learning models in the same way the humans think and perform actions. They emulate hierarchical structure of the nervous systems with layers and layers of neurons connected to each other to enable lightning fast transmission of signals. The layers of processing units learn new features and remember them just like the human brain. This helps in identifying complex patterns and instances and transforms them into outputs. They are used in various fields like healthcare, credit and risk management or any information retrieval systems like cellphone speech recognition systems.

Although the artificial neural network is not as complex as the human nervous system, it processes the input features just like the human brain. These input features are multiplied by a weight based on different algorithms. They are then sent to the processing units which transform the features to develop relationships and pass them to the next layer of units. The final output will have the processed and transformed result based on the model developed by the neural network based on certain hidden relationships. The most simple neural network model adds up the values of the input features and passes them as output values. Complex neural networks are further built on this basic ideology.

The drawback of the neural network system lies within the structure. There is no set methodology to determine the number of layers between the inputs and the output and also the number of processing elements in each layer. So, this is based on a trial and error algorithm.

Constructing a neural network involves setting up the layers and elements and deciding the weights of the interconnections between the layers for the model to learn. Just like the human brain learns from experience, neural networks fine tune the weights on the elements through learning from the errors occurring on the output. This is called training.

Chapter 5 - Experimental Results and Analysis

In this chapter, the implementation of the methodology described in the previous chapter along with the experimental results is explained. The data that was obtained was at the driver - week level from May 2013 to December 2013 from one of the centers of a food distribution company.

5.1 Linear Regression

The simplest type of regression, linear regression, was used to develop a model with MPG as the dependent variable. Among all the variables listed earlier, weight, backhaul, stop time, idle time, stop count, idle fuel, age, and experience of drivers in food service and in the company under discussion can be controlled and are truly independent. Hence, only these variables were used as the independent variables to build the model and predict the dependent variable, which is MPG.

The linear regression method provided the following output:

$$\begin{aligned} MPG = 0.304 * Age + 0.0528 * Exp_{company} - 0.0014 * Exp_{foodservice} - 1.82 \\ * Stopcount - 0.913 * Stoptime + 0.0029 * Idletime + 0.0032 \\ * Idlefuel - 0.0031 * Weight - 0.0029 * Backhaul \\ + 2.391 \end{aligned} \tag{5-1}$$

The model was validated using the data from January 2014 - March 2014 (first two weeks). The histogram represents the percentage of data within the error groups.

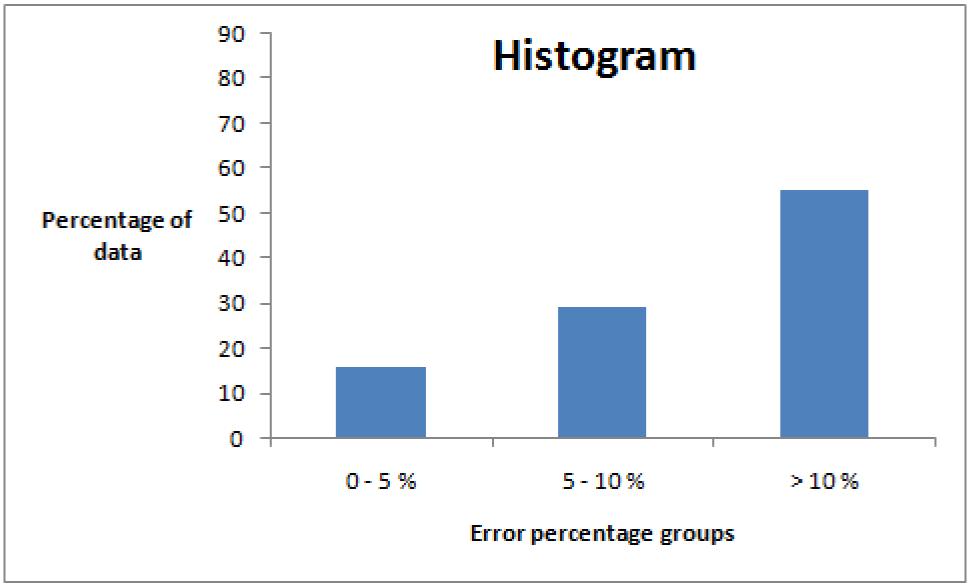


Figure 5.1 Percentage of data in different error groups for linear regression

The R-squared (adjusted) value was 61.3 %. This means that only 61.3% of the variation in the MPG can be explained by the current model. Also, $Exp_{foodserivce}$, idle time, and idle fuel had a p-value greater than 0.05. This means that these three variables are insignificant in terms of having an influence on the MPG. Statistically speaking, null hypothesis that the co-efficients are zero for the above three variables is not rejected. The lack of fit p-value of 0.001 (less than 0.05) indicates that the linear predictors are not sufficient for predicting the MPG accurately. The mean absolute percentage error was 15.52%. All the above results indicated the current linear model performs poorly in predicting the MPG values and the error values would be high.

5.2 Stepwise Linear Regression

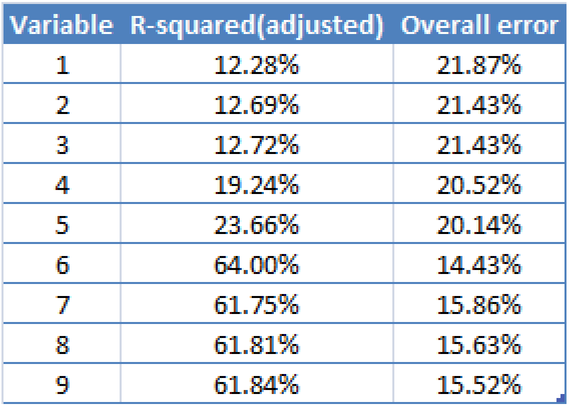
The next method to be used is the stepwise linear regression. Forward stepwise regression was used with an alpha_to_enter and alpha_to_remove value of 0.05. Weight, backhaul, stop time, idle time, stop count, idle fuel, age, and experience of drivers in food service and in the company under discussion are the variables that go into the model.

$$\begin{aligned} MPG = {} & 3.421 * Age + 0.0924 * Exp_{company} \\ & - 1.562 * Stopcount - 1.014 * Stoptime - 0.093 * Weight \\ & - 0.051 * Backhaul + 4.824 \end{aligned} \qquad (5\text{-}2)$$

This method automatically includes only the most significant variables, which means it includes those variables with a p-value < 0.05 (Bendel and Afifi, 1977). The R-squared value obtained for this data was 64%. This was not good enough when the threshold is considered as 90%. The mean absolute percentage error is 14.43%. This is slightly better than the linear regression model. The lack of fit p-value was 0.0023. This clearly showed that there was no linear relationship between the independent and the dependent variables. It also clearly showed that more information was needed in terms of the independent variables.

5.3 Best Subset

An alternate method to the stepwise regression, which is best subset method, was implemented to identify a better performing model. Weight, backhaul, stop time, idle time, stop count, idle fuel, age, and experience of drivers in food service and in the company under discussion were the variables that went into the model. Various combinations of the predictor (independent) variables were used to find the model with the highest R-squared (adjusted) value. The method starts with one variable and finishes when all the variables are used. Table 5.1 provides the R-squared (adjusted) value.



Variable	R-squared(adjusted)	Overall error
1	12.28%	21.87%
2	12.69%	21.43%
3	12.72%	21.43%
4	19.24%	20.52%
5	23.66%	20.14%
6	64.00%	14.43%
7	61.75%	15.86%
8	61.81%	15.63%
9	61.84%	15.52%

Table 5.1 R-squared (adjusted) values for best subset

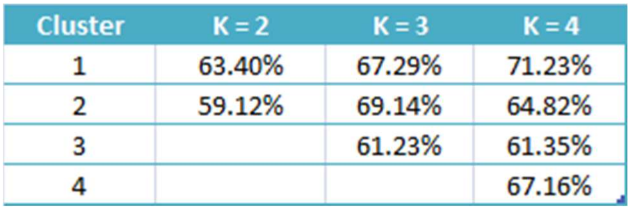
This table shows that the model with 6 variables, which must be the one that was obtained from the stepwise regression method, was the best model with the highest R-squared (adjusted) value.

$$\begin{aligned} MPG = {} & 3.421 * Age + 0.0924 * Exp_{company} \\ & - 1.562 * Stopcount - 1.014 * Stoptime - 0.093 * Weight \\ & - 0.051 * Backhaul + 4.824 \end{aligned} \tag{5-3}$$

This model also had the least overall error. From this it can be confirmed that the best subset method is redundant if the stepwise method is already performed.

5.4 K-means Clustering

In all the methods of regression explained earlier, it was clear that the R-squared (adjusted) values very low. It can be inferred that the models that were developed do not capture the majority of variations in the MPG. The models required more information to detect these variations. The need for more information also leads to the fact that the division of data into smaller chunks would capture small variations which would not be the case when the model is built on a larger data. Cluster analysis was used to divide the data. This was done using a combination of K-means cluster analysis and linear regression. Weight, backhaul, stop time, idle time, stop count, idle fuel, age, and experience of drivers in food service and in the company under discussion are the variables that go into the model. The process started with a K-value of 2, which is the smallest possible number of clusters. The highest K-values used was 4 since any value greater than that produced with very small sized clusters with rows of value below 100 out of 2018, which is not feasible. When there were 9 clusters a stepwise linear model was built on each one of them. Table 4.2 represents the R-squared (adjusted) values for all the clusters.



Cluster	K = 2	K = 3	K = 4
1	63.40%	67.29%	71.23%
2	59.12%	69.14%	64.82%
3		61.23%	61.35%
4			67.16%

Table 5.2 R-squared (adjusted) values for K-means clustering

It can be clearly seen that the R-squared (adjusted) values are low compared to the 90% threshold value. This tells us that the linear models built on these clusters do not perform better or in other words, do not capture complete variation in the MPG. The model with the highest R-squared value (71.23%) had an overall error of about 14.12%. As we can see, it has slightly improved compared to the linear and stepwise models, but is still producing predictions with high error percentages.

5.5 Implementation of Modifications

There were some modifications that were performed on the trucks, which were lesser tire rolling resistance and a smaller gap between the truck trailer and the cabin. Tire rolling resistance is the amount of force that resists the movement of the truck in the forward direction. It exerts an opposite force in the reverse direction. So, higher the rolling resistance, greater is the amount of energy utilized by the engine to drive the truck forward. This brings down the fuel efficiency of the truck. The next modification which was lessening the gap between the truck trailer and the cabin resulted in a decrease in the aerodynamic drag due to the wind flowing in the gap. High gap length decreases the fuel efficiency. Both these modifications tend to improve the performance of the truck in terms of fuel efficiency, and were tested statistically. These modifications were performed during July 2013. In the Figure 4.2, there are two simultaneous trends depicted. The blue line represents the MPG performance for 2012. Since the Dallas center was started during April 2012, the data points start from week 16. The red line represents the MPG performance for 2013. It can be clearly seen that the MPG performance in 2012 was much better than it was in 2013 up until week 33. This is around the same time when the modifications were introduced for the trucks.

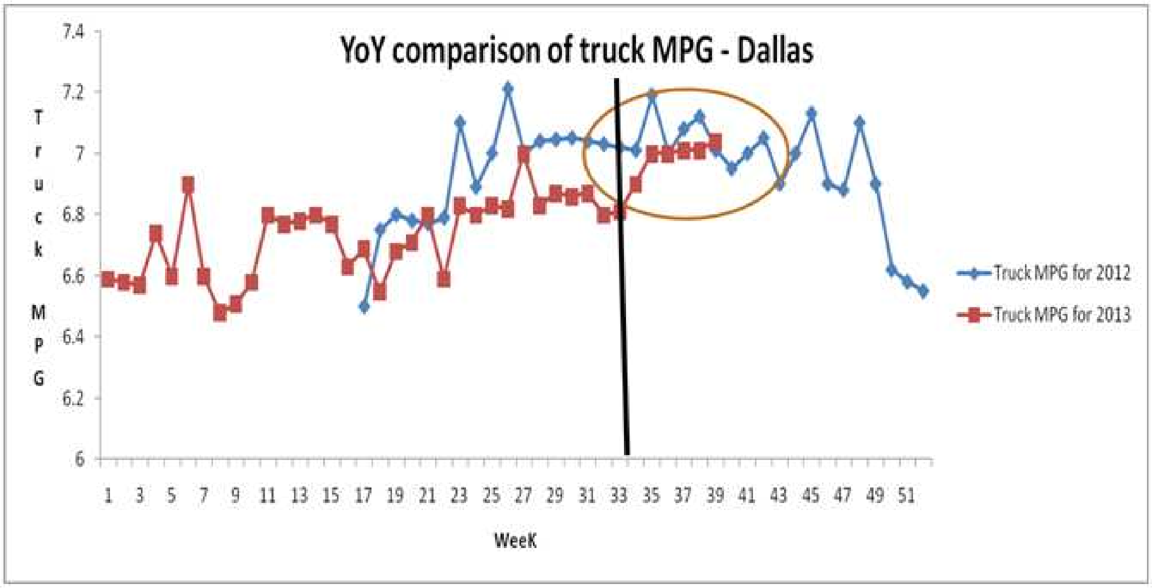
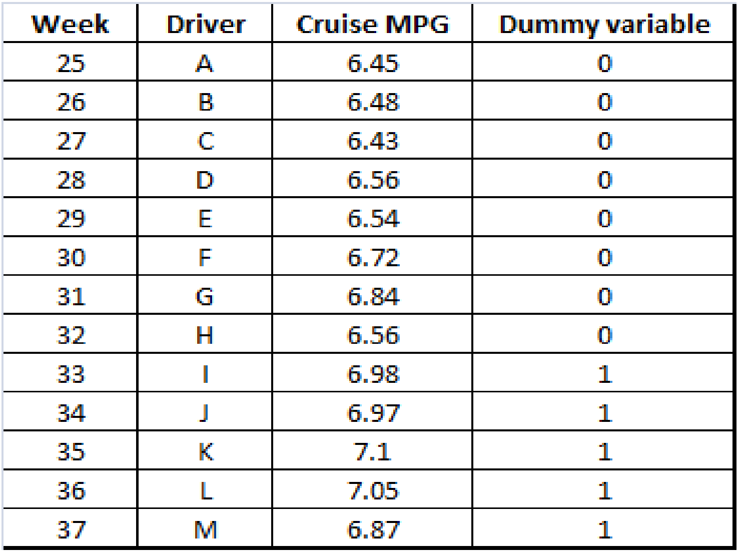


Figure 5.2 YoY comparison of truck MPG

It was necessary to test for the statistical significance of these modifications on the MPG performance. A hypothesis test was performed to find any significant difference in the variables including the MPG between dataset of June-July 2013 and the dataset of August-September 2013. There was a statistical difference in the MPG, but no such difference in the other variables. This suggested that the modifications might have had an influence on the MPG.

This was included in the dataset in the form of an encoded variable. If the MPG was for any week after week 33, then to denote the inclusion of modifications, the value of the encoded variable was 1 and zero otherwise.



Week	Driver	Cruise MPG	Dummy variable
25	A	6.45	0
26	B	6.48	0
27	C	6.43	0
28	D	6.56	0
29	E	6.54	0
30	F	6.72	0
31	G	6.84	0
32	H	6.56	0
33	I	6.98	1
34	J	6.97	1
35	K	7.1	1
36	L	7.05	1
37	M	6.87	1

Table 5.3 Sample illustration of the encoded variable

After the inclusion of the encoded variable, a regression model had to be built to examine its effect. This was done in four ways, which were i) A general linear regression model (since it includes a binary variable), ii) Stepwise linear regression, iii) Best subset method, iv) A combination of K-means cluster analysis and stepwise linear regression. The results for all the methods are shown below.

General linear regression model:

$$MPG = 0.197 \ * Age + 0.0318 * Exp_{company} - \ 0.0028 * Exp_{foodservice} - 0.41 * Stopcount - \ 0.584 * Stoptime + 0.001 * Idletime + 0.00193 * Idlefuel - 0.323 * Weight \ - 0.0145 * Backhaul - 0.081 \ * \ Encoded\_variable \ + \ 4.98 \qquad (5\text{-}4)$$

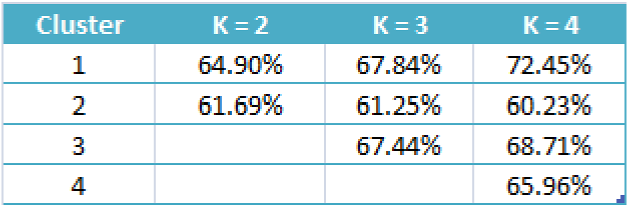
This model gave a R-squared (adjusted) value of 64.8%. The p-value for $\mathrm{Exp}_{\mathrm{foodservice}}$, $\mathrm{Exp}_{\mathrm{company}}$, idle time, idle fuel, backhaul, and encoded variable are greater than 0.05. Also, the lack of fit p-value was 0.0021. This suggested that the current linear combination of variables do not perform well in predicting the MPG.

Stepwise regression model:

$$MPG \ = \ 2.391 * \ Age \ \ - \ 1.562 \ * \ Stopcount \ - \ 1.014 \ * \ Stoptime - 0.093 \ * \ \ Weight \ + \ 4.824 \qquad (5\text{-}5)$$

The R-squared (adjusted) value was 64.1%, which was not much different from the linear model built in the previous method. Another important take away from this model was that the encoded variable, which was thought to be influential on the MPG, was not as statistically significant as it was thought before.

K-means clustering: Table 5.4 shows the R-squared (adjusted) values for the models on all 9 clusters.



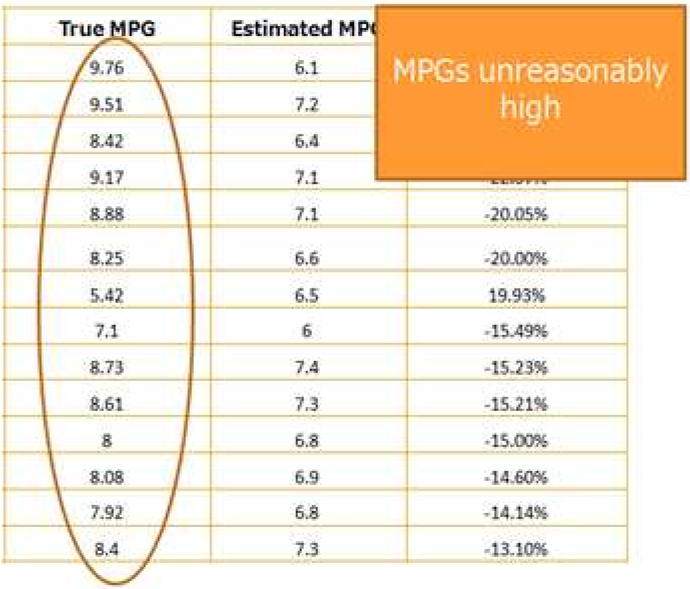
Cluster	K = 2	K = 3	K = 4
1	64.90%	67.84%	72.45%
2	61.69%	61.25%	60.23%
3		67.44%	68.71%
4			65.96%

Table 5.4 K-means cluster analysis with R-squared (adjusted values) for model with encoded variable

This also showed that the encoded variable does not improve the performance of the model. This problem suggested that the variations in the data were not completely being captured, which was why the R-squared values were low. In other words, it meant that the model did not account for the complete behavior of the data. All these issues indicated the need for the data at a more granular level. The daily level data was more granular than the weekly level data and hence, it was extracted in the same way as the weekly level data.

5.6 Anomalies in the Daily Level Data

The daily level data was obtained from May to December 2013. While examining the data, there were some anomalies that were found, as shown in Table 5.5.

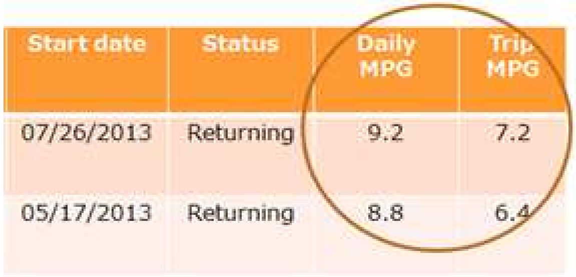


MPGs unreasonably high

True MPG	Estimated MP	
9.76	6.1	
9.51	7.2	
8.42	6.4	
9.17	7.1	[illegible]
8.88	7.1	-20.05%
8.25	6.6	-20.00%
5.42	6.5	19.93%
7.1	6	-15.49%
8.73	7.4	-15.23%
8.61	7.3	-15.21%
8	6.8	-15.00%
8.08	6.9	-14.60%
7.92	6.8	-14.14%
8.4	7.3	-13.10%

Table 5.5 Depiction of erroneous MPGs at daily level

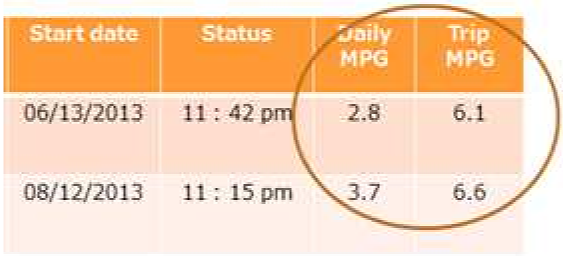
The MPGs for some of the drivers were unreasonably high. Normally, the MPGs were found to hover between 5 and 7.5. Anything beyond this boundary suggests that the numbers were incorrect. When this issue was investigated, the following results were obtained. In Table 5.6, the daily MPGs are very high compared to the MPG for the trips. This is because when the truck is returning, it is almost empty because the goods have already been delivered. This increases the MPG values drastically. Sometimes the returning process will happen on a single day and hence, the daily MPG for that day will be very high.



Start date	Status	Daily MPG	Trip MPG
07/26/2013	Returning	9.2	7.2
05/17/2013	Returning	8.8	6.4

Table 5.6 High daily MPGs

Similarly, there were low MPGs < 5 found in the data. The reason is that when the trip had just started, for example at 11:42 pm, as shown in Table 5.7, the truck is fully loaded. If the goods are not delivered before midnight, the weight will be very high, and so the MPGs will be very low.



Start date	Status	Daily MPG	Trip MPG
06/13/2013	11 : 42 pm	2.8	6.1
08/12/2013	11 : 15 pm	3.7	6.6

Table 5.7 Low daily MPGs

Both these cases suggested that the data needed to be at the trip level and not the daily level to have the correct picture of the MPG performance. As a result of this realization, trip level data was extracted. The trip level data was taken from May 2013 to December 2013. Weight, backhaul, stop time, idle time, stop count, idle fuel, age, and experience of drivers in food service and in the company under discussion are the variables that go into the model.

General linear regression model:

$$MPG = 0.146 * Age + 0.0121 * Exp_{company} - 0.0004 * Exp_{foodservice} - 0.932 * Stopcount - 0.414 * Stoptime - 0.143 * Idletime - 0.258 * Idlefuel - 0.051 * Weight - 0.0097 * Backhaul - 0.025 * Encoded\_variable + 2.507 \quad (5\text{-}6)$$

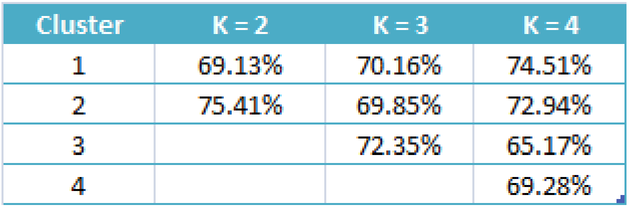
This model gave a R-squared (adjusted) value of 74.35%. The p-value for $\text{Exp}_{\text{foodservice}}$, $\text{Exp}_{\text{company}}$ and encoded_variable were greater than 0.05. Also, the lack of fit p-value was 0.0013. The R-squared (adjusted) value had certainty increased from 64% to 74.35%. This showed the effect of changing the level of data to the trip level.

Stepwise regression model:

$$MPG = 3.491 * Age - 1.389 * Stopcount - 2.352 * Stoptime - 0.097 * Weight - 0.006 * Backhaul - 0.194 * Idletime - 0.002 * Idlefuel + 9.42 \quad (5\text{-}7)$$

The R-squared (adjusted) value is 76.89%. The model is certainly performing better. However, the lack of fit value is 0.004, which is less than 0.05. This again shows that a linear combination of the variables does not capture the complete variation of the MPG.

K-means clustering: The following figure shows the R-squared (adjusted) values for the models on all 9 clusters. It can be seen that the R-squared values have improved, but are still below the 90% threshold.



Cluster	K = 2	K = 3	K = 4
1	69.13%	70.16%	74.51%
2	75.41%	69.85%	72.94%
3		72.35%	65.17%
4			69.28%

Table 5.8 R-squared (adjusted) values for K-means cluster analysis at trip level

It can be clearly seen that the R-squared (adjusted) values are low. This means that the predictions obtained were not very close to the actual values.

5.7 Polynomial and Nonlinear Regression

As observed earlier, any method that involves a linear model for the data does not provide a good fit and a good performing model. The next step was to look at the nonlinear relationships and polynomial relationships between the variables.

The different combinations of variables tried were:

1. Quadratic terms

2. Cubic terms

3. Cross terms

4. Inverse terms

5. Log-log transformations

There were 1083 variations of models for all the above possible methods. The R-squared adjusted value is not the appropriate performance measure for nonlinear models. Histograms were built for all the models, which depicted the percentage of data in each error groups (0-5%, 5-10%, and >10%). The following model had the highest percentage of data in the 0-10% error group, and thus had the least errors and was the best model.

$$LN(MPG) = {}^{1}\Big/\left(\begin{array}{c}(2.13e+17 * Stopcount) + (1.04e+17 * Stoptime) \\ - (5.12e+17 * Weight)\end{array}\right) + 4.926 \qquad (5\text{-}8)$$

This shows that stop count, stop time, and weight had the highest impact on the MPG and were the most influential factors on the MPG. The histogram for the percentage of data in the error groups is shown below.

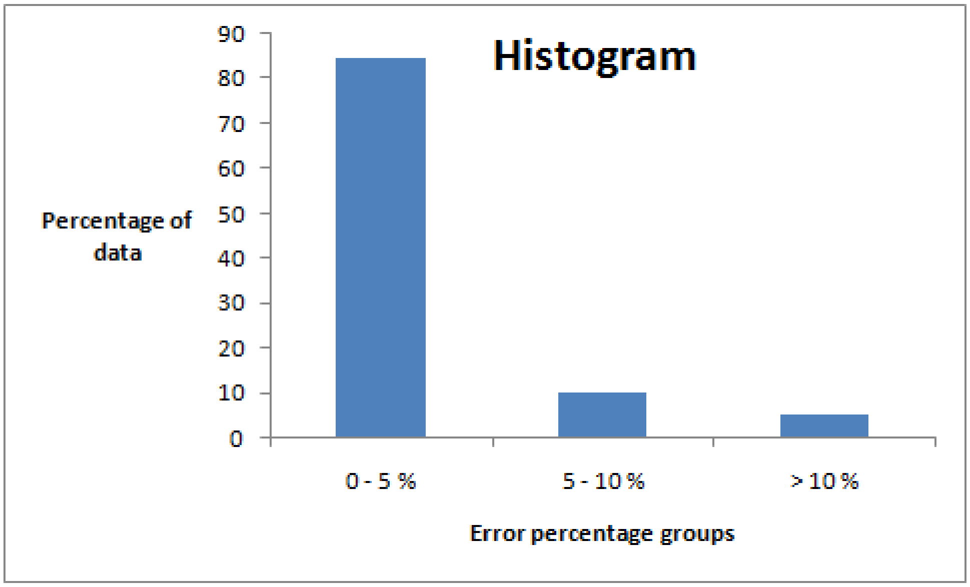


Figure 5.3 Percentage of data in the error groups for nonlinear model

The histogram shows that for the data from May 2013 through December 2013, on which the model was built, around 84% of the data is within the 0-5% error group. When the model was validated for the data from Jan 2014 through March 2014 (first two weeks), the histogram appears as follows.

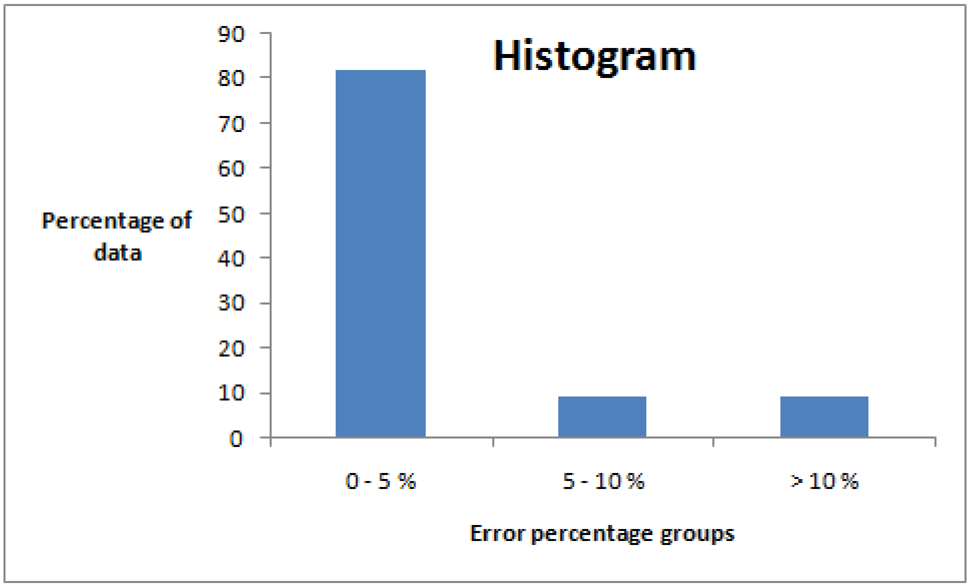
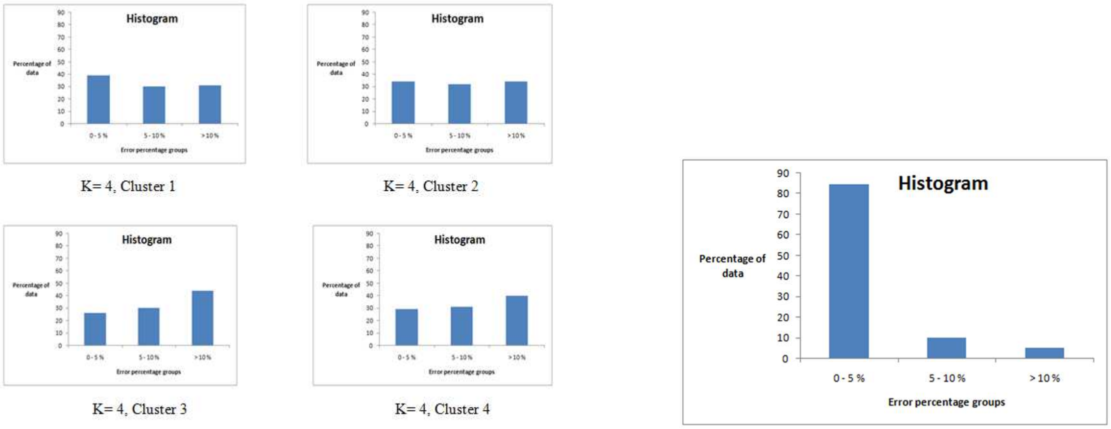


Figure 5.4 Percentage of data in the error groups for nonlinear related to validation data

It can be observed that a high percentage of the data is present in the least error band and hence the model had performed better.

The error charts were developed and compared with the results from clustering. It can be clearly seen that the performance of the model was improved by using a nonlinear method as opposed to clustering. The percentage of data within the 0-5% error group is around 34% for clustering, whereas it is around 84% for the nonlinear model. This indicates nonlinearity in the interaction between the independent variables and the MPG. The major improvement took place in the overall error, which is as low as 10.05%, much lower than the 14% for the clustering method.



K= 4, Cluster 1

K= 4, Cluster 2

K= 4, Cluster 3

K= 4, Cluster 4

Figure 5.5 Comparison of histograms of clustering and nonlinear model

5.8 Sensitivity Analysis

Once the model is obtained, it must be validated. Hence, sensitivity analysis techniques were used to test the robustness of the model in extreme conditions. The two main techniques used were percentage sensitivity and maximum to minimum variation.

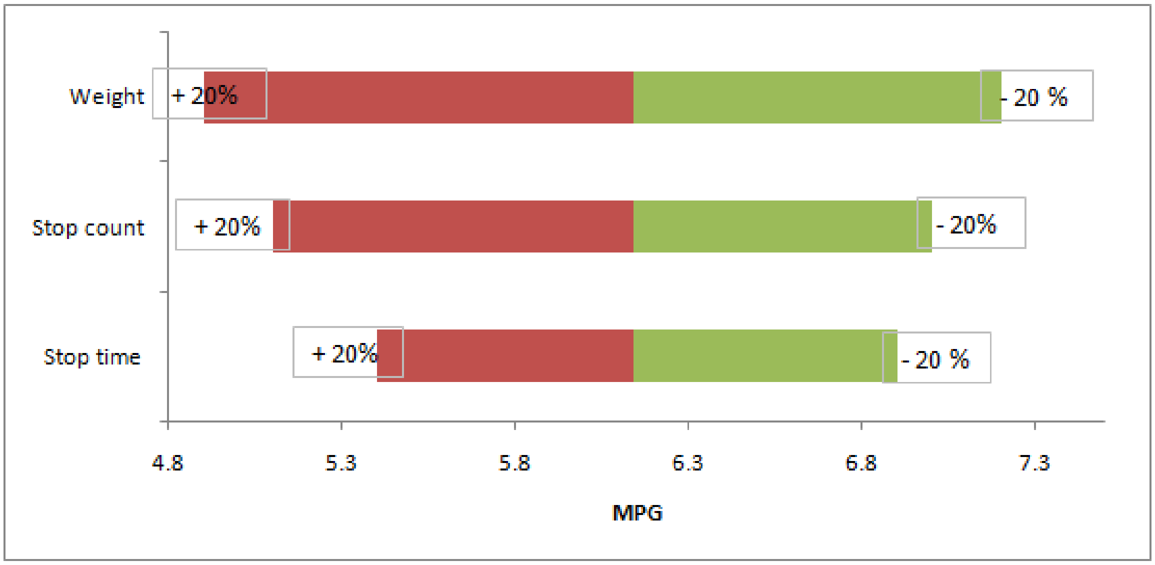


Figure 5.6 Percentage sensitivity analysis

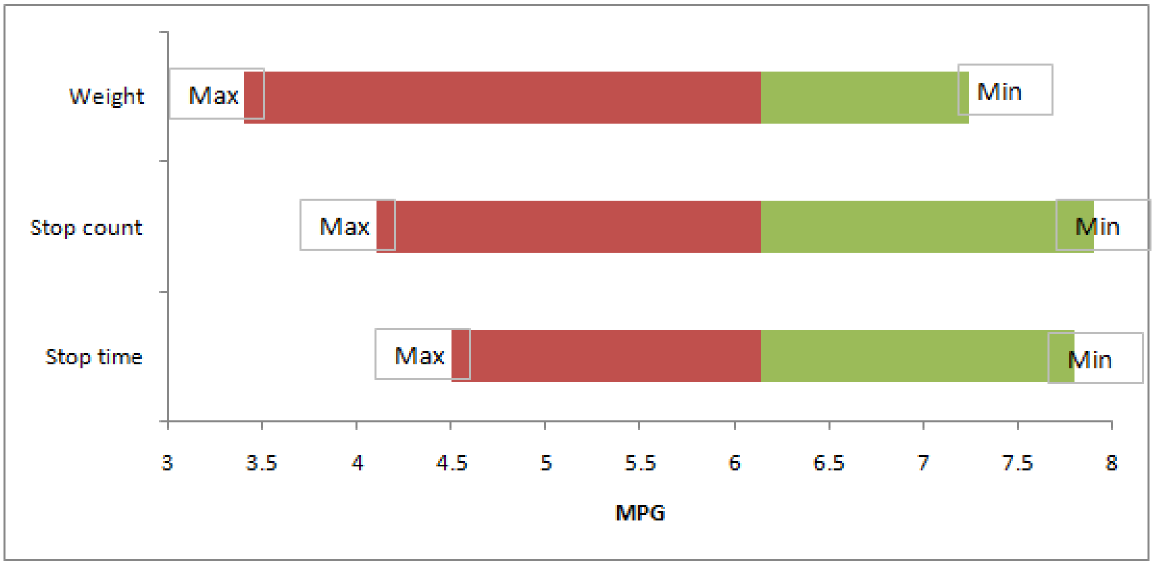
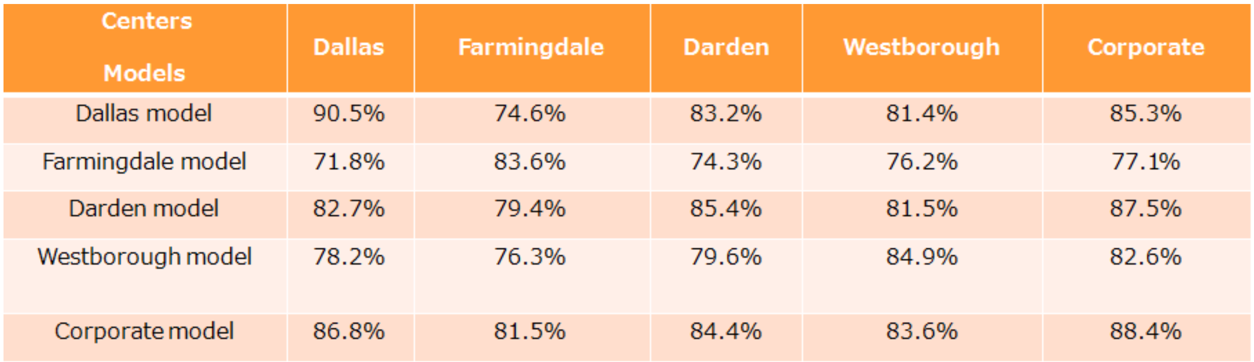


Figure 5.7 Maximum to minimum variation

The above figures represent the variation of the MPG when a particular variable is varied from the lowest to the highest values. The figure suggests that in both the methods, weight has the highest influence on the MPG. In other words, the MPG is more sensitive to the weight. It also suggests an inverse relationship between the MPG and the independent factors.

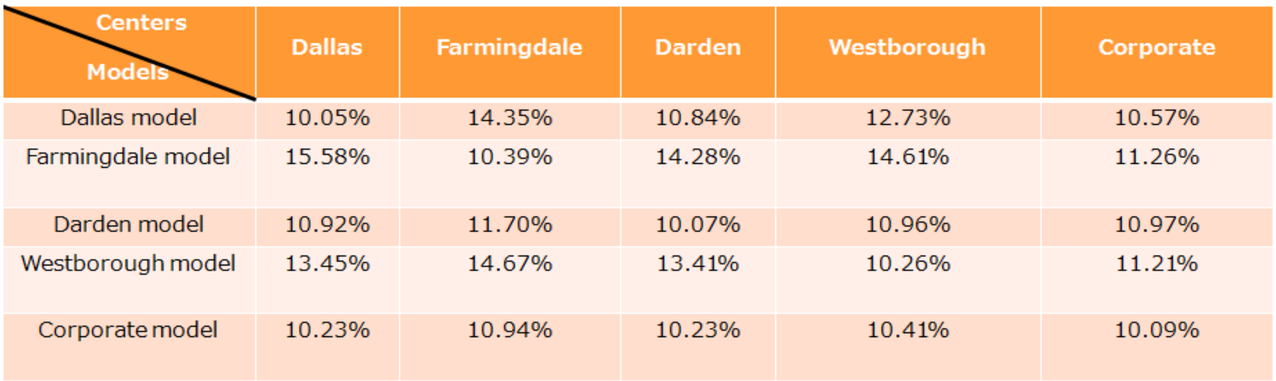
5.9 Most Consistent Model

Consistency is necessary so that a model that is the most generic can be used for predictions. The data was collected from five centers - Dallas, Darden, Westborough, Farmingdale and Corporate. The timeline was from May 2013 to Dec 2013. Five different models were built using the data from the five centers. These models were tested across the different centers, to obtain the most consistent model. After comparing the five models, it was observed that the Corporate model, as shown in Table 4.9 was the most consistent one.



Centers Models	Dallas	Farmingdale	Darden	Westborough	Corporate
Dallas model	90.5%	74.6%	83.2%	81.4%	85.3%
Farmingdale model	71.8%	83.6%	74.3%	76.2%	77.1%
Darden model	82.7%	79.4%	85.4%	81.5%	87.5%
Westborough model	78.2%	76.3%	79.6%	84.9%	82.6%
Corporate model	86.8%	81.5%	84.4%	83.6%	88.4%

Table 5.9 Percentage of data within 10% error



Centers Models	Dallas	Farmingdale	Darden	Westborough	Corporate
Dallas model	10.05%	14.35%	10.84%	12.73%	10.57%
Farmingdale model	15.58%	10.39%	14.28%	14.61%	11.26%
Darden model	10.92%	11.70%	10.07%	10.96%	10.97%
Westborough model	13.45%	14.67%	13.41%	10.26%	11.21%
Corporate model	10.23%	10.94%	10.23%	10.41%	10.09%

Table 5.10 Average percentage errors for all five centers and models

The analysis showed that stop time, stop count, and weight were the most significant factors that affected the MPG. These factors can be controlled and can be used to study predictions of the MPG. There was an 88.4% accuracy for the predictions when compared with the actual MPG values, which means that 88.4% of the data was within the 10% error band.

$$LN(MPG) = {}^{1}\Big/\left(\begin{array}{c}(1.76e + 17 \ast Stopcount) + (1.24e + 17 \ast Stoptime) \\ - (4.67e + 17 \ast Weight)\end{array}\right) + \quad 5.237 \tag{5-9}$$

It was also seen that the Corporate model produced similar errors at the trip level for all the centers as shown in Table 5.10. Hypothesis tests were conducted between the error sets of all the centers and it was found that there was no difference in the errors statistically. This proved that the Corporate model was indeed the most consistent. The individual errors at the trip level were calculated for the Corporate model on the datasets of all five centers. The overall average error was around 10%. It was found that the maximum error at the trip level was 18.57%.

5.10 Neural Network

The neural network training was done using MATLAB Neural Network Toolbox. The training process is the most sensitive and time consuming part of the algorithm. A well trained network gives good results given any input. A not well trained network on the other hand will give erratic results if the inputs given are different from the input used to train. Because the datasets have a significant size and because the network also needs inputs to validate and test the network, the proper division of the data set would have to be made. Ideally, the bigger the training dataset, the more accurate the network is. In practice a network with a too big training dataset will predict very well the relation between the output and trained input but will give erroneous values for inputs not included in the training dataset. This process is called over fitting and it occurs when the network is too trained.

The independent variables – weight, stop time and stop count from the best nonlinear model for Corporate center were included as inputs to the neural network. The next task was to find out the number of hidden layers in the neural network and the number of activation units in each layer. The training data was taken from the Corporate center because the most consistent nonlinear model was built on this data between May 2013 and Dec 2013. The testing data was from Jan 2014 to March 2014 (first two weeks).

5.10.1 Number of activation units in a layer

For the purpose of this experiment, there was one hidden layer included with number of activation units increasing from 2 until the root mean squared error for the training data or the testing data increased beyond the errors obtained from the nonlinear model.

Number of Activation Units	2	3	4	5	6
MAPE values for training data	8.53%	8.46%	8.41%	8.38%	8.38%
MAPE values for testing data	9.43%	9.46%	9.97%	10.07%	10.07%

Table 5.11 MAPE values for training and testing data with different number of activation units

It can be seen from Table 5.11 that as the number of activation units increased, the MAPE values for the training data decreases but the same for testing data increases and goes beyond 10% at 5 activation units .This is because of overfitting explained earlier which causes variance and does not provide accurate outputs for the testing data. So, from the above experiment, it was decided that the number of activation units in one hidden layer would be 4.

5.10.2 Number of hidden layers

This experiment included 4 activation units in each layer whose number was varied from 1 until the MAPE values increased beyond 10%.

Number of Hidden Layers	1	2	3
MAPE values for training data	8.41%	8.40%	8.35%
MAPE values for testing data	9.97%	9.97%	10.04%

Table 5.12 MAPE values for training and testing data with different number of hidden layers

It can be clearly seen from Table 5.12 that the MAPE values for training data decreases with the number of hidden layers but it clearly overfits the data at 3 layers where the MAPE values goes beyond 10%.

So, with the results of the above experiments, it was decided that the neural network architecture would include 3 input units (3 features), 2 hidden layers (4 activation units in each layer) and one output unit. This architecture gave the best predictions for the MPG and can be used to improve fuel economy.

5.10.3 Nonlinear Model versus Neural Network

Data Mining Method	Nonlinear Model	Neural Network
MAPE values for training data (May 2013 - Dec 2013)	10.09%	8.40%
MAPE values for testing data (Jan 2014 to March 2014 (first two weeks))	14.35%	9.97%

Table 5.13 MAPE values comparison between nonlinear model and neural network

To depict this pictorially, a random set of 100 samples were picked from the testing data i.e. data from Jan 2014 to March 2014. The chart below shows a comparison of predictions from both neural network and nonlinear models. It can be clearly seen that the predictions from the nonlinear model are farther from the actual data compared to the neural network predictions.

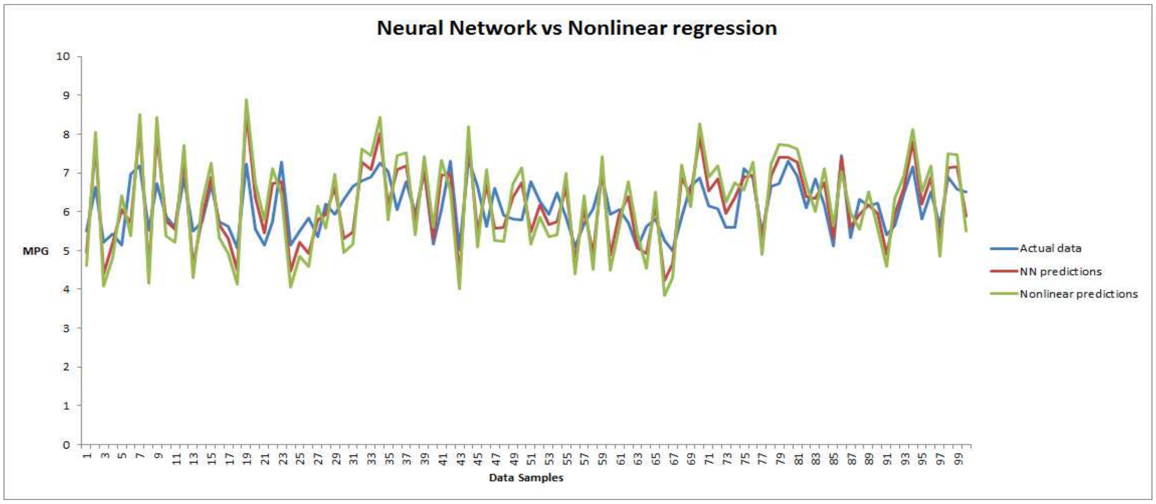


Figure 5.8 Error comparison between nonlinear model and neural network

5.11 Chapter Summary

This chapter explains the implementation of the methodology in detail with descriptions of the experiments and a discussion of results. The experimental results show that stop time, weight, and stop count are the most significant factors to affect the MPG. It is also established that there exists a nonlinear relationship between these variables and the MPG. The linear model, stepwise linear, best subset, clustering, and nonlinear regressions were the different types of methods used to build the model. It was found that the a linear model with clustering provided a model with an overall error of about 14%, while the nonlinear model provided with an overall error of around 10%. This suggests that the nonlinear model performs better than the linear models. Also, the percentage of data within the 0-5% error band is about 84%, while for clustering it was about 34%. All these results show that there exists a nonlinear relationship between the variables and the MPG.

The sensitivity analysis tests also suggest that MPG is most sensitive towards the aforementioned variables. The most consistent model across all the five centers was developed. It was found that the Corporate model was the most consistent one across all the five centers. This was found by testing for the statistical difference between the average errors of the predictions from the models on the data from all the five centers. Finally, neural network architecture was built to improve the accuracy of the nonlinear relationship which gave better performance while predicting the MPGs.

Chapter 6 - Conclusions and Future Work

6.1 Conclusions

Weight, backhaul, stop time, idle time, stop count, idle fuel, age, and experience of drivers in food service and in the company under discussion were the variables put into the model. The initial approach focused on building a linear model with linear regression, stepwise, best subset, and cluster with linear regression. The models obtained from all these approaches provided a least overall error value of 14%. In the best model, which was obtained from the results of the methodology in the previous chapter, it seems that weight, stop count, and stop time are the most significant factors to affect the MPG, or the fuel efficiency of the trucks. A consistent model that could provide relatively accurate predictions across five distribution centers - Dallas, Darden, Westborough, Farmingdale and Corporate, had to be developed. It was found that the model built using the data from the Corporate center was the most consistent one. The timeline of the data used to build the model was from May 2013 through December 2013. The predictive model provided predictions of which about 88% of the data that was used, was within the 0-10% error group. This was an improvement on the lesser 43% obtained for the linear regression and K-means clustering models. The model was also validated on the data for January 2014 through the first two weeks of March 2014 and it provided an accuracy of 81% of the data that was within the 0-10 % error group. The average overall error was around 10%, which was the least for the approaches explored in this research as shown in Table 6.1

	Linear approaches				Nonlinear
	Linear regression	Stepwise linear	Best subset	Clustering	
Overall error	15.56%	14.43%	14.43%	14.12%	10.08%

Table 6.1 Comparison of overall errors for different approaches

It is also clear from the sensitivity analysis performed that the greater the value of these variables, the lesser is the MPG. The main objective of this research was to identify the key factors that influence the MPG. The other objective was to develop a predictive model to predict the future MPG during instances of re-routing. Based on the model under discussion, if a re-routing is performed on a particular region, it is possible to estimate the change in MPG by changing the values of these variables accordingly. The model can be used to predict the average change in MPG for a set of trips whenever a re-route is performed. Since, the aim of re-routing is to reduce the miles and trips; extra load will be added to the remaining trips. Although, the MPG would decrease because of this extra load, it would be offset by the savings due to the drop in miles and trips. The net savings in the fuel can now be translated into the amount of money saved.

Since the three controllable variables which have significant impact on MPG are known, it is possible to vary their values to study the effect on MPG. This can lead to the development of setting up boundaries for the values of these variables in order to maintain the fuel efficiency at the desired level. Every year, the competition in the distribution industry increases. Companies yearn for more satisfied customers in order to increase revenue and profits. This pressurizes them to push the load capacity of the trucks to the extreme which can lead to safety issues. Thus, with the variables identified in this research, boundaries can be set up to restrict the companies from compromising on safety issues due to load capacity. An optimal solution needs to be found in order to balance fuel efficiency, profits, load capacity and safety issues. There is some discussion in the United States in relation to the surface transportation authorization bill about possibly increasing the load limits for trucks where the current mass limit is 80 000 lbs (36.3 tonnes).

6.2 Future Work

The best model that was obtained from this research had about 81% of the data within the 0-10 % error group. Ideally, this number should be close to 100%. However, a value above 90% would provide us with a better performing model. Since this area of research involves human interactions and weather conditions, many uncertainties that cannot be controlled are included (Feng and Figliozzi, 2013). However, if more controllable factors related to the engine performance of the trucks and truck specifications are available, a better performing model could be developed because a higher percentage of variations in MPG could be captured. Other variables that might be helpful are the values for the gap between the cabin and trailer, tire pressure, and tire rolling resistance (Sharpe and Roeth, 2014).

In the future, there are different kinds of data mining techniques that can be used to build a better performing model. Association rule mining can be used to study the interactions between the independent variables themselves (Aggarwal and Phillip, 1999). Once these interactions are accounted for, then the model that is obtained will provide better predictions with lower error percentages.

Decision tree method can be used to divide the dataset into different groups based on certain relationships, which are learned by the method automatically (Quinlan, 1990). Since there is no prior knowledge on how the data points are related, this would be a very appropriate method for building the model. A step ahead would be the random forest method, which is a group of decision trees that can be used to produce relationships between the variables (Yand and Gu, 2014).

Support vector machines can be used along with association rule mining to detect patterns in the dataset. It divides the dataset into clear groups with defined boundaries (Subhasi, 2013). This could help in developing different regression models in order to obtain better predicitions for the MPG.

REFERENCES

Abbass, Hussein A. "An evolutionary artificial neural networks approach for breast cancer diagnosis." *Artificial Intelligence in Medicine* 25.3 (2002): 265-281.

Acuna, Edgar, and Caroline Rodriguez. "The treatment of missing values and its effect on classifier accuracy." In *Classification, Clustering, and Data Mining Applications*, pp. 639-647. Springer Berlin Heidelberg, 2004.

Adderley, Richard, Michael Townsley, and John Bond. "Use of data mining techniques to model crime scene investigator performance." Knowledge-Based Systems 20, no. 2 (2007): 170-176.

Aggarwal, Charu C., and S. Yu Philip. "Data mining techniques for associations, clustering and classification." In *Methodologies for Knowledge Discovery and Data Mining*, pp. 13-23. Springer Berlin Heidelberg, 1999.

Agrawal R., and Psaila, G. 1995. Active Data Mining. In Proceedings of the First International Conference on Knowledge Discovery and Data mining(KDD-95), 3-8. Menlo Park, Calif.: American Association for Artificial Intelligence.

Ahmad, Iftikhar, Azween B. Abdullah, and Abdullah S. Alghamdi. "Application of artificial neural network in detection of probing attacks." In *Industrial Electronics & Applications, 2009. ISIEA 2009. IEEE Symposium on*, vol. 2, pp. 557-562. IEEE, 2009.

Amado, V. Expanding the Use of Pavement Management Data. In: Transportation Scholars Conference 2000, University of Missouri, 2000, www.ctre.iastate.edu/mtc/papers/amado.pdf (Accessed Jan 2014)

Arrieta, Angélica González, et al. "Neural Networks Applied to Fingerprint Recognition." *Distributed Computing, Artificial Intelligence, Bioinformatics, Soft Computing, and Ambient Assisted Living*. Springer Berlin Heidelberg, 2009. 621-625.

Barai, S. V., and Yoram Reich. "Weld classification in radiographic images: data mining approach." In Proceedings of National Seminar on Non Destructive Evaluation. 2002.

Barai, Sudhir Kumar. "Data mining applications in transportation engineering." *Transport* 18, no. 5 (2003): 216-223.

Bartz-Beielstein, Thomas, and Sandor Markon. "Tuning search algorithms for real-world applications: A regression tree based approach." In *Evolutionary Computation, 2004. CEC2004*. Congress on, vol. 1, pp. 1111-1118. Institute of electrical and electronics engineers, 2004.

Bendel, Robert B., and Abdelmonem A. Afifi. "Comparison of stopping rules in forward “stepwise” regression." *Journal of the American Statistical Association* 72, no. 357 (1977): 46-53.

Borio, Claudio. "Rediscovering the Macroeconomic Roots of Financial Stability Policy: Journey, Challenges, and a Way Forward." *Annual Revenue Finance Economy* 3, no. 1 (2011): 87-117.

Bradley, Paul S., and Usama M. Fayyad. "Refining Initial Points for K-Means Clustering." In *ICML*, vol. 98, pp. 91-99. 1998.

Cao, Huiping, Nikos Mamoulis, and David W. Cheung. "Discovery of periodic patterns in spatiotemporal sequences." *Knowledge and Data Engineering,* Institute of electrical and electronics engineers *Transactions* on 19, no. 4 (2007): 453-467.

Cohen, Eliot A. "Distant battles: Modern war in the third world." *International Security 10*, no. 4 (1986): 143-171.

Colin Cameron, A., and Frank AG Windmeijer. "An< i> R</i>-squared measure of goodness of fit for some common nonlinear regression models." *Journal of Econometrics* 77, no. 2 (1997): 329-342.

Czitrom, Veronica. "One-factor-at-a-time versus designed experiments." *The American Statistician 53*, no. 2 (1999): 126-131.

Dantzig, G.B. and Ramser, J.H. (1959), “The truck dispatching problem”, Management Science, Vol. 6 No. 1, pp. 80-91.

Deng, Jiamei, Richard Stobart, and Bastian Maass. "The Applications of Artificial Neural Networks to Engines."

Edelkamp, Stefan, and Stefan Schrödl. "Route planning and map inference with global positioning traces." In *Computer Science in Perspective*, pp. 128-151. Springer Berlin Heidelberg, 2003.

Fan, Wenhui, Huayu Xu, and Xin Xu. "Simulation on vehicle routing problems in logistics distribution." *COMPEL: The International Journal for Computation and Mathematics in Electrical and Electronic Engineering 28*, no. 6 (2009): 1516-1531.

Fayyad, U. M.; Piatetsky-Shapiro, G.; Smyth, P. and Uthursamy, R. Advances in knowledge discovery and data mining, Association for the Advancement of Artificial Intelligence Press/The MIT Press, Cambridge, MA, 1996.

Foslien, Wendy, Valerie Guralnik, and Karen Zita Haigh. "Data Mining For Space Applications." In *Eighth International Conference on Space Operations*, p. E8. 2004.

Frey, H. Christopher, Amirhossein Mokhtari, and Tanwir Danish. "Evaluation of selected sensitivity analysis methods based upon applications to two food safety process risk models." Dept. of Civil, Construction, and Environmental Eng., North Carolina State Univ., Raleigh, NC (2003).

Godek, Paul E. "Regulation of Fuel Economy and the Demand for Light Trucks, The." *JL & Econ. 40* (1997): 495.

González, Pamela Castellón, and Juan D. Velásquez. "Characterization and detection of taxpayers with false invoices using data mining techniques." *Expert Systems with Applications* (2012).

Gray, J. R., C. Homescu, L. R. Petzold, and R. C. Alkire. "Efficient Solution and Sensitivity Analysis of Partial Differential-Algebraic Equation Systems Application to Corrosion Pit Initiation." *Journal of The Electrochemical Society 152*, no. 8 (2005): B277-B285.

Greene, David L., James Kahn, and Robert Gibson. "Fuel economy rebound effect for US household vehicles." *Energy Journal-Cambridge MA Then Cleveland OH- 20* (1999): 1-31.

Hall, J.; Mani, G.; and Barr, D. 1996. Applying Computational Intelligence to the Investment Process. In *Proceedings of CIFER-96: Computational Intelligence in Financial Engineering. Washington, D.C.*: Institute of Electrical and Electronics Engineers Computer Society.

Hamby, D. M. "A comparison of sensitivity analysis techniques." *Health Physics 68*, no. 2 (1995): 195-204.

Hartigan, John A., and Manchek A. Wong. "Algorithm AS 136: A k-means clustering algorithm." *Applied statistics* (1979): 100-108.

Hasheem, Soukaena Hassan. "An Association Rules Analysis to enhance solving the Congestion Problem." *Journal of Emerging Trends in Computing and Information Sciences 2*.

Hayashi K, Yano Y, "Future city logistics in Japan from the shippers' and carriers' view - prospects and recent measures to develop them." *Logistics systems for sustainable cities proceedings of the 3rd International conference on city logistics* (Madeira, Portugal, 25 June, 2003), pp. 263-277.

He, Chao, Yunshan Ge, Chaochen Ma, Jianwei Tan, Zhihua Liu, Chu Wang, Linxiao Yu, and Yan Ding. "Emission characteristics of a heavy-duty diesel engine at simulated high altitudes." *Science of the Total Environment 409*, no. 17 (2011): 3138-3143.

Helton, Jon C., Jay Dean Johnson, Cedric J. Sallaberry, and Curt B. Storlie. "Survey of sampling-based methods for uncertainty and sensitivity analysis." *Reliability Engineering & System Safety 91*, no. 10 (2006): 1175-1209.

Ji, Qiang, Zhiwei Zhu, and Peilin Lan. "Real-time nonintrusive monitoring and prediction of driver fatigue." *Vehicular Technology, Institute of Electrical and Electronics Engineers Transactions* on 53, no. 4 (2004): 1052-1068.

Kanungo, Tapas, David M. Mount, Nathan S. Netanyahu, Christine D. Piatko, Ruth Silverman, and Angela Y. Wu. "An efficient k-means clustering algorithm: Analysis and implementation." *Pattern Analysis and Machine Intelligence, Institute of Electrical and Electronics Engineers Transactions on 24*, no. 7 (2002): 881-892.

Koh, Hian Chye, and Gerald Tan. "Data mining applications in healthcare." *Journal of Healthcare Information Management—Vol 19*, no. 2 (2011): 65.

Kohavi, Ron, Dan Sommerfield, and James Dougherty. "Data Mining Using a Machine Learning Library in C++." *International Journal on Artificial Intelligence Tools 6*, no. 04 (1997): 537-566.

Kovács, Gábor. "Possible methods of application of electronic freight and warehouse exchanges in solving the city logistics problems." *Periodica Polytechnica: Transportation Engineering 38*, no. 1 (2010): 25-28.

Kusiak, A. "Data mining in manufacturing: a review." *Journal of Manufacturing Science and Engineering 128*, no. 4 (2006): 969-976.

Kydes, Andy S. "Energy intensity and carbon emission responses to technological change: the US outlook." *The Energy Journal 3* (1999): 93-121.

Li, Shing-Han, David C. Yen, Wen-Hui Lu, and Chiang Wang. "Identifying the signs of fraudulent accounts using data mining techniques." *Computers in Human Behavior 28*, no. 3 (2012): 1002-1013.

Liao, Shu-Hsien, Pei-Hui Chu, and Pei-Yuan Hsiao. "Data mining techniques and applications–A decade review from 2000 to 2011." *Expert Systems with Applications 39*, no. 12 (2012): 11303-11311.

Little, Roderick JA, and Donald B. Rubin. "The analysis of social science data with missing values." *Sociological Methods & Research 18*, no. 2-3 (1989): 292-326.

Loh, Wei-Yin. "Classification and regression trees." Wiley Interdisciplinary Reviews: *Data Mining and Knowledge Discovery 1*, no. 1 (2011): 14-23.

Mannila, H.; Toivonen, H.; and Verkamo, A.I. 1995. Discovering Frequent Episodes in Sequences. In *Proceedings of the First International Conference on Knowledge Discovery and Data Mining(KDD-95)*, 210-215. Menlo Park, Calif.: American Association for Artificial Intelligence

Ngai, E. W. T., Yong Hu, Y. H. Wong, Yijun Chen, and Xin Sun. "The application of data mining techniques in financial fraud detection: A classification framework and an academic review of literature." *Decision Support Systems 50*, no. 3 (2011): 559-569.

Passing, H., and W. Bablok. "A new biometrical procedure for testing the equality of measurements from two different analytical methods. Application of linear regression procedures for method comparison studies in clinical chemistry, Part I." *Clinical Chemistry and Laboratory Medicine 21*, no. 11 (1983): 709-720.

Piatetsky-Shapiro, Gregory, Ronald J. Brachman, Tom Khabaza, Willi Kloesgen, and Evangelos Simoudis. "An Overview of Issues in Developing Industrial Data Mining and Knowledge Discovery Applications." In *KDD*, vol. 96, pp. 89-95. 1996.

Pisters, Patricia. "Logistics of perception 2.0: multiple screen aesthetics in Iraq War films." *Film-Philosophy 14*, no. 1 (2010): 232-252.

Pradhan, Biswajeet. "A comparative study on the predictive ability of the decision tree, support vector machine and neuro-fuzzy models in landslide susceptibility mapping using GIS." *Computers & Geosciences* (2012).

Privé, Julie, Jacques Renaud, Fayez Boctor, and Gilbert Laporte. "Solving a vehicle-routing problem arising in soft-drink distribution." *Journal of the Operational Research Society 57*, no. 9 (2006): 1045-1052.

Quinlan, J. Ross. "Decision trees and decision-making." *Systems, Man and Cybernetics, Institute of Electrical and Electronics Engineers Transactions on 20*, no. 2 (1990): 339-346.

Saltelli, Andrea, and Paola Annoni. "How to avoid a perfunctory sensitivity analysis." *Environmental Modelling & Software 25, no. 12* (2010): 1508-1517.

Singh, Satnam, Halasya Siva Subramania, Steven W. Holland, and Jason T. Davis. "Decision Forest for Root Cause Analysis of Intermittent Faults." *Systems, Man, and Cybernetics, Part C: Applications and Reviews, Institute of Electrical and Electronics Engineers Transactions on 42*, no. 6 (2012): 1818-1827.

Solberg, J.J. (1977), “A mathematical model of computerized manufacturing systems”, Proceedings of the 4th International Conference on Production Research, Tokyo, Japan, pp. 1265-75.

Spencer, Kevin Simões. *Fuel Consumption Optimization using Neural Networks and Genetic Algorithms*. Diss. Master Thesis, Instituto Superior Técnico, Lisbon, Portugal, 2011.

Stergiou, Christos, and Dimitrios Siganos. "Neural Networks. 1996." (2010).

Subasi, Abdulhamit. "A decision support system for diagnosis of neuromuscular disorders using DWT and evolutionary support vector machines." *Signal, Image and Video Processing* (2013): 1-10.

Ulrich, Keith. "DHL Open Innovation: Program for the Development, Deployment and Promotion of Innovative Solutions in Logistics." In *Strategies and Communications for Innovations*, pp. 305-317. Springer Berlin Heidelberg, 2011.

van Duin J H R, Kneyber J C, "Towards a matching system for the auction of transport orders". Logistics systems for sustainable cities proceedings of the 3rd International conference on city logistics (Madeira, Portugal, 25 June, 2003), pp. 163-177

Vernon, David, and Alan Meier. "Identification and quantification of principal–agent problems affecting energy efficiency investments and use decisions in the trucking industry." *Energy Policy* (2012).

Wong, Y. K., and W. L. Woon. "An iterative approach to enhanced traffic signal optimization." *Expert Systems with Applications 34*, no. 4 (2008): 2885-2890.

Yang, Wei, and C. Charles Gu. "Random forest fishing: a novel approach to identifying organic group of risk factors in genome-wide association studies." *European Journal of Human Genetics 22*, no. 2 (2014): 254-259.