PREDICTION OF FUEL CONSUMPTION OF LONG HAUL HEAVY DUTY TRUCKS USING MACHINE LEARNING AND COMPARISON OF THE PERFORMANCE OF VARIOUS LEARNING TECHNIQUES

A thesis submitted to the Delft University of Technology in partial fulfilment of the requirements for the degree of

Master of Science in Mechanical Engineering with specialization in Vehicle Engineering

by

Akshay Bhoraskar
Student Number: 4702298

August 30, 2019

Akshay Bhoraskar: *Prediction of fuel consumption of long haul heavy duty trucks using machine learning and comparison of the performance of various learning techniques* (August-2019)

The cover picture is taken from the collection of Matthew T Rader.

The work in this thesis was done in collaboration with:

Department of Cognitive Robotics
Faculty of Mechanical, Maritime and Materials Engineering
Delft University of Technology, The Netherlands

Sustainable Transport and Logistics, TNO
Den Haag, The Netherlands

Supervisors:	Prof. Dr. ir. Martijn Wisse	TU Delft
	Dr. Wei Pan	TU Delft
	Dr. Victor Knoop	TU Delft
	Ir. Emiel van Eijk	TNO

For Mumma, Daddy, Shozab, Shantanu, Sylvia ...

Abstract

This study aims at a possible solution to predict the fuel consumption of heavy duty diesel trucks, particularly, the tractor-semitrailer for their long haul operations using various machine learning techniques. It intends to provide a possible alternative to simulation or physics based models, which often are very complicated. The stringent laws on emission control set by the Paris Agreement [1] and the fact that heavy duty trucks contribute to almost 27% of CO_2 emissions from road transport and their dependence on diesel for operations (in long haul) makes it the need of the hour to first, have an estimate on the emissions being produced and second, to develop technologies to reduce those emissions.

This study focuses specifically on the first part i.e., estimating the amount of fuel consumed by heavy duty trucks in the European Union and thereby determine the emissions being produced. The main objective is to examine whether an approach of machine learning could be a viable option to predict fuel consumption. This thesis is part of the AEROFLEX project [2] and was done in collaboration with TNO, which provided all the data-sets required for the study.

The idea was to explore the regime of machine learning for one time step ahead prediction of fuel consumption. Furthermore, this study also focused on the development of another model by not using any variables affected by the driver as input into the training model. This exclusion was necessary to make sure the model remained adaptive to new routes and new trucks, especially because large scale on-road testing of the newly developed trucks is impossible and also because this way would help predict the fuel consumed by a truck without the necessity of it driving on a road. The study concludes with a comparison with an existing simulation model at TNO and provide an alternative machine learning solution. It also provides a comparison between different machine learning techniques and suggest the most accurate one.

It was found that machine learning could potentially be used to predict the amount of fuel consumed by a long haul heavy duty truck driving on a motorway. It was also found that engine torque was the variable that affected the fuel consumption of the truck the most. Furthermore, Neural Network was the most potent algorithm among all the other learning techniques for both the models developed in this study with it performing better than the simulation tool by a factor of approximately 3.8 in the model where the driver/drive influenced inputs were not considered in the training data-set. The results obtained from this work at a sampling frequency of 10 Hz. (i.e., 0.1 seconds) are comparable to the ones reported by other sources at a sampling rate of 0.016 Hz. (i.e., 1 minute) or 0.0016 Hz. (i.e., 10 minutes). This goes on to say that the machine learning algorithms are also potent at much higher sampling frequencies.

Contents

LIST OF FIGURES

LIST OF TABLES

Chapter 1

BACKGROUND

1.1 Aeroflex

A possible solution to tackle the problem of increasing emissions from heavy duty trucks in the European Union (EU) is to develop new technologies that would be able to meet the requirements set by law. One such project in the EU is the aforementioned AEROFLEX project which aims at developing new technologies to meet the changing operating conditions while simultaneously implementing measures to achieve the new emission standards set by the European Commission. It is an EU project that aims at 'Aerodynamic and Flexible Trucks for Next Generation of Long Distance Road Transport.' It targets to achieve 18-33% more efficient heavy duty vehicles by 2030 by conceptualising and implementing new technologies for advanced power-trains, better aerodynamics and using loading space in the vehicle more efficiently, using innovative loading units. It also focuses on the demonstration and assessment of the new technologies developed in terms of both, technical capabilities and cost gains. More specifically, the main targets of the AEROFLEX project are as follows [3]:

1. 4-6% energy savings by using the loading space in the truck more efficiently.
2. 5-12% energy efficiency by integrating more flexible and advanced power-trains.
3. 5-10% reduction in energy consumption using improved truck aerodynamics.
4. 4-5% energy saving by separate platforms.

This work will be useful in the assessment phase of the project, where the new technologies developed would be tested on demonstrator vehicles in a limited number of on-road test-cases. The results can be used to check if the technologies developed adhere to the emission laws. It could also help with the fuel estimation of new technologies in new vehicle configurations and in the evaluation of those vehicles on roads it has not been tested on, thereby extrapolating the results from the on-road tests. The results from those tests will be helpful to know if the technology developed does help in attaining the required targets. The fact that this study could possibly be a feasible replacement for fuel prediction using simulations also makes it a worthwhile research topic.

1.2 Similar Work

This work in an extension of various research projects that have been carried out towards fuel consumption estimation. The two major approaches followed for such a prediction

are simulation techniques and machine learning techniques. The main questions that are targeted in the literature study of this work are:

1. Why is a machine learning approach necessary if a simulation based approach already exists and is known to be very accurate?
2. Which machine learning techniques have already been used for a prediction of the amount of fuel consumed by a vehicle? Which of them have been the most accurate ones?
3. What are the drawbacks of the studies in the regime of machine learning for the prediction of fuel consumption?
4. How will a study on long haul heavy duty trucks add to the novelty of the existing literature?

1.2.1 Simulation based study

Various simulation based studies predict the fuel consumption of heavy duty trucks [4] [5]. Another study by CM Silva et al. [6] on gasoline based light duty vehicles used various simulation models for the prediction of fuel consumption. A mathematical model developed by Kyoungho Ahn et al. [7] aimed at estimating vehicle fuel consumption based on their instantaneous speed and acceleration. Despite the high accuracy of all these models, a major disadvantage of this approach is the fact that it takes a considerable amount of time to design and run a simulation. Along with that, it also becomes increasingly difficult to adapt to new vehicle configurations and in most cases, stays particular to one specific type of vehicle. Altering the vehicle configurations/ adding more variables in the model will only make the simulation slower. Furthermore, as stated by [6], the fuel consumption prediction in their models highly depended on the engine power demand and to model the relation of other factors on the fuel consumption would be increasingly difficult. Thus, the simulation based study is not only prone to higher computation times but also higher modelling complexities and intricate design and requires profound system knowledge. This can be tackled by using machine learning approaches.

1.2.2 Machine Learning based study

There is a lot of existing literature on methods to predict the vehicle fuel consumption using machine learning techniques. The fact that this estimation technique is able to provide predictions for new vehicle configurations without many alterations makes it quite a lucrative method of predicting fuel consumption of passenger vehicles or even buses, trucks and airplanes.
Henrik Almer [8] worked on the prediction of fuel consumption of heavy duty vehicles using a data-set acquired by 62 vehicles. His model of fuel consumption was based on input attributes of slope, vehicle speed, weight of the vehicle, weather conditions, driving behaviour and speed limit of the road. He also went on to study whether the driving vehicle was in platoon or by itself. He used Linear Regression, Random Forests [9], Support Vector Regression (SVR) [10] and shallow Artificial Neural Networks (ANN) for his modelling. The results showed slope of the road, vehicle speed and vehicle weight being the most important variables for fuel estimation. He also concluded that a data collection sampling time of 10 minutes was preferable to a 1 minute collection sampling time. Although this study was very rich in its data-set, the fact that it was sampled at 10 minutes is an issue. A lot of parameters would change a lot in a period of 10 minutes. A shorter sampling frequency would make the model more realistic and applicable to real scenarios. In the work of Zheyuan Cheng et al. [11], a big data based deep leaning speed prediction

model was used which could predict speeds of vehicles in both urban traffic conditions and also on freeways. They incorporated factors like route types, route curvature, driver behaviour and weather and traffic conditions to predict vehicle speed. ANFIS [12], which combines the positives of both Fuzzy Systems and Neural- Networks was used in this study for their prediction. Although it is able to achieve good accuracy, this model does not account for other important variables that might affect the speed of the vehicle. Since the training and test data is the same, this work might also be prone to over-fitting, i.e., high accuracy on the test track only and very low accuracy on new, unseen data.
Another work by Abril Galang [13] used neural networks to predict the amount of fuel consumed by a hybrid vehicle, based on data collected at 1 Hz. His model was based on input parameters including Hybrid State (On/Off), Velocity, Acceleration, Engine Speed, Throttle Position, Brake Position and Elevation of the road. He concluded that his approach outperformed the physics based and simulation based models developed at the Colorado State University. But, he used a very small training set of 70 minutes for his Neural Network. This would also lead to over-fitting on the available data-set thereby leading to poor adaptive and reproducibility properties on new, unseen data-sets. Also, his network was shallow, i.e., with only one hidden layer and he did not explore deep neural networks, which could provide better results.
Other works on predicting fuel consumption of passenger vehicles by Predić, Bratislav et al. [14] aimed at using techniques like Random Forest , Neural Networks using the Levenberg-Marquardt backpropagation algorithm. This predicted the fuel consumed by using input attributes like elevation of the road, vehicle speed and vehicle acceleration with the models being not just fast but also computationally quite cheap. The results from them also predicted the fuel consumption of the vehicles quite accurately. This work too, like the previous one by Abril Galang [13] is also prone to over-fitting as the training data-set contains data from only one vehicle which hinders the model's reproducibility to other cars.
Federico Perrotta et al. [15] also applied machine learning techniques like SVR , Random Forest and ANN for fuel consumption prediction of trucks. Using a model that incorporated about 14,000 records, measured at 0.016 Hz, they concluded that Random Forest method slightly over-performed the other techniques. Along with that, the work done by S. Wickramanayake et al. [16] on the fuel consumption of a bus, also had results coinciding with that of Federico Perrotta et al. [15]. They too concluded that among the different techniques studied, Random Forest was able to outperform the others. Although the results from these two studies were comparable, the fact that less training data was used for the model of Federico Perrotta et al. [15] makes the model prone to over-fitting, i.e., the model unable to achieve high accuracy on new data due to lack of training data. Along with that, the sampling frequency in the study was too low. One sample every 60 seconds could lead to very drastic changes in the variables measured, thereby severely affecting the quality of data used for training the models.
A study by Lev Ertuna [17] uses neural networks to obtain a relation between vehicle properties and its fuel consumption in different driving conditions. The author analysed the effect of the number of hidden layers and hidden units in the model and how the accuracy of the model depends of those factors. Although quite insightful, the study shows no relation between the driving style and the fuel consumption. The author concluded that adding more hidden layers or more units in the hidden layer does not necessarily improve performance. This could be attributed to the fact that the author considered only a few input variables and the model could learn well from a shallow network. This is something that will be analysed in this research as the error posted by the author was quite high. An interesting insight was given by Thomas R. Waters [18] in his Master thesis, in which he classified the drivers ranging from conservative to aggressive, based on their acceleration

and braking behaviour. This classification helped him correlate the effect of the driver and his behaviour on the road to the amount of fuel consumed by the vehicle. Driver behaviour based on aggressiveness has also been studied by other authors [19]. The way in which the driver accelerates and decelerates during a driving cycle on a road would also make a lot of difference in the amount of fuel consumed by the vehicle [20] [21]. Another popular method to classify driver behaviour is to classify different road types such as highways, steep roads or rural roads and then recognising driving behaviour on those roads [22] [23].
The possibility of using machine learning to predict fuel consumption is not just limited to road vehicles. Other studies also extend this knowledge base to other means of transport. For instance, there have been studies on airplanes and the estimation of their fuel consumption using machine learning and ANN. Schilling [24] was able to use ANN to predict the fuel consumed by an aircraft by training the model on input parameters that described the aircraft as well as the surrounding weather conditions. He was able to conclude that ANN could be equally accurate as various physical calculations in predicting fuel consumption. The subsequent studies by Trani et al. [25] also confirmed the accuracy of such a model and also concluded that this approach was computationally much faster and less complex.
All these studies go on to say that machine learning could not just be a viable tool to predict the fuel consumption of vehicles but also a very accurate one. Some researchers have also stated that this technique, provided one has access to data, is more accurate than simulation based modelling techniques. This makes machine learning for the prediction of fuel consumption of vehicles a topic worth looking into for further research. The collective results of these studies also imply that vehicle velocity, acceleration, vehicle weight and slope of the road are among the most important factors affecting the prediction of the fuel consumed.
The major area where the literature does not provide any light on is the question whether these machine learning models are predictive for the entire route. The dependence of the driving style and all the input variables affected by the driver on the amount of fuel consumed by the vehicle is always very high. This study also delved into the prediction of the fuel consumed by the vehicle without any influence of the driver related input variables. This made the study adaptive to new roads and new vehicles without the vehicle actually driving on the new road thereby translating its results to unseen data and provide fuel consumption estimations.
Another area where the literature lacks is exploring deep neural networks for the prediction of fuel consumption. Certain variables in the input data-set have no direct correlation to the fuel consumed by the vehicle. Deep Neural Networks help in providing sophisticated modelling options, which should lead to better predictions. Along with that, this study also encompasses a much larger data-set than the ones considered in most other studies. That would help in overcoming shortcomings due to over-fitting of the model to one particular type of vehicle which in turn would make the model more adaptive and reproducible to new vehicles and new motorways. The data-set considered for this study is measured at 10 Hz. which would make the study more applicable to real-life applications unlike other sources in the literature where the sampling frequency was much lower.
As also mentioned by Henrik [8] in his work, the literature for heavy duty vehicles is still sparse and further research is needed in that sphere. This work focuses specifically on the long-haul heavy duty trucks which have a much higher payload capacity and are used for the transportation of freight over longer distances. These vehicles have not been studied in any of the literature stated, which adds to the novelty of this work. Furthermore, with the new laws on heavy duty vehicles introduced by the European Parliament [26], it becomes increasingly important to study the emissions of heavy duty trucks and build strategies to control them. This study continues to use various machine learning techniques to pre-

dict the fuel consumption of long haul heavy duty trucks and then compare the results obtained to those of a simulation based tool used at TNO.
The literature for road vehicles can be summarised in the table below:

Table 1.1: Summary of the literature for road transport

Author/s	**Vehicle Type**	**Approach**	**Input Variables/ Design Specifications**	**Best algorithm**
Henrik Almer [8]	Trucks	Machine Learning	Vehicle Properties, Platooning, Weather, Driver Behaviour	Random Forest
Zheyuan Cheng et.al [11]	Car	Machine Learning (Only ANFIS)	Route Types, Route Curvature, Driver Behaviour, Weather and Traffic Conditions	-
Ramadoni Syahputra [27]	Car	Machine Learning (Only ANFIS)	Engine Specifications, Vehicle Weight, Acceleration and Manufacture Year	-
Abril Galang [13]	Hybrid Car	Machine Learning (only Neural Networks)	Hybrid State (On/Off), Velocity, Acceleration, Engine Speed, Throttle Position, Brake Position, Elevation	-
Predić, Bratislav et al. [14]	Car	Machine Learning (Only Neural Network)	Road elevation, Velocity, Acceleration, Hour of day, Day of week	-
Federico Perrotta et al. [15]	Rigid Truck	Machine Learning	Vehicle Weight, Velocity, Acceleration, Geographical Position, Cruise Control	-
S. Wickramanayake et al. [16]	Bus	Machine Learning	Speed, Distance, Location, Ignition status, Elevation, Battery Voltage	Random Forest
Thomas R. Waters [18]	Electric Car	Machine Learning	Driver Aggressiveness, Vehicle Properties	Support Vector Machines (SVM)
Wawrzyniec Gołebiewski et al. [28]	Car	Simulation based, validated from measurements	design for velocity estimation and power consumed	-
Tony Sandberg [4]	Trucks	Simulation based, validated with real data	design of Vehicle Body, Engine, Powertrain, Wheel, Driver and Brake Control, Road, Weather	-
Kanit Wattanavichien et al. [5]	Light Duty Truck	Simulation Based, validated with measurement data	design of total Vehicle's Aerodynamic Drag, Grade, Rolling Resistance, Engine Performance Characteristics, Drive-train details, Vehicle Properties, Road, Tires	-
CM Silva et al. [6]	Light Duty Vehicles	Simulation Based and validated with real data	design for Vehicle Characteristics, Transmission Type, Engine Characteristics, Exhaust After-Treatment, Ambient Temperature, Road Topography and Vehicle Occupancy	-
João C Ferreira et al.[21]	Bus	Data-Warehousing and Data Mining	Driving style, Weather, GPS location	-
Kyoungho Ahn et al. [7]	Light Duty Vehicles and Heavy Duty Trucks	Simulation Based and validated with real data	design for Traffic, Vehicle, Road, Driver and the travel	-
Chris Bingham et al. [20]	Electric Car	Statistical Data Analysis	Driver Behaviour/ Driving Style	-
Ahmet Gürcan Çapraz et al. [29]	Car	Machine Learning	Speed, Acceleration, Engine RPM, Volumetric Efficiency, Slope of road	Support Vector Machines (SVM)
Michael Ben-Chaim [30]	Car	Simulation based with validation from experimental data	fuel consumption analytically related to vehicle properties, engine specifications, tire properties and driving style	-
Lev Ertuna [17]	Gasoline Cars	Machine Learning (only Artificial Neural Network)	Engine Specifications, Drivetrain Specifications, Transmission Specifications, Passenger and Luggage information	-

The questions that the literature study for this work aimed at can now be answered as follows:

1. *Why is a machine learning learning approach necessary if a simulation based approach already exists and is known to be very accurate?*

- Although the simulation based models are quite accurate and have been developed for a long time, their accuracy is compromised when new variables are introduced in the analysis of fuel consumption. Any new addition in the model is not just time consuming but also requires deep theoretical knowledge of the system. Along with that, simulation models are computationally very expensive as compared to the machine learning models. If the machine learning algorithms provide comparable results to the simulation based models, computation would be a lot faster and the model would be adaptive to new variables/ addition of new variables.

2. *Which machine learning techniques have already been used for a prediction of the amount of fuel consumed by a vehicle? Which of them have been the most accurate ones?*
   - Various authors have studied the effect of various variables on the fuel consumption of the vehicles. Majority of them gave a prediction using Linear Regression, Support Vector Regression, Random Forest and Neural Networks. They explored shallow neural networks and not deep neural networks. The conclusion of which method performs the best was not clear. Some [14] found Aritifical Neural Network to outperform the other techniques while others, including Federico Perrotta et.al, [15] and S. Wickramanayake et al. [16] conluded that Random Forest performed the best, although only slightly better than neural networks. Some others, like Ahmet Gürcan Çapraz et al. [29] conclude SVM outperforms all other techniques. This requires further attention to provide better results as to which technique provides the best predictions. This discrepancy can be explained due to the differences in the design of training models of various authors and the quality of the data obtained by each one of them.

3. *What are the drawbacks of the studies in the regime of machine learning for the prediction of fuel consumption?*
   - A major downside of machine learning models is the availability of quality data. They are always dependent on the training data, unlike simulation models. If there is ample data available, the option of developing a machine learning model is suitable. But, once the training has been done on a good data-set, the machine learning model is easily translatable to unseen data-sets for providing predictions.

4. *How will a study on long haul heavy duty trucks add to the novelty of the existing literature?*
   - As previously mentioned, the literature on long haul heavy duty trucks is quite sparse and needs further research, especially in the sphere of long haul heavy duty trucks. These types of vehicles are capable of carrying much higher payloads than the rigid trucks and are used to carry freight over very long distances. Along with that, the new laws set by the European Parliament to reduce the average emissions of heavy duty trucks by 15% in 2025 as compared to 2019 [26] makes this study relevant to requirements set by the law. Coupled with that, acquiring data-sets for long haul heavy duty trucks is also difficult which also adds novelty to the existing literature.

1.3 Previous Study of this work

As part of the previous study of this work, a route profiling model was developed at TNO. The work aimed at extracting latitude, longitude and elevation information from OpenRouteService [31]. It also extended to finding tunnels and bridges along the route specified and thereby adjusting the elevation along the length of the bridge/tunnel. The

elevation obtained at various data-points along the route was then used to find the slope of the route at those data-points using the formula:

$$Slope_i(\%) = \left[\frac{dE_i}{dD_i}\right] * 100\% \tag{1.1}$$

Where, dE_i is the change in elevation of the road between two consecutive instances and dD_i is the change in distance between two consecutive instances. The issue with calculating slope using this approach is the fact that the data-points received from OpenStreetMap (OSM) [32] are not separated by a constant distance. They are randomly distributed along the length of the route. For instance, there were two consequent data-points separated by a distance of 50 meters and there were also consequent data-points separated by a distance of over a kilometer. Some researchers, like Lindberg [33] have used these data-points straightaway to find the slope profile. But, this approach has its shortcomings as the data-points gathered from the online services were not separated by equal distances. This sometimes leads to the elevation changing over a very small distance which leads to the slope being very high for a very small instance. To counter this, the road was divided into segments of constant lengths and the elevation interpolated at those points to find the slope. This was preferable as it could overcome the problems created by data-points with unequal distances between them. The interpolation helped in removing this discrepancy and make the slope profile of the route more realistic. The procedure for this can be found in Appendix A

This study will act as the preliminary basis for profiling of the route on which the study has been done. The inputs obtained from OpenStreetMap contain information about the distance, latitude, longitude, altitude and speed-limit of the road. This information is used for two purposes in this work: First, the necessity of adding speed-limit into the data-set as inputs. The vehicle should not exceed the speed-limit on the motorway and hence it becomes an important limiting factor. The second purpose of using the information obtained from OpenStreetMap is for route-profiling. This becomes an important part when the test-set is being made for the model in which all the driver-related inputs are not considered in the model. The idea of testing vehicles on unseen routes without actually driving on it would only be fulfilled if the route could be modelled. This is made possible with the help of inputs from OpenStreetMap.

The above mentioned work will also help in calculating distance and the slope profile of the entire route, which can later be used to check if the model developed in this work is capable of performing well on unseen routes and thereby to assess if the model is in fact adaptive to new routes and new vehicles.

1.4 Summary

This chapter gave an insight into the background information of where this work fits in the bigger picture at TNO. It elaborated on the questions that needed to be answered from literature and also provided information on similar work done in this field and how this work would add to the novelty of the existing literature. This chapter concluded with the previous work of this study which was necessary to establish how route profiling (information about the route) was done from OpenStreetMap.

Chapter 2

INTRODUCTION

This chapter directs the focus to this research and provides an insight into the motivation behind this study and also elaborates on specific outcomes, impact, scope and the gaps targeted. It also proposes the main questions that this research aims to answer.
The Paris Agreement has made the reduction of CO_2 of prime importance. It has also stressed on the importance of developing new technologies that would help in the reduction of greenhouse gases and keeping the emissions under check. Another major motivation is the first ever CO_2 emission performance standards for new heavy duty vehicles in the EU laws set by the European Parliament on May 17, 2018. The laws state that the average emissions from the new trucks should be 15% lower in 2025 as compared to 2019. Heavy-duty vehicles are responsible for 27% of road transport CO_2 emissions and it has increased by almost 25% since 1990 mainly due to increase in road freight traffic and is set to increase further in the absence of new policies [26]. Along with all this, the electrification of heavy duty trucks is not realistic on a short notice due to the requirement of better infrastructure and better charging options [34] which drives the focus to make the diesel powered vehicles/ hybrid vehicles more efficient.

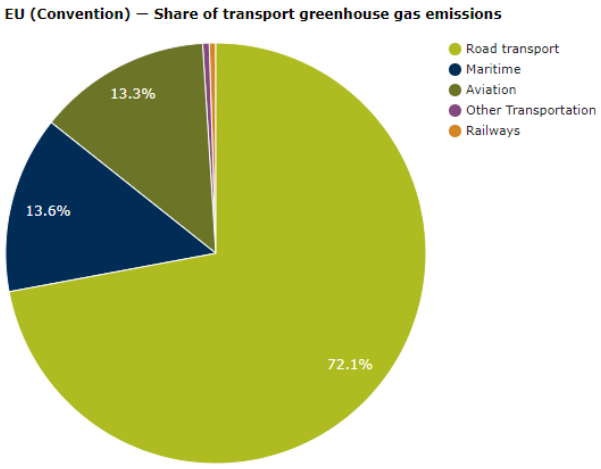


Figure 2.1: Share of Transport Greenhouse Gas emissions

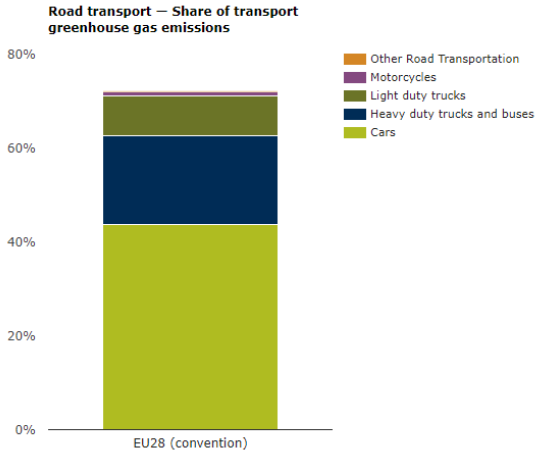


Figure 2.2: Road Transport- Share of Transport Greenhouse Gas emissions

Figure 2.1 and Figure 2.2 taken from the European Environment Agency [35] states that road transport is responsible for about 72% of the greenhouse gases produced from all the modes of transport. Heavy duty trucks account for about 19% in that. The possibility of using machine learning techniques to examine the fuel consumed by a vehicle in L/km (fuel consumed in liters per kilometer) is examined in this study. The attributes considered for this prediction include the variables influenced by the driver (namely, velocity, acceleration, brake and throttle position etc.), road attributes (namely latitude, longitude, altitude, road slope and the speed limit of the road), vehicle parameters (namely, engine power, vehicle weight, payload etc.) and weather attributes (ambient temperature, humidity etc.) The data is collected at a frequency of 10 Hz. from a Tractor-Semitrailer. The tractor-semitrailer, unlike the rigid truck has a higher maximum combination weight, i.e., combined weight of the empty vehicle and the payload of 40 tonnes. These kind of vehicles are used for the long haul freight transport.

Figure 2.3: One possible configuration of the Tractor Semitrailer

The motivation of this work is to find which of the aforementioned variables affect the fuel consumption of the truck the most. The work would also be a good method to predict the fuel consumption of long haul heavy duty trucks. The increasing need to transport more payload with a single truck has lead to the development of these trucks. The emission standards need to be checked for them and made sure that such trucks are feasible, not just in terms of their physical deployment but also that the emissions produced by them are under standard limit set by the European Commission. This tool will help in that prediction. Finally, the development of a tool that can act as a universal one, unlike the simulation tool which needs to be modified for each kind of truck and needs intricate knowledge about the vehicle and the road it is driven on, is the main motivation behind this work.

Although the literature suggests that the velocity and acceleration of a vehicle are among the major contributors towards the prediction of fuel consumption, the idea of not using any factor that is influenced by the driver for the training data was derived from the motive to make the model predictive and adaptive to new roads and new trucks. If the model could predict the fuel consumed by a vehicle only by using characteristics of the vehicle and the information about the route, it would be helpful in predicting the fuel consumption of any vehicle on a given route without the need for actually driving the vehicle on that route. Instead of it being just a 'regression fit' model, it should also help in predictions of vehicle fuel consumption without actual road tests.

2.1 Expected Impact of the work

This work primarily aims at providing an alternate solution to simulation models that are used to predict the fuel consumption of vehicles. With the increasing need to keep the emissions of the vehicles under constant check, it is imperative to develop methods that could accurately predict the fuel consumed and thereby the emissions produced. Along with that, modelling of a computationally less expensive and faster tool which could provide comparable results to the models in use also makes it lucrative.
This work will help in two non-trivial tasks; first, to predict the fuel consumed during test runs of the various vehicle manufacturers and second, to use it as a monitoring tool to make the auditing of fuel consumption easy. It can also be useful in the detection of anomalous vehicles by identifying irregular fuel consumption.
Furthermore, if a tool that could predict the fuel consumed by a heavy duty vehicle without driving the vehicle on an actual road is developed, it can go a long way in the testing of new technologies in newer vehicles. The idea that the time required for testing will be less is lucrative and worthy of further research. The fact that the tool is adaptive and will be able to predict the fuel consumption of different trucks on various roads could also be a way to keep the emissions of the vehicles under constant check. This will be of great importance in assuring the emission levels and having a definite idea of the actual emissions without actually performing road tests.

2.2 Gaps targeted in this study

The main questions that this research aims to answer are:

1. Is machine learning a possible option to predict fuel consumption of heavy duty vehicles?
   (a) If yes, which learning technique is the most accurate one? and;
   (b) Which variables in the measurement data-set affect the fuel consumption of the vehicle the most?
2. Is excluding driver affected input variables from the training data a viable option for the prediction of total fuel consumption? Also, is this approach adaptive to new roads and new truck configurations? Does it compare to the simulation tool developed at TNO?

2.3 Scope

This study, due to the nature of data used, will be limited to only long haul heavy duty diesel vehicles driving on motorways. According to the European Automobile Manufacturers' Association (ACEA) [36], about 96% of the heavy duty vehicles run on diesel. This

makes the study relevant to a large fraction of heavy duty vehicles, especially because of the new laws that must be adhered to in the coming years.

2.4 Summary

This chapter dealt with the motivation for this work and the impact that this work would have on the existing literature. This chapter concluded with the questions that this work aims to answer and established the scope of this work.

Chapter 3

METHODOLOGY

This chapter gives an insight into machine learning. It also defines the approaches used in the prediction of the fuel consumption. It also specifies the techniques used in this work and the process of evaluating results.

3.1 Machine Learning

One of the most popular formal definitions of machine learning, given by Tom Mitchell is as follows:

> *A computer program is said to learn from experience E with respect to some class of tasks T and performance measure P if its performance at tasks in T, as measured by P, improves with experience E.*

In simple words, a program that could learn and improve with experience and perform better at some given task. In this study, supervised learning algorithms are used in which the computer program learns from a set of inputs and desired output. This trained algorithm can then be used for new, unseen input sets and the computer program computes the outputs for those new inputs.

3.2 Approaches used in this work

There are two approaches of modelling used in this work. The first in which all the variables from the measurement data are used as inputs to the model and the second in which only road attributes and vehicle characteristics are used as inputs in the model. The first method would act as the prediction model to be used as a monitoring and faulty fuel consumption tool by predicting the fuel consumption of the vehicle for the next time step/ distance step whereas the second method would act as a more adaptive version, which could help predict the total fuel consumption of the vehicle for the entire trip without it actually driving on the road. For the sake of clarity, the input variables have been classified into 4 different types:

1. Driver/Drive Influenced parameters
2. Weather Attributes
3. Vehicle Parameters

4. Road Characteristics

The variables in each of these groups are shown in the table below:

Table 3.1: Input variables

Group	Variables
Driver/Drive Influenced Parameters	Velocity, Acceleration, Engine Speed, Accelerator Pedal, Brake Pedal, Driving Gear, Fan Speed, Engine Torque, Oil, Engine, Fuel and Coolant Temperature
Weather Attributes	Ambient Temperature, Asphalt Temperature, Humidity
Vehicle Parameters	Weight, Payload, Front Area, Rated Engine Power, Number of Axles
Road Characteristics	Slope, Elevation, Speed-Limit, Distance, Longitude, Latitude

The sources of these input variables is further elaborated in Section 4.1.4. The classification was only done for the sake of clarity. The measurement data does not make this classification. It was made for the author's convenience of explaining the models developed.

3.2.1 Model 1: Modelling using all the input variables, namely driver/ drive influenced, vehicle parameters, road parameters and weather parameters

A possible way to predict the amount of fuel consumed is by using the knowledge of physics based equations. A model used by Hilde Huisman [37] in her work showed that the power needed at the wheel of a vehicle can be calculated as follows:

$$\begin{aligned} P_{wheels} &= P_{rrc} + P_{drag} + P_{inertia} + P_{gradient} \\ &= MC_{rr}Agv\cos\theta + \frac{1}{2}\rho C_d A v^3 + Mav + Mgv\sin\theta \end{aligned} \tag{3.1}$$

Where,

P_{wheels}	Power needed at the wheel (kW)
P_{rrc}	Power required to overcome rolling resistance (kW)
P_{drag}	Power required to overcome drag force due to air resistance (kW)
$P_{inertia}$	Power required to overcome the inertia of the vehicle (kW)
$p_{gradient}$	Power required by the vehicle to overcome gradient forces (kW)
M	Combined mass of the vehicle and the payload (kg)
C_{rr}	Coefficient of rolling resistance (-)
A	Front area of the truck (m^2)
g	Acceleration due to gravity (m/s^2)
v	Velocity of the vehicle (m/s)
θ	Gradient of the road (rad)
ρ	Density of air (kg/m^3)
C_d	Drag Coefficient (-)
a	Acceleration of the vehicle (m/s^2)

The power delivered to the wheels has a direct relation to the amount of fuel consumed by the vehicle:

$$Fuel\ Consumed \propto P_{wheels} \tag{3.2}$$

In the Equation 3.1, the fuel consumed is a function of velocity, slope and acceleration. But, it is known from the knowledge of the system that the acceleration at some iteration would

also depend on the previous acceleration due to the fact that the vehicle can accelerate only upto a certain limit. It also has a dependence on the previous value of velocity as the acceleration is the change in velocity in a set time difference. This can be represented as:

$$a_i = f(v_i, v_{i-1}, a_{i-1}) \tag{3.3}$$

This goes on to say that the fuel consumed at some point would depend not just on the present value of the acceleration of the vehicle but also the previous value. The velocity that the vehicle drives with is also governed by the speed-limit of the road. This also acts as one of the variables that would influence the amount of fuel consumed. The final equation that could be now be used for the training of the hybrid model is as follows:

$$\begin{aligned} \mathit{Fuel\ Consumed}_i = f(&\mathit{Velocity}_i, \mathit{Elevation}_i, \mathit{Slope}_i, \mathit{Velocity}_{i-1}, \\ &\mathit{Front\ Area}, \mathit{Vehicle\ Weight}, \mathit{Payload}, \mathit{Speedlimit}_i, \\ &\mathit{Acceleration}_i, \mathit{Acceleration}_{i-1}) \end{aligned} \tag{3.4}$$

The above equation gives the relation between fuel the driving characteristics, road characteristics and vehicle characteristics with the fuel consumption. The final model would have fuel consumption dependent on driver influenced factors, weather attributes, vehicle parameters and the road characteristics.

$$\begin{aligned} \mathit{Fuel\ Consumed}_i = f(&\mathit{Driver/Drive\ Influenced\ Factors}, \mathit{Weather\ Attributes}, \\ &\mathit{Vehicle\ Parameters}, \mathit{Road\ Characteristics}) \end{aligned} \tag{3.5}$$

The data collected, as mentioned earlier, is done at 10 Hz. For the sake of prediction in this model, the same data is used without any alterations to estimate the fuel consumed for the testing set and the outcome would also be at a frequency of 10 Hz, i.e., the amount of fuel consumed would be predicted at intervals of 0.1 seconds.

3.2.2 Model 2: Using only road and vehicle parameters as inputs and excluding all driver/drive and weather related inputs

This type of modelling would act as the one in which the fuel consumption of the vehicle could be predicted without any influence of the driver. The amount of fuel consumed by a vehicle is heavily dependent on the driver and his driving style. If the model is able to predict the fuel consumed by the vehicle without taking into account the driver inputs, it could be put to use on other roads to predict the fuel consumption of other vehicles. For the model to work on roads where tests have not been conducted, it is imperative to remove those input variables that can only be obtained from the measurements from the road. This was the main motivation behind this type of modelling. The fuel consumed by the vehicle in this model could be expressed as the following:

$$\begin{aligned} \mathit{Fuel\ Consumed} = f(&\mathit{Engine\ Power}, \mathit{Front\ Area}, \mathit{Vehicle\ Weight}, \\ &\mathit{Payload}, \mathit{Road\ Elevation}, \mathit{Road\ Slope}, \mathit{Speed\ Limit}, \\ &\mathit{Distance}) \end{aligned} \tag{3.6}$$

The final model would have the fuel consumption dependent on vehicle parameters and road characteristics only.

$$\mathit{Fuel\ Consumed} = f(\mathit{Vehicle\ Parameters}, \mathit{Road\ Characteristics}) \tag{3.7}$$

It is understood that driver behaviour would impact the fuel consumption but due to lack of labelled data on the driver driving pattern, it cannot be incorporated into this study. A possible solution is to build different models for different driving styles and then implement that model for fuel prediction. This is beyond the scope of this work but could

be performed with the help of labelled data on the driving patterns of different drivers, as explored by some authors [18] [20] [21] [23]. This study only dives into the possibility of estimating the amount of fuel consumed by a particular long haul vehicle on any motorway without the need for actually driving on it.
Since in this type of modelling there is no estimate of time, the model is trained using a data-set which is not time dependent but distance dependent. The fuel consumed at intervals of a specific distance (2 meters in this case) is used as the training set. This helps the model translate easily to help in the prediction of the fuel consumption for new routes. Since these routes have a distance measure, it is easy to use the trained model for this new testing data. The model would then provide the prediction of the amount of fuel consumed for every 2 meters of distance travelled.

3.3 Machine Learning Models used in this work

The machine learning techniques that will be used in this work are elaborated in this section. Since the model requires the prediction to be a continuous set of values form a given set of inputs, the techniques described are all supervised, regression learning techniques.

3.3.1 Linear Regression

Linear Regression is the process of fitting a straight line to a set of input variables to predict an output. A univariate linear regression includes the prediction of one output from one set of input values. It can be illustrated as follows:

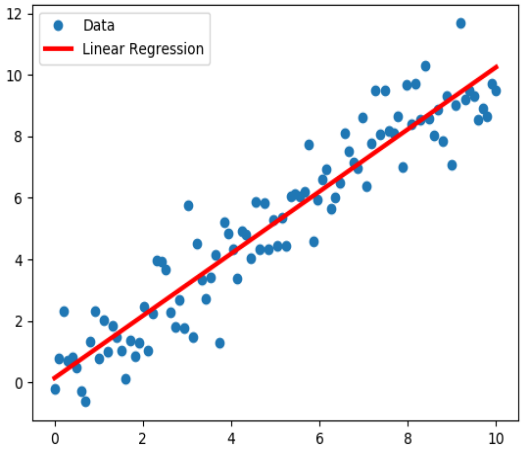


Figure 3.1: An illustration of uni-variate Linear Regression

This work is a case of using multiple inputs to predict an output, which is called the multivariate linear regression, the hypothesis of which can be represented as follows:

$$y = h_\theta(x) = \theta_0 + \theta_1 x_1 + \theta_2 x_2 + \cdots + \theta_n x_n \tag{3.8}$$

Where, y is the output and $x_1, x_2, \ldots x_n$ are the input variables. $\theta_1, \theta_2, \ldots \theta n$ are the coefficients that need to be learned. In linear regression, the aim is to find the values of the coefficients to provide the best fit between the input and output. This is done by minimising a cost function. The hypothesis above can be represented in their vectorized version along with the cost function as:

$$h_\theta(x_j) = \theta^T x_j = \sum_i \theta_i x_{j,i} \tag{3.9}$$

$$\theta = min \sum_{j} Cost\ Function(y_j, h_\theta) \tag{3.10}$$

The above equations help in finding the values of the weights (or, coefficients) by using the gradient descent method to find an appropriate solution for the prediction of the output. The cost function used is the least squares method which can be represented as:

$$Cost\ Function\ J(\theta) = \sum_{j=1}^{n} (h_\theta(x_j) - y_j)^2 \tag{3.11}$$

3.3.2 Support Vector Regression (SVR)

Support Vector Regression (SVR) is a regression technique which is used for the continuous prediction of a variable from a given set of input variables. A characteristic of the SVR algorithm is the fact that it is not very harsh on outliers, as long as the outliers are within a predefined limit ϵ. This goes on to say that the algorithm allows the points to be within an ϵ deviation.
To find an estimate function, the following method is used in SVR [10]:

$$f(x) = \langle w^T, x \rangle + b \tag{3.12}$$

where, $w \in \mathbb{R}^N$ *and* $n \in \mathbb{R}$ This function needs to make sure that the points lying in the vicinity ϵ around it are not penalised, i.e., $f(x) - y \leq \epsilon$. The optimisation problem for this function that needs to be minimised is as follows [10]:

$$\begin{aligned} min \quad & \left(\frac{1}{2} \|w\|^2 + C \sum_{i=1}^{M} (\xi_i + \xi_i^*) \right) \\ such\ that\ & y_i - \langle w^T, x_i \rangle - b \leq \epsilon + \xi_i^* \\ & \langle w^T, x_i \rangle + b - y_i \leq \epsilon + \xi_i^* \\ & \xi_i, \xi_i^* \geq 0, \forall i = 1 \dots M \end{aligned} \tag{3.13}$$

In the above Equation 3.13, w is the hyper plane, ξ_i, ξ_i^* are slack variables, b is the bias, ϵ is the allowable error and C is the penalty factor for points lying outside the allowable deviation ϵ. The following figure shows a graphical representation of the SVR:

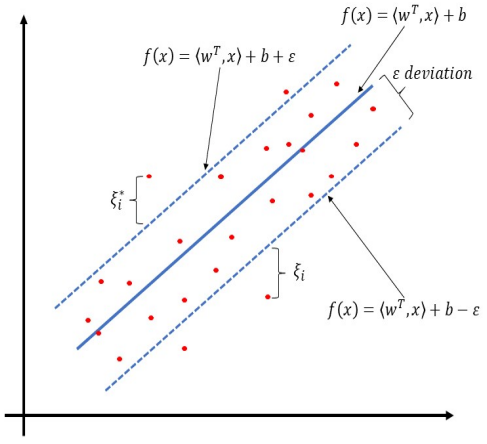


Figure 3.2: An illustration of the SVR algorithm

The objective functions in Equation 3.12 and 3.13, are solved to get a regressive function of the following form:

$$y = f(x) = \sum_{i=1}^{M} (\alpha_i - \alpha_i^*) \langle x, x_i \rangle + b \tag{3.14}$$

$\alpha_i, \alpha_i^* > 0$ are the support vectors. Equation 3.14 cannot be used for higher dimension feature space which brings in the necessity to use 'kernel trick' which helps transform the equation into a higher dimension feature space. The choice of the kernel is part of tuning the SVR algorithm. The final equation for multiple feature space can hence be written as:

$$y = f(x) = \sum_{i=1}^{M} (\alpha_i - \alpha_i^*) k\langle x, x_i \rangle + b \tag{3.15}$$

where, k is the kernel function.
The choice of the kernel varies from it being linear, polynomial or a Gaussian Radial Bias Function (RBF). The following equations give some depth into the functions that represent these kernels:

$$\begin{aligned} \textit{Linear Kernel:} \quad & k(x, x_i) = x^T x_i \\ \textit{Polynomial Kernel:} \quad & k(x, x_i) = (1 + x^T x_i)^d \\ \textit{Gaussian Radial Bias Kernel:} \quad & k(x, x_i) = \exp \frac{- \parallel x - x_i \parallel^2}{2\sigma^2} \end{aligned} \tag{3.16}$$

In Equation 3.16, for polynomial kernels, d is the degree of the polynomial. Hence $d > 0$. The σ in RBF kernel is a parameter defining the width of the kernel and therefore $\sigma > 0$. The major disadvantage of using the SVR is the number of variables that need to be tuned for it to work appropriately. The tuning requires the need to find appropriate values for all the free variables mentioned in Equations 3.12, 3.13 and 3.14. The tuning was done via a grid search method which helped finding the right values for ϵ and C. The values of ϵ can change the accuracy of the SVR. With smaller values, the model is prone to overfitting (i.e., model performing well on the training set but not so good on the testing data) while with larger values of it, the model is prone to underfitting (i.e., the model not able to learn enough from the training set itself and not able to perform well on it). The values of the parameter C, on the other hand, determines the penalty that the model would incur. With a very low value of C, the model is more prone to underfitting while at very high values, the model is prone to overfitting. The method used for the choice was the Mean Squared Error. The values for which the Mean Squared Error for the cross-validation set was least was considered as the appropriate value.
The equations for the application of SVR have been taken from the paper on Support Vector Machines by Corinna Cortes [10] and the implementation and tuning of the SVR algorithm on Python was done with the help of the scikit-learn [38]. Further information about the SVR algorithm and its equations can be found in Appendix B

3.3.3 Random Forest Regression

Random forest is a supervised learning algorithm. It is used for both classification and regression problems. In the context of this study, the Random Forest Regression, as the name suggests is used as a regression algorithm. Random forests are derived from decision trees first developed by Tin Kam Ho [9].
Decision Trees can be explained as a learning method in which the output predictions are obtained through a series of questions which narrow the range of possible values to arrive at a possible prediction. Each of these questions has a true or false answer and the algorithm derives a solution for each of the answer. A simple analogy, relevant to this study, can be explained via the following figure:

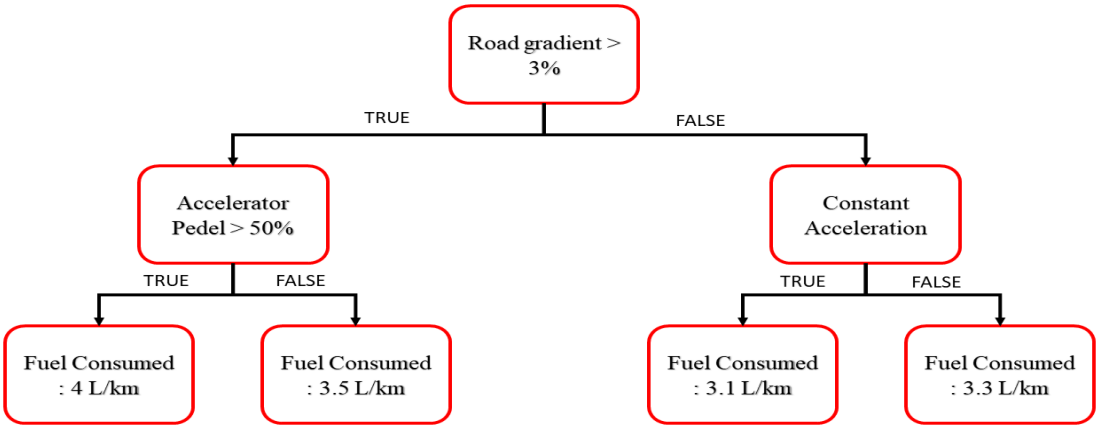


Figure 3.3: An illustration of a decision tree

As can be seen in Figure 3.3, a series of questions help the algorithm to arrive at a possible prediction. Since this is a supervised learning algorithm, the algorithm is able to predict the set of values with the training provided earlier and the questions framed are also a result of the algorithm learning from the data provided. Each node in a decision tree is further subdivided into branches and the last node in a branch is called a leaf. The tuning of the algorithm needs to be done to find the optimal values for each of these elements.
A random forest is a compilation of many of these decision trees in which each tree can come up with its own prediction. This helps make the algorithm robust as the final output is an average of the predictions made by each tree. A random forest can be illustrated as following:

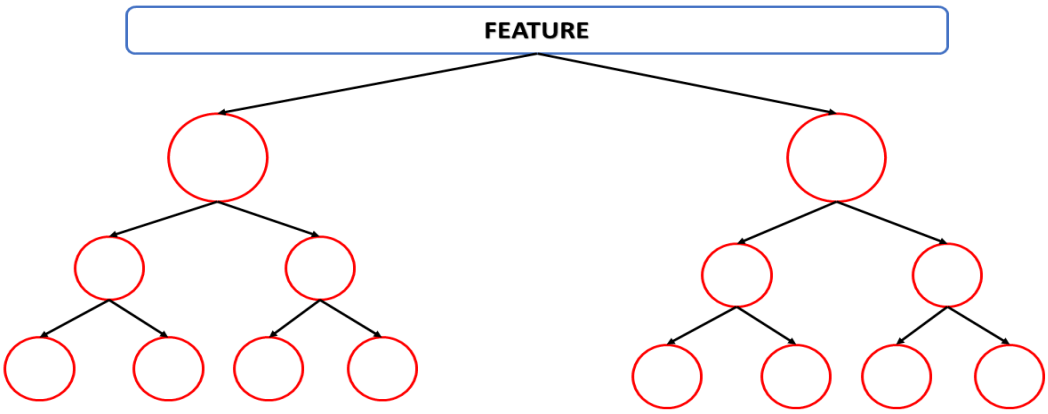


Figure 3.4: An illustration of a random forest with 2 trees

Figure 3.4 is a typical random forest with two decision trees which helps in the prediction of the fuel consumption in this context. Each leaf at the end makes a prediction and the final value is the average of all the predictions made. A major advantage of a random forest regressor is the fact that it is relatively very easy to tune and implement. Along with that, random forest regressors can overcome over-fitting of data due to the robustness achieved by different predictions of different trees. The flip side is the fact that with increasing the number of trees, the algorithm also becomes more computationally expensive.
For its implementation, scikit-learn was used [38], which helps in tuning and finding the right parameters for the algorithm.

3.3.4 Neural Networks

Artificial Neural Networks are derived from the biological neuron models. An algorithm which is used both for classification and regression tasks and in the scope of this study it is used for a regression task as the output is a continuous variable.
A neuron in the scope of an artificial network is called a node [39]. The working of an artificial neuron can be explained as the neuron (node) getting activated when a particular threshold is exceeded. Each input is associated with a weight. The summation of different inputs (multiplied by their weights) determines whether the threshold is exceeded or not. The following figure shows the arrangement of an artificial neuron or a node:

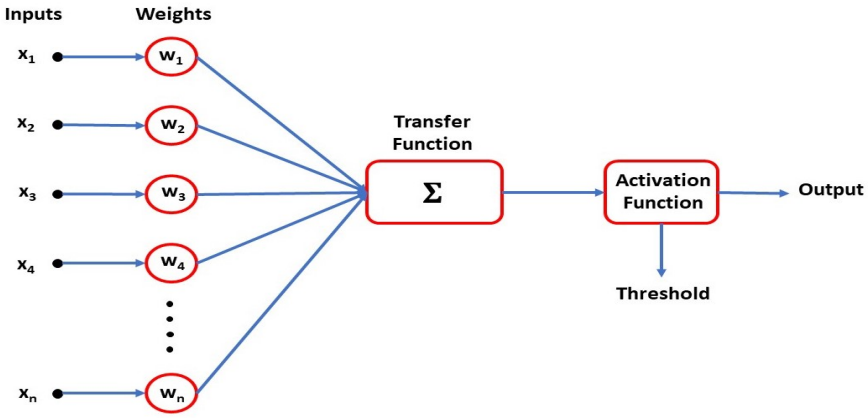


Figure 3.5: An illustration of an Artificial Neuron

The Figure 3.5 shows the different inputs $(x_1, x_2 \dots x_n)$, each associated with a weight $(w_1, w_2 \dots w_n)$. The product of the input with its corresponding weight is fed into the activation function to check if the threshold is exceeded. If yes, the node is activated and a corresponding output is received.
An artificial neural network is made by arranging the nodes in layers to make it into a network. Each node is connected to the other by a link which is associated with its individual weight. All the layers in between the input and output are called the hidden layers. The figure below shows an illustration of an artificial neural network:

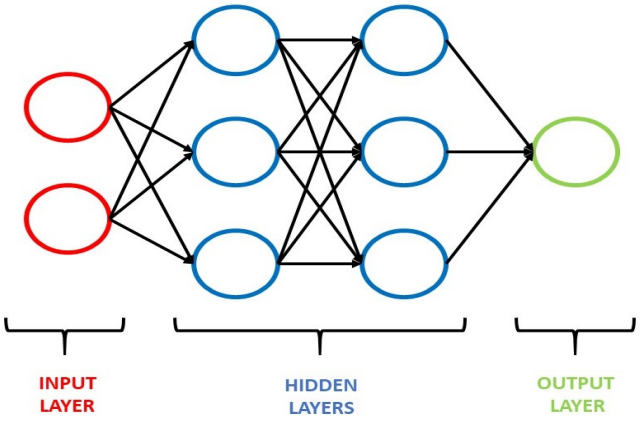


Figure 3.6: An illustration of an Artificial Neural Network with 2 nodes in Input layer, 3 nodes in each of the 2 Hidden Layers and 1 node in the Output Layer

The Neural Network used in this work is the feed-forward neural network in which the information is carried only in the forward direction. Each node receives an input from a node in the previous layer and the information propagates through the network. In such a case, the network is also able to back-propagate the error to adjust the weights associated with the inputs to make sure the outputs have the least possible error. This is called back-propagation. The back-propagation is implemented through a gradient descent method. Neural Networks involve a lot of tuning and the performance of both shallow (networks with one hidden layer) and deep (networks with more than 2 hidden layers) will be evaluated in this work. A possible downside of Neural Networks is the algorithm over-fitting (i.e., algorithm performing well on the training model but not so well on the test set). This problem was countered in this study by using dropouts. Dropout is a technique in which the model is forced to learn even when all the input features are not present. This is implemented with the help of a probability value. For instance, if the dropout is set at 0.5, 50% of the input is 'dropped' and the algorithm is forced to learn with only remaining 50% of the input features. A simple illustration of this technique on the network of Figure 3.6 would look as follows:

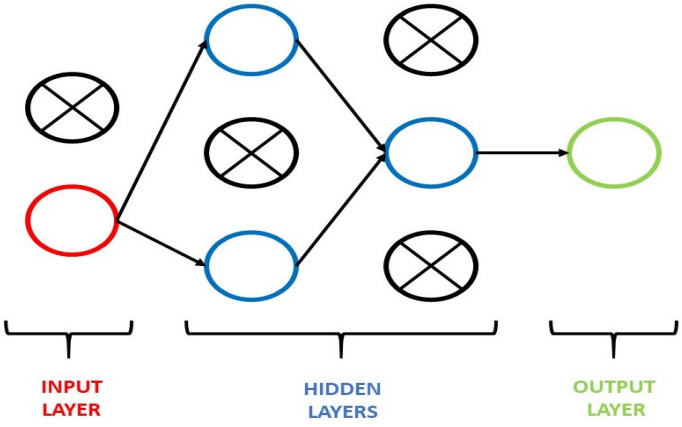


Figure 3.7: An illustration of an Artificial Neural Network with dropout of 0.5

Figure 3.7 shows an illustration of how a network would function with a dropout of 0.5. The algorithm is forced to learn with only 50% of the available inputs. For the implementation of a neural network, the values of all these variables are found by the tuning of the network. In Python, the implementation of this algorithm was done using Keras [40] and TensorFlow [41].

3.4 Evaluation Process

This section deals with the procedure to evaluate the performance of the developed algorithms and the various metrics used for that purpose.

3.4.1 Root Mean Squared Error

The root mean square error (RMSE) is the root of the arithmetic mean of the difference between the actual and estimated value. It is given by the following formula:

$$RMSE = \sqrt{\frac{\sum_{i=1}^{N}(y_{predicted,i} - y_{reference,i})^2}{N}} \tag{3.17}$$

In the context of this study, $y_{predicted}$ is the predicted value of the fuel consumed while $y_{reference}$ is the fuel consumption value at the same instant from the measurement data

and N is the number of measurements. RMSE itself is not a very intuitive measure of error. In this study, the RMSE error will be reported in Liters per kilometer (L/km). which is more intuitive and easier to compare.

3.4.2 Total Fuel Consumed

The total fuel consumed will also be used as a measure for comparison between the actual measurements and the predictions from the algorithm. It will also act as a vital measure for comparison between the simulation tool and the machine learning algorithms. The total fuel consumed will be reported in liters (L) for the entire trip.

3.4.3 Absolute Error in total fuel consumed

The absolute error (AE) is the error between the predicted value and the reference value.

$$Absolute\ Error\ in\ total\ fuel\ consumed = \mid \sum_{i=1}^{N} y_{predicted,i} - \sum_{i=1}^{N} y_{reference,i} \mid \tag{3.18}$$

In the context of this study, as previously mentioned, $y_{predicted}$ is the predicted value of the fuel consumed while $y_{reference}$ is the fuel consumption value at the same instant from the measurement data. Since AE is also not a very intuitive measure of error, it will also be reported over 1 km. to make the value more intuitive.

3.4.4 Comparison with simulation tool

The simulation tool used at TNO, called the advanced vehicle model is used to predict the amount of fuel consumed by the vehicle. The tool has been implemented on Simulink and includes a design of trailer-chassis, truck-chassis, truck powertrain, vehicle body and the driver. The simulation model requires the same input as the input fed into Model 2 as described in Section 3.2.2 of this work. The Model 2 was specifically designed in this way to ensure a measure for comparison.
The advanced model has been made to include the design of the following:

1. Trailer Chassis: Includes the design of brakes, tyres and axles.
2. Truck Chassis: Includes the design of front tyres and brakes and rear tyres and brakes.
3. Truck Powertrain: Includes the design of clutch, engine, final drive, gearbox and the fuel consumption model.
4. Body of the vehicle
5. Driver model

As can be established, the advanced model is an elaborate simulation tool which would act as the source of comparison for the machine learning models developed. The comparisons would be made on the basis of total fuel consumed for a particular test ride in liters of fuel consumed (L) and liters of fuel consumed per kilometer (L/km).

3.5 Summary

This chapter introduced the different machine learning techniques that would be used in this study, namely, Linear Regression, Support Vector Regression, Random Forests and Neural Networks. It also gave an insight into the two kinds of modelling used in this work; First, in which all the input variables from the measurement data-set were used for the prediction of the fuel consumption, namely, vehicle-related inputs, driver-influenced inputs, road-related inputs and weather-related inputs and; Second, in which the driver-influenced inputs were excluded from the model and the fuel consumption of the truck was estimated from only vehicle-related inputs and road-related inputs.
The chapter also introduced the metrics that would be used to establish the performance indices of the various machine learning techniques for the two different kinds of modelling. The chapter concluded with providing information about the Simulation tool used at TNO for the prediction of fuel consumption.

Chapter 4

DATA COLLECTION

4.1 Data Collection

4.1.1 Measurement Data from road tests

The Aeroflex project required trucks to be driven on a test route in Spain, for the road tests. A 242 km motorway, this test was the sole source of data used in this study. The fact that the vehicle was driven on an actual road takes into account that the data-set incorporates the effect of traffic and also the variations brought about by changes in weather.
The road test was conducted for the one vehicle type as mentioned in Chapter 2, the tractor semi-trailer. This vehicle was driven on the road carrying the average European payload (which is 13598 kg. for Tractor-semitrailer). The table below elaborates on the kind of readings obtained from the road test. The data collection from this test was done at 10 Hz. i.e., ten readings for each of the mentioned variables in one second.

Table 4.1: Variables obtained from the road tests

Variable Name	**Description (unit)**
Time	Time of the measurements (s)
Distance	Distance travelled by the vehicle (km)
Velocity	Velocity of the vehicle at the current and previous instance (km/h)
Longitude	GPS longitude position of the vehicle (-)
Latitude	GPS latitude position of the vehicle (-)
Altitude	GPS altitude at the vehicle's position (m)
Ambient Temperature	Temperature of the air ($°C$)
Humidity	Humidity in the air (%)
Actual gear ratio/number	The driving gear of the vehicle at that point (-)
Asphalt Temperature	Temperature of the surface of he road ($°C$)
Tire Temperature	Temperature of the tire of the vehicle ($°C$)
Exhaust Temperature	Temperature of the exhaust gases ($°C$)
Engine Speed	Speed of the engine (rpm)
Engine Torque	Torque of the engine (%)

Continued on next page

Variable Name	Description (unit)
Oil Temperature	Temperature of the oil ($°C$)
Accelerator Pedal	Actuation of the accelerator pedal (%)
Brake pedal	Actuation of the brake pedal (%)
Coolant Temperature	Temperature of the coolant in the engine ($°C$)
Fuel Temperature	Temperature of the fuel in the vehicle ($°C$)
Fan Speed	Speed of the radiator fan (rpm)
Fuel Flow Measurements	The amount of fuel used by the vehicle at that point (L/s)

All the variables mentioned in the table above will be used in the training of the model. An advantage of using machine learning techniques lies in the fact that the model can accommodate many variables without having to specifically understand the individual effect of them on the target variable. In this case, the target variable (Output) is the Fuel Consumption which would be estimated with the help of all the other variables.

4.1.2 Data from OpenStreetMap

OpenStreetMap act as the replacement of the GPS data for Model 2 used in this work. The data obtained from the OpenStreetMap contained information about the geographical location in terms of latitude, longitude and elevation at points along the length of the route. Along with that, the OpenStreetMap also contain an important factor that needs to be incorporated in the training set, even for Model 1. This is the speed-limit of the road. Once the route is selected on the map, using the overpass API, the speedlimit for various links along the way can be found. This input variable would act as a limiter. Since the vehicle should never exceed the speed-limit of the road, the model should know the the maximum speed at which it can drive on the road.
The table below shows the variables obtained from OpenStreetMap and their description.

Table 4.2: Variables obtained from OpenStreetMap

Variable Name	Description (unit)
Longitude	OSM longitude position of the vehicle (-)
Latitude	OSM latitude position of the vehicle (-)
Distance	Distance travelled by the vehicle (km)
Altitude	OSM altitude at the vehicle's position (m)
Speed limit	Speed Limit of the link of the road (km/h)

The OpenStreetMap therefore act as the source for geographical location for Model 2 and the information about the speed limit of the road acts as an important input variable for both models. Route profiling of the route has already been described as part of previous work in Section 1.3 and the inclusion of data obtained from the OpenStreetMap into the measurement data-set is further elaborated in Section 5.1.2.

4.1.3 Vehicle Data

The data from vehicles is another important factor in the consolidation of the data-set. The amount of fuel consumed by any vehicle also depends on what kind of vehicle is driving on the road. The specific attributes of the driven vehicles was necessary to make sure the model is tunable to newer vehicles with different attributes.
This study focuses on the type of heavy duty trucks mentioned in Chapter 2 and the attributes of that vehicle are also incorporated in the model. The following table gives an insight into the variables considered.

Table 4.3: Variables obtained from the Vehicle

Variable Name	Description (unit)
Engine Power	Rated power of the engine of the vehicle (kW)
Front Area	Frontal area of the truck (m^2)
Axles	Number of axles in the vehicle (-)
Weight	Net weight of the empty vehicle (kg)
Payload	Payload the vehicle is carrying (kg)

The power of the engine is necessary to know how much power the vehicle can produce. This information is necessary to know how the vehicle would respond to the amount of weight it is carrying. The acceleration potential of the vehicle would be determined by this factor. The frontal area of the truck is used to know how much resistance to the truck is produced due to the air drag. The vehicle data is necessary to make the model adaptive to newer vehicle configurations and be able to provide accurate predictions for the same.

4.1.4 Consolidation of the data for Model 1

The first model used in this study as described in Section 3.2.1 will help predict the fuel consumption of the vehicle with all the inputs derived from the three aforementioned sources. The model would therefore predict the effect of the variables mentioned in the table below. The one in red is the target variable (output variable) while the non-highlighted ones are the input variables. This data-set would be used for the training and testing of the model developed.

Table 4.4: Consolidated list of variables used for Model 1

Variable Name	Description (unit)
Time	Time of the measurements (s)
Distance	Distance travelled by the vehicle (km)
Velocity	Velocity of the vehicle at the current and previous instance (km/h)
Longitude	Longitude position of the vehicle (-)
Latitude	Latitude position of the vehicle (-)
Altitude	Altitude at the vehicle's position (m)
Ambient Temperature	Temperature of the air ($°C$)
Humidity	Humidity in the air (%)
Actual gear number	The driving gear of the vehicle at that point (-)
Engine Power	Rated Power of the engine of the vehicle (kW)
Front Area	Frontal area of the truck (m^2)
Axles	Number of axles in the vehicle (-)
Weight	Net weight of the vehicle (kg)
Payload	Payload the vehicle is carrying (kg)
Speed limit	Speed Limit of the link of the road (km/h)
Asphalt Temperature	Temperature of the surface of he road ($°C$)
Tire Temperature	Temperature of the tire of the vehicle ($°C$)
Exhaust Temperature	Temperature of the exhaust gases ($°C$)
Engine Speed	Speed of the engine (rpm)
Engine Torque	Torque of the engine (%)
Oil Temperature	Temperature of the oil ($°C$)

Continued on next page

Variable Name	Description (unit)
Accelerator Pedal	Actuation of the accelerator pedal (%)
Brake pedal	Actuation of the brake pedal (%)
Coolant Temperature	Temperature of the coolant in the engine ($°C$)
Fuel Temperature	Temperature of the fuel in the vehicle ($°C$)
Fan Speed	Speed of the radiator fan (rpm)
Fuel Flow Measurements	The amount of fuel used by the vehicle at that point (L/s)

The Equation 3.4 incorporating the variables for predicting the fuel consumption in Chapter 3 can now be extended to incorporate all the variables that will be used in the model to predict the amount of fuel consumed by the vehicle.

4.1.5 Consolidation of the data for Model 2

The second type of modelling considered in this study will deal only with inputs from the road and the vehicle as described in Section 3.2.2. The data-set used here would specifically exclude any variable that is affected by the driver. The table below summarises the variables used in this model. The one in red is the target variable (output variable) while the non-highlighted ones are the input variables. This data-set would be used for the training and testing of the model developed.

Table 4.5: Consolidated list of variables used in Model 2

Variable Name	Description (unit)
Distance	Distance travelled by the vehicle (km)
Longitude	Longitude position of the vehicle (-)
Latitude	Latitude position of the vehicle (-)
Altitude	Altitude at the vehicle's position (m)
Engine Power	Rated power of the engine of the vehicle (kW)
Front Area	Frontal area of the truck (m^2)
Axles	Number of axles in the vehicle (-)
Weight	Net weight of the vehicle (kg)
Payload	Payload the vehicle is carrying (kg)
Speed limit	Speed Limit of the link of the road (km/h)
Fuel Flow Measurements	The amount of fuel used by the vehicle at that point (L/s)

The Equation 3.6 incorporating the variables for predicting the fuel consumption in Chapter 3 can now be extended to incorporate all the variables that will be used in the model to predict the amount of fuel consumed by the vehicle.

4.2 Summary

This chapter was dedicated to the collection of data. It gave an insight into the different sources from where the data was obtained. The data obtained was segregated into three sources; one, obtained from the road-tests; two, obtained from the vehicle; and three, obtained from OpenStreetMap.
The chapter also dealt with the consolidation of the data that would be used for each of the two models used in the study.

Chapter 5

DATA PRE-PROCESSING AND FILTERING

5.1 Data Pre-Processing

This section elaborates on the pre-processing of the data obtained from the measurements. It deals with the creation of new input variables from the ones obtained from the data-set and finally, the filtering techniques used to make the data-set usable.

5.1.1 Adding input variables

The data collected from the various sources mentioned in the previous chapter could be used to include more input variables. These are inputs that could be used to estimate the fuel consumed by the truck more efficiently. The two important variables that can be obtained are acceleration of the vehicle and the slope of the road.
From Equation 1.1 stated earlier, the slope of the road can be calculated. Along with that, the acceleration of the vehicle can be calculated in a similar way:

$$Acceleration_i(m/s^2) = \frac{dV_i(m/s)}{dT_i(s)} \tag{5.1}$$

Where,
dV_i is the change in velocity of the vehicle between two consecutive instances and dT_i is the change in time between two consecutive instances.

5.1.2 Inclusion of the OpenStreetMap data in measurement data

As stated in Section 1.3, the GPS data obtained from the road test measurements need to be incorporated with the road data obtained from OpenStreetMap. Since the data obtained from the OpenStreetMap is not obtained at a fixed sampling frequency, there is a need to interpolate the data in a way that the distances are equally spaced, as already mentioned in Section 1.3. Along with that, the data obtained from the measurements does not include an important parameter, the speedlimit of the road. The data from OpenStreetMap is used for route profiling purposes.
It also contains information about the latitude, longitude and elevation of points along the route. This helps in the profiling of the route, specially to be used as inputs into

Model 2 as explained previously in Section 3.2.2. Since Model 2 has no measurement information, it becomes imperative for the model to know the driving route. The data from OpenStreetMap helps with this and provides a route for which the fuel prediction has to be made.
It is also imperative to know that since the training of the model was done for every 2 meters for Model 2, the sampling of the route profiling also needs to be done at 2 meters, i.e., one data-point for every 2 meters along the way. This is also achieved with the help of interpolation of the data-points obtained from OpenStreetMap. The detailed procedure for that is elaborated in Appendix A.

5.2 Filtering of Data

Filtering of data is the removal of noisy elements in the data which would help in improving the quality of the data. Although it might seem like an important step before the training of any machine learning algorithm, there have been studies that suggest the opposite. According to Guozhong An [42] and Chuan Wang et al. [43], some noise in the data helps in developing a better training set for the algorithms developed as it helps to reduce the over-fitting of the model, which in turn helps the model adapt better to new, unseen data-sets. This is the reason why this study does not use any filtering techniques for the measurement data. The randomness in the data-set due to noise would help in the generalisation of the models developed which in turn would be useful to adapt the model to newer routes and newer trucks.
However, this study also includes some input variables which are not directly obtained from the measurement data. These variables have been calculated and need to be filtered to make sure they do not contain unrealistic values. These variables are the slope of the road and the acceleration of the vehicle. For instance, the slope of the road as calculated from road measurements is as follows:

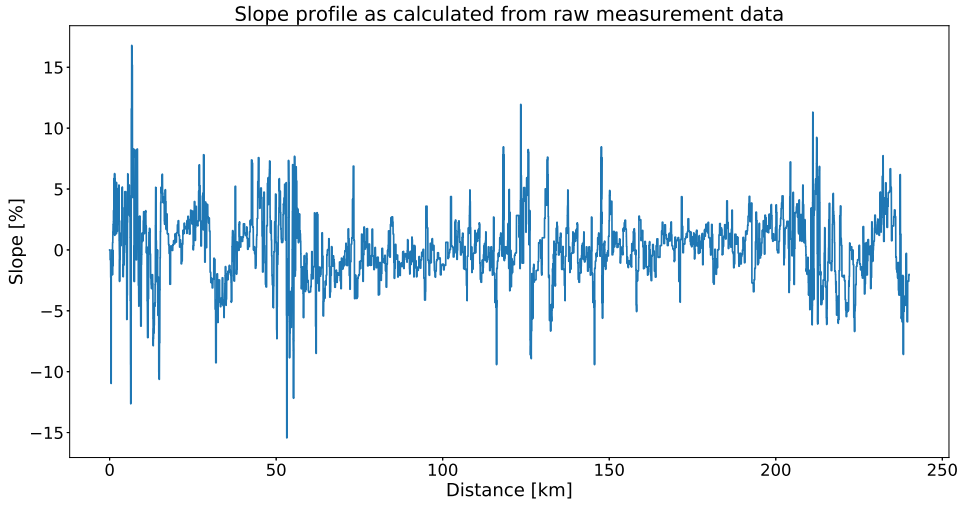


Figure 5.1: Unfiltered slope profile from measurement data

As can be seen in Figure 5.1, the slope plot shows unrealistic values. These values are not possible on a real road and can be explained due the small errors in measurements when the elevation changes by a very small value but the distance does not change. Slopes on motorways should not exceed 7% [44] and the Figure 5.1 clearly does not follow that. This is why this input variable needs to be filtered. The technique used for filtering is the

butter-worth low pass filter [1].
To filter out the high-frequency noise, the frequency spectrum of the slope profile was analysed. Frequency Spectrum of any signal is the range of frequencies contained in the signal. The only difficulty is the choice of the cutoff frequency. The frequency spectrum of the slope profile is hence analysed in order to find a cutoff frequency. It looks as follows:

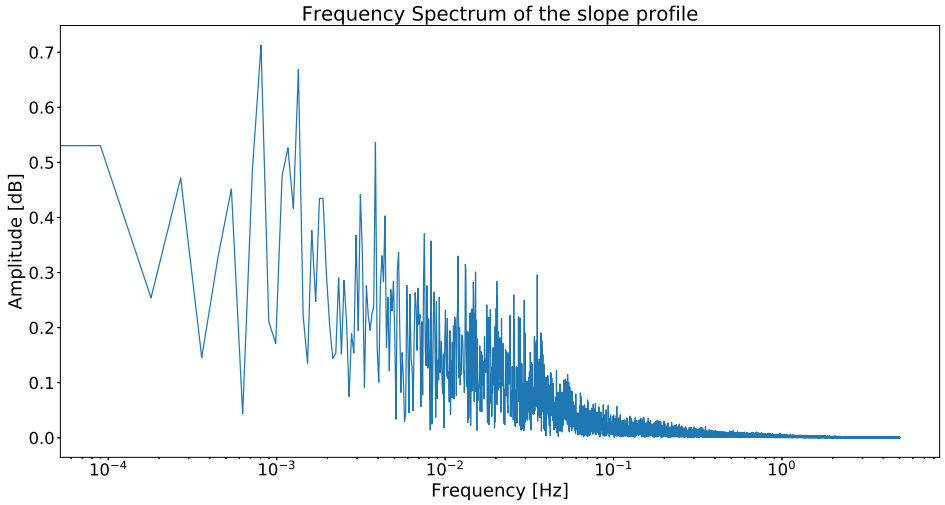


Figure 5.2: Frequency Spectrum of the slope profile

From Figure 5.2, a cutoff frequency needs to be chosen from this to remove the effects of high-frequency noise. There is always a trade-off to choose the right value of cutoff frequency as the higher the frequency one chooses as the cutoff, the more is the noise retained in the signal, whereas if the value of cutoff is too low, vital information in the signal is lost. The main aim was to smoothen the slope profile in order to make it realistic while keeping the elevation profile of the road as close to the original as possible. This can be done using a statistical measure called the explained variance score. It is the ratio of variance in the dependent variable as predicted/estimated from the independent variable and can be expressed mathematically as:

$$Explained\ Variance = \left[1 - \frac{Var(y_{predicted} - y_{actual})}{Var(y_{actual})}\right] \tag{5.2}$$

In this context, the $y_{predicted}$ and y_{actual} can be treated as the elevation values and Var is the variance which can be expressed as:

$$Var(x) = \frac{1}{n}\sum_{i=1}^{n}\big(x_i - mean(x)\big) \tag{5.3}$$

The higher the explained variance score, the higher is the amount of information retained in the predicted signal when compared to the actual signal with the values always lying between 0 and 1. Since slope is derived from the elevation at each instance, it is imperative that the information contained in the elevation profile is not lost whilst smoothening the slope profile. As seen from Figure 5.2, the amplitude seems to start oscillating from a frequency of about 0.007 Hz. The variance score of the elevation profile for the cutoff frequencies around that value can be expressed in the table below:

[1]More Information about the butterworth low pass filter can be found in Appendix C

Table 5.1: Explained Variance Score for elevation at different cutoff frequencies

Cutoff Frequency [Hz.]	Explained Variance Score
0.003	0.916
0.005	0.959
0.007	0.975
0.009	0.983
0.012	0.989

As mentioned earlier, it is sort of a trade-off to choose the correct cutoff frequency value. The value of cutoff frequency in this study was chosen to be 0.007 Hz. The variance in the elevation profile is not too high as to lose vital information about the road. For lower cutoff frequencies, although the slope profile was smooth, vital information was lost as seen from the variance score of the elevation. For higher cutoff frequencies, although the variance score was very high, the slope profile was not smooth enough and was not within the limits. The corresponding smoothened slope profile for a cutoff frequency of 0.007 Hz. looks as follows:

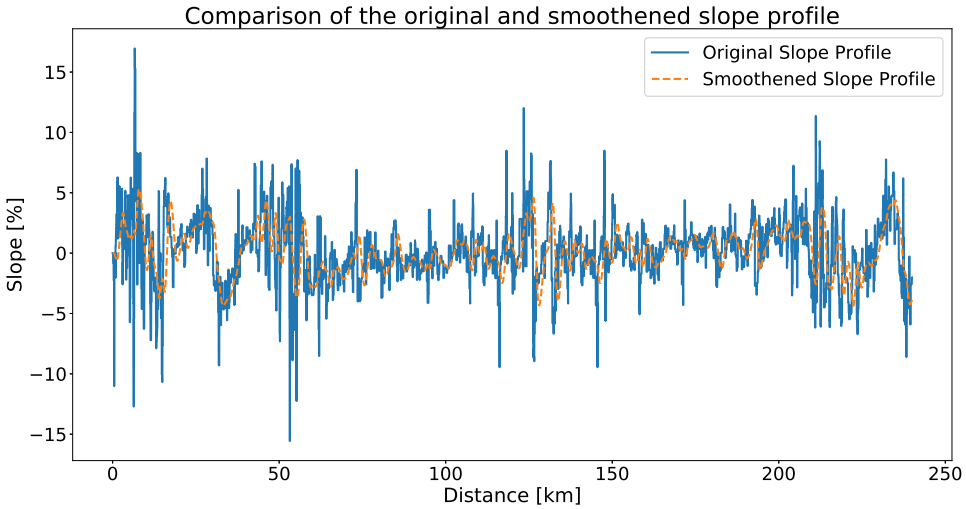


Figure 5.3: Comparison of the smoothened and original slope profiles

Figure 5.3 shows that the slope profile has been smoothened and is within the limits as suggested by literature.
A similar process is used for the smoothening of the acceleration profile. Since acceleration is derived from the velocity profile, the variance score of the velocity is aimed at maintaining high while simultaneously bringing the acceleration within acceptable limits. Acceleration of a loaded heavy duty truck cannot exceed 0.4 m/s^2 at lower velocities (30 km/h to 50 km/h) and 0.18 m/s^2 for higher velocities ($>$ 60 km/h) [45]. This would be treated as limits for the acceleration profile. Although Woodrow Poplin's paper titled, "Acceleration of Heavy Trucks" suggested a similar value for a heavy truck, it was not cited here as his paper was not subjected to a peer-review. The acceleration values obtained were found to be extremely unrealistic and needed to be dealt with to make sure the acceleration is within acceptable limits. The acceleration plot calculated from the raw measurement data looked as follows:

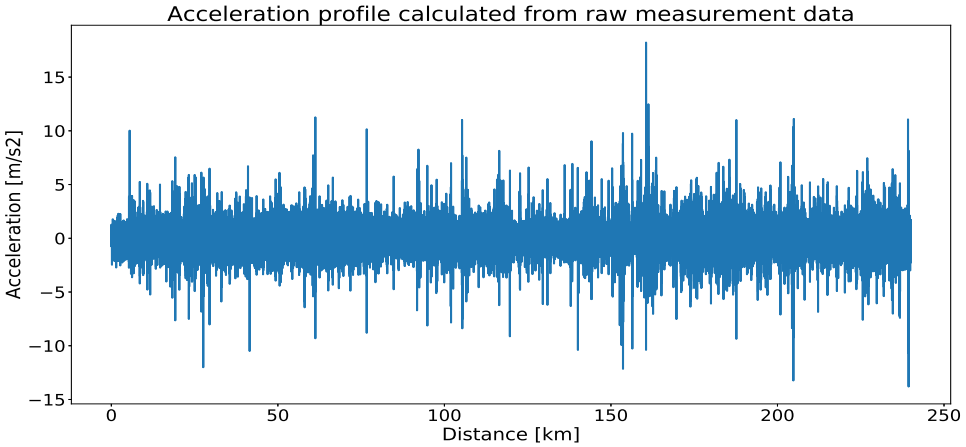


Figure 5.4: Unfiltered acceleration profile calculated from measurement data

As seen from Figure 5.4, the acceleration values exceed those that are realistically possible. To filter these unrealistic values, a similar procedure as the slope smoothening is performed and the following table gives the values for the explained variance scores for different cutoff frequencies in the velocity profile:

Table 5.2: Explained Variance Score for velocity at different cutoff frequencies

Cutoff Frequency [Hz.]	**Explained Variance Score**
0.01	0.796
0.015	0.848
0.02	0.88
0.025	0.89

The same dilemma as in the previous case occurs here. The cutoff frequency in this case was chosen to be 0.02 Hz. as the variance from velocity is not so high, which goes on to say that the vital information in the signal is not lost whilst making sure the acceleration profile is smoothened. Although higher cutoff frequencies do give higher variance score, the acceleration was still not realistic for those cutoff frequencies. A cutoff frequency of 0.02 Hz. seemed the most appropriate choice for this data-set. The final acceleration plot looks like the following:

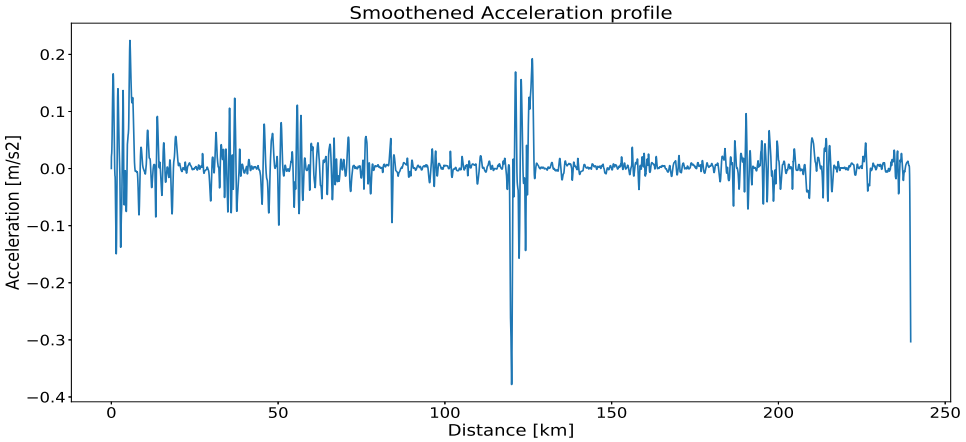


Figure 5.5: Smoothened acceleration profile after filtering

As can be seen from Figure 5.5, the acceleration is within limits as stated in literature and hence can now be used as a valid input for the model.
Once the data-set has been filtered, it can be used as input for the various learning algorithms.

5.3 Summary

This chapter dealt with the pre-processing of the data, especially on the adding of new input variables in the data-set and the inclusion of input variables from one source into the training set (i.e., from OpenStreetMap to the training-set). Finally, it concluded with the filtering of some of the input variables (slope and acceleration). The butterworth low-pass filter was used for the filtering of the two input variables to make sure the values are within a range that is possible in reality.

Chapter 6

RESULTS

This chapter gives an insight into the results obtained during this study. It elaborates on shuffling, splitting and scaling of data followed by the analysis of the consolidated data-set and detailed analysis of the results obtained from each of the machine learning algorithms and their comparisons. The chapter also provides light on the variable importance, i.e., the influence of each of the input variables on the output variable (fuel consumption) and how strongly each one of them affects the fuel consumption of the vehicle.

6.1 Data Splitting and Shuffling

The measurement data is first split into training, cross-validation and test data-sets. The training set is the largest of the lot which is used to train the models, the cross-validation is one which is used to help tune the models while the test data-set is used to test the performance of the developed models. Although the choice of splitting is arbitrary, this work splits the measurement data as 75% training set and 25% test-set. The 75% training set is further divided into 60% training data and 15% cross-validation set. Once the training is done on the training set, the cross-validation set is used to tune the parameters of the models and the test-set is finally used to check the performance. Following the splitting, the data is shuffled so that the learning of the algorithm is not affected due to continuous time evolving variables. This shuffling is done randomly to make sure the model does not learn the general trend of evolution of the output variable.

6.2 Scaling data

The values of the data measured from the road tests vary in magnitudes and units. It is important to know that machine learning algorithms rely on the euclidean distance between two data-points to arrive at a possible conclusion or prediction. This leads to a lot of trouble especially when variables are in different units and have varying magnitudes. An important step to implement in any machine learning algorithm is feature scaling wherein the features are scaled to a set range of values. This removes any discrepancy that might occur due to differences in magnitude of values or their units. The method used for scaling in this work is the standard scaling. It can be represented as follows:

$$x' = \frac{x - \mu}{\sigma} \tag{6.1}$$

Where,

x'	Feature scaled value
x	Feature raw value
μ	Mean value of the feature array
σ	Standard Deviation of the feature array

The Equation 6.1 makes sure the features are scaled so that they are not affected by differences in magnitude or units.
For the sake of building good training ans testing sets, the measurement data is first filtered (acceleration and slope, with regard to this study - to make sure the values do not exceed a certain limiting value) followed by splitting, shuffling and scaling of data. It is important to follow this chronological order as this way, the models do not just learn the evolution of the trend of the fuel and also, the testing data-set is a completely new, unseen data-set.

6.3 Results from each learning algorithm

This section delves into the results obtained from each learning algorithm for both aforementioned models as mentioned in Section 3.2 and their performance on the test data-set.

6.3.1 Linear Regression

Linear Regression is the easiest of the models to build and requires no tuning. It is usually implemented first to get an idea about the data and the performance of machine learning algorithms on a new application. In this case, linear regression was tested both for Model 1 and Model 2 and the results obtained are as follows:

Model 1: Results when model is trained using all the input variables, namely driver/drive influenced, vehicle parameters, road parameters and weather parameters

Model 1 is to be used as an online prediction model which could help in the prediction of the amount of fuel consumed by the vehicle in the next time step/distance step. The test set of 60 kms (which is approximately 28000 data-points) was used to check the performance of the model. The comparison of the estimates by Linear Regression and the Measurement data is as follows:

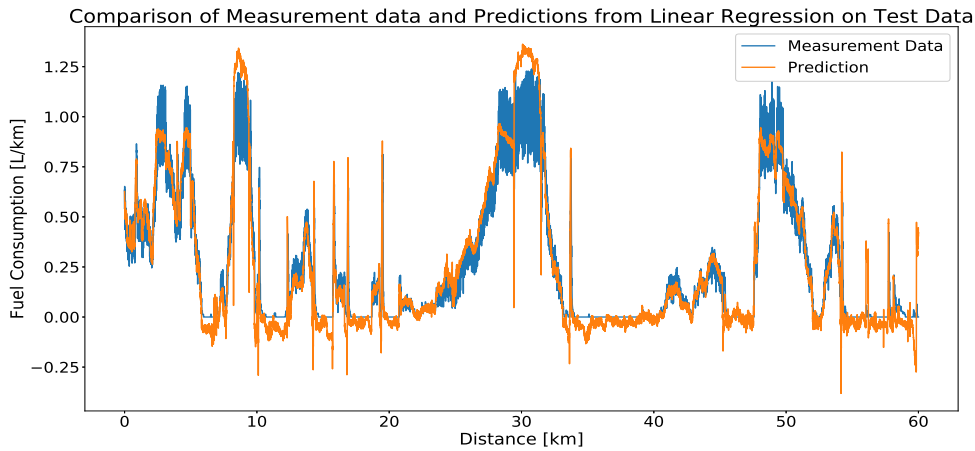


Figure 6.1: Comparison of fuel consumed as predicted by Linear Regression for Model 1

The Figure 6.1 shows that the model is able to follow the measurements well except for the high variations. This could be attributed to the noise in the measurement. But, the model

is able to give a good aggregate of the fuel consumption. The cumulative fuel consumed for the entire trip is as below:

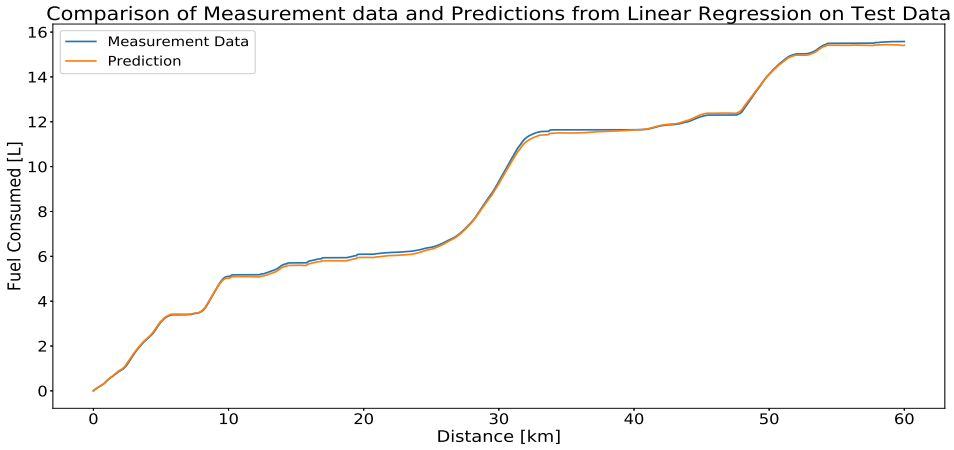


Figure 6.2: Comparison of the cumulative fuel consumed as predicted by Linear Regression for Model 1

The Figure 6.2 shows the cumulative fuel consumed by the truck and that predicted by the model. The model is quite accurate in predicting the amount of fuel consumed. The following table further elaborates on the results obtained:

Table 6.1: Results from Linear Regression for Model 1

Model	**RMSE ±2 % [L/km]**	**Total Fuel Consumed ±2 % [L/km]**	**Absolute Error in total Fuel Consumed ±2 % [L/km]**	**Total Fuel Consumed [L]**
Measurement Data	-	0.259	-	15.58
Linear Regression	0.097	0.256	0.003	15.48

The ±2 % variation is to account for the learning of the algorithm. The weights assigned by the algorithm during training to the different input variables will not be the same for each run. The ±2 % accounts for that difference.
The Table 6.1 shows the values for the fuel consumed by the truck as measured and as predicted by the model. The absolute error in the total fuel consumed of 0.003 shows that the model is accurate in predicting the amount of fuel consumed. The Figures 6.1 and 6.2 also show that the prediction seems to be working well. The RMSE value is the most important as this would denote how much the model deviates from the measurement data.

Model 2: Results when the model is trained using only vehicle and road parameters

The potential of the model on a new route can be tested with Model 2. It would show how adaptive the model is, to newer, unseen routes. The exclusion of the diver related inputs from the model is a lucrative method to ensure the predictions of fuel consumed could take place without the necessity of driving on the road. The following figure shows how well the model can predict the total amount of fuel consumed during the entire trip (test-set) when the driver-influenced input parameters are not considered.

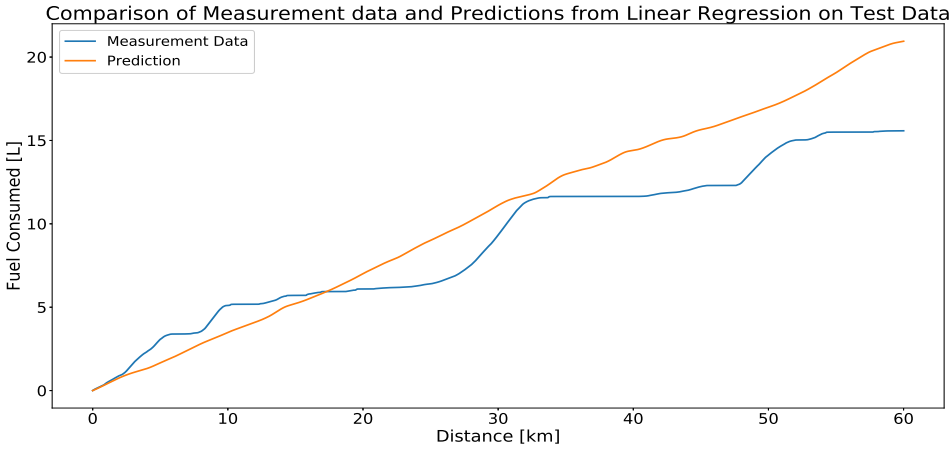


Figure 6.3: Comparison of the cumulative fuel consumed as predicted by Linear Regression for Model 2

The following table further elaborates the results obtained on the test set of 60 kms.:

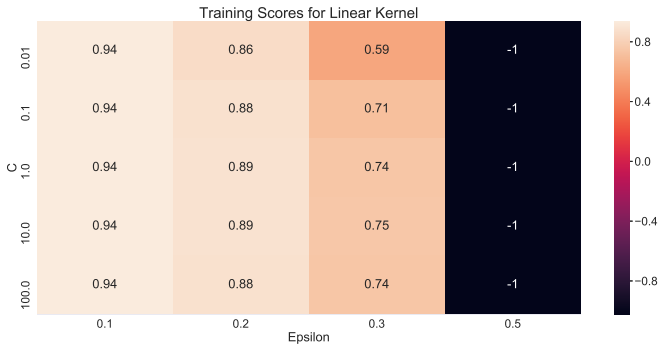
Table 6.2: Results from Linear Regression for Model 2

Model	**RMSE** ±2 % **[L/km]**	**Total Fuel Consumed** ±2 % **[L/km]**	**Absolute Error in total Fuel Consumed** ±2 % **[L/km]**	**Total Fuel Consumed [L]**
Measurement Data	-	0.259	-	15.58
Linear Regression	0.35	0.349	0.089	20.95

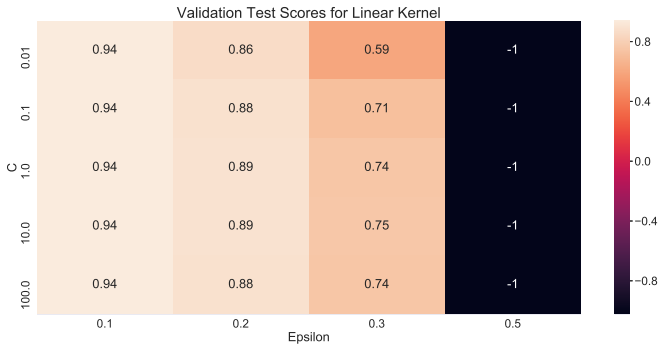
As mentioned earlier, the ±2 % is to account for the differences in weights assigned by the algorithm to the different input parameters for different runs.
Since Model 2 is to be used for the overall fuel prediction of the entire route, it is less significant to report the RMSE errors in this case as the model is not to be used as an on-board prediction model but as a model that could predict the amount of fuel consumed by the truck on a route without actually driving on it. The high RMSE value also proves that this modelling is not good for on-board fuel consumption estimation. The total fuel consumed during the entire trip is the target the model sets to achieve. The linear approximation goes on to say that this algorithm is incapable of learning for this kind of modelling.
Linear Regression seems to be working well for Model 1 but fails for Model 2. Hence, Linear Regression is not capable of performing well when the model becomes more complex and learning is more difficult.

6.3.2 Support Vector Regression (SVR)

SVR, unlike linear regression, needs a lot of tuning. The three parameters that need to be tuned as mentioned earlier are Epsilon, C and the choice of kernel. This is done with the help of an exhaustive grid search method. A range of values for epsilon and C are tested on the training and validation set for each of the kernels and the one with the highest score is chosen as the appropriate parameter.
The training and cross-validation scores for the linear kernel are as follows:



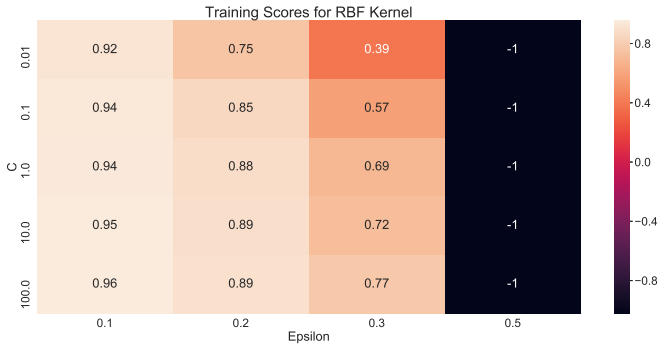
(a) Training Scores



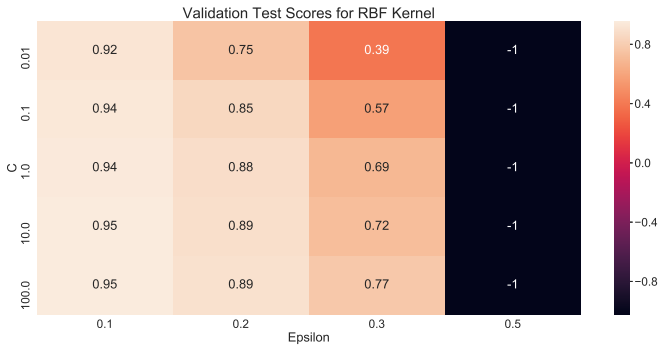
(b) Cross-Validation Scores

Figure 6.4: Training and Cross-Validation scores for Linear Kernel

The training and cross-validation scores for the RBF kernel are as follows:



(a) Training Scores



(b) Cross-Validation Scores

Figure 6.5: Training and Cross-Validation scores for RBF Kernel

It is evident from Figures 6.4 and 6.5 that the Gaussian Radial Bias Kernel (RBF) is the better choice for the kernel. It scores better on the cross-validation set. The values of epsilon and C for the RBF kernel were chosen to be 0.1 and 100 respectively. Having chosen the suitable tuning parameters, the algorithm can now be used for the prediction of the fuel consumption on the test set.

Model 1: Results when model is trained using all the input variables, namely driver/drive influenced, vehicle parameters, road parameters and weather parameters

The performance of the SVR algorithm was checked on the test-set of 60 kms. The following figure shows the comparison of the predictions made by the algorithm and the measurement data:

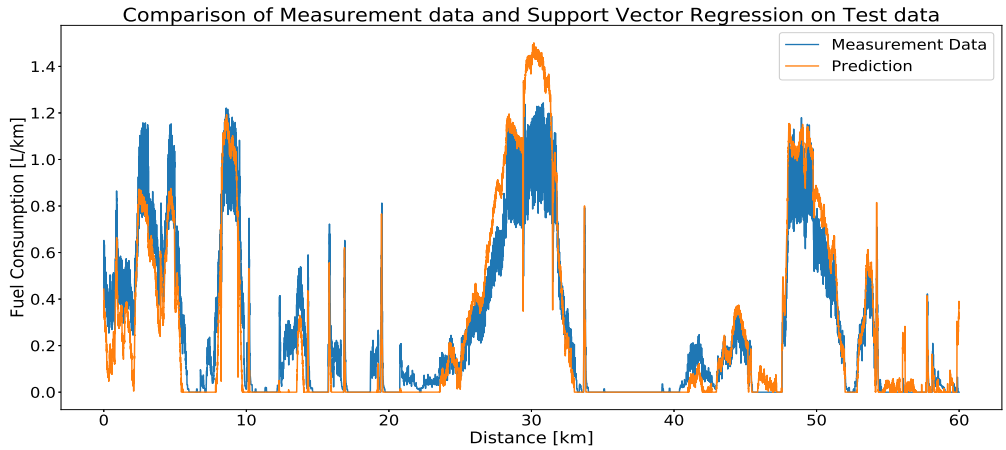


Figure 6.6: Comparison of fuel consumed predicted by Support Vector Regression algorithm for Model 1

The Figure 6.6 shows that the SVR algorithm tends to underestimate the fuel consumption for most of the route while it also overestimates for a certain part. The prediction of the cumulative fuel consumption by the truck would give better insight into the performance of the algorithm.

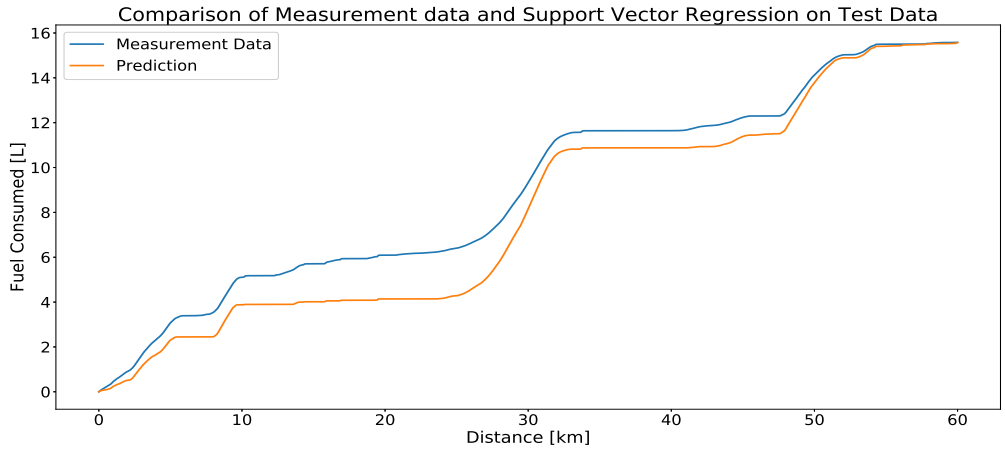


Figure 6.7: Comparison of the cumulative fuel consumed predicted by Support Vector Regression algorithm for Model 1

Figure 6.7 shows that the model underestimates for most part of the route. Although the

algorithm seems to be working fine, it looks to perform worse than Linear Regression. The following table gives a better insight into the performance of the algorithm:

Table 6.3: Results from Support Vector Regression for Model 1

Model	**RMSE ±2 % [L/km]**	**Total Fuel Consumed ±2 % [L/km]**	**Absolute Error in total Fuel Consumed ±2 % [L/km]**	**Total Fuel Consumed [L]**
Measurement Data	-	0.259	-	15.58
SVR	0.13	0.259	0.0001	15.57

As mentioned earlier, the RMSE value is a good measure of the performance of the algorithm when compared to the measurement data and the value is a little higher than the one calculated from Linear Regression.

Model 2: Results when the model is trained using only vehicle and road parameters

Model 2 tests the potential of the algorithm on the ability to predict the fuel consumption of the truck without it actually driving on the road. The following figure shows how well the model can predict the cumulative fuel consumed during the entire trip (test-set):

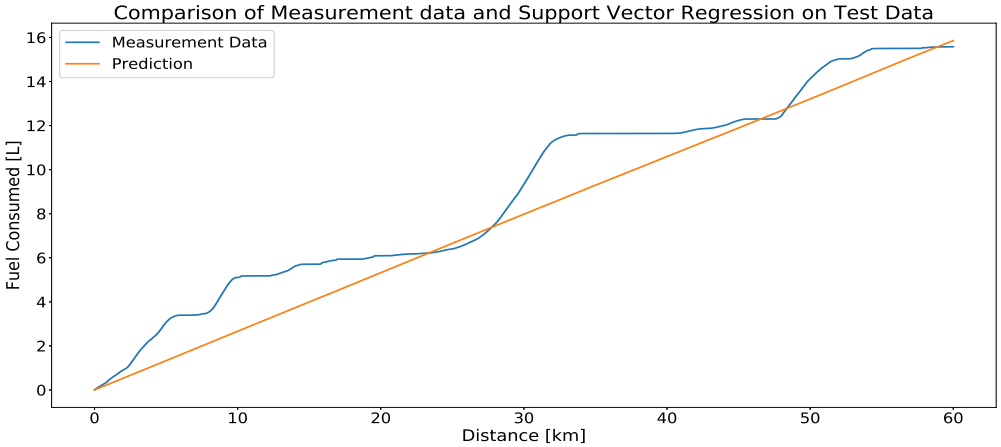


Figure 6.8: Comparison of the cumulative fuel consumed predicted by Support Vector Regression algorithm for Model 2

The following table further elaborates on the results obtained:

Table 6.4: Results from Support Vector Regression for Model 2

Model	**RMSE ±2 % [L/km]**	**Total Fuel Consumed ±2 % [L/km]**	**Absolute Error in total Fuel Consumed ±2 % [L/km]**	**Total Fuel Consumed [L]**
Measurement Data	-	0.259	-	15.58
SVR	0.33	0.264	0.004	15.85

As can be seen from Figure 6.8, the SVR does not perform so well as it only provides a linear approximation of the measurement data. Although the prediction for the total fuel consumed by the truck is predicted well by the algorithm, the fact that it only provides a linear approximation would be a problem.
Just like Linear Regression, the SVR algorithm also seems to be a decent option for the prediction of fuel consumption as done in Model 1 but does not translate well to Model 2. It goes on to say that SVR is unable to perform well when the conditions becomes more difficult and the learning is more complex but is a possible solution when the fuel consumption is to be predicted for the next time step.

6.3.3 Random Forest

Random Forest is also relatively easy to tune as there is need for only one parameter tuning, the number of estimators. The number of estimators is the number of trees in the model as described earlier in Section 3.3.3.
The tuning of this parameter is done by increasing the number of trees and checking the performance for each iteration. This is done in the following way:

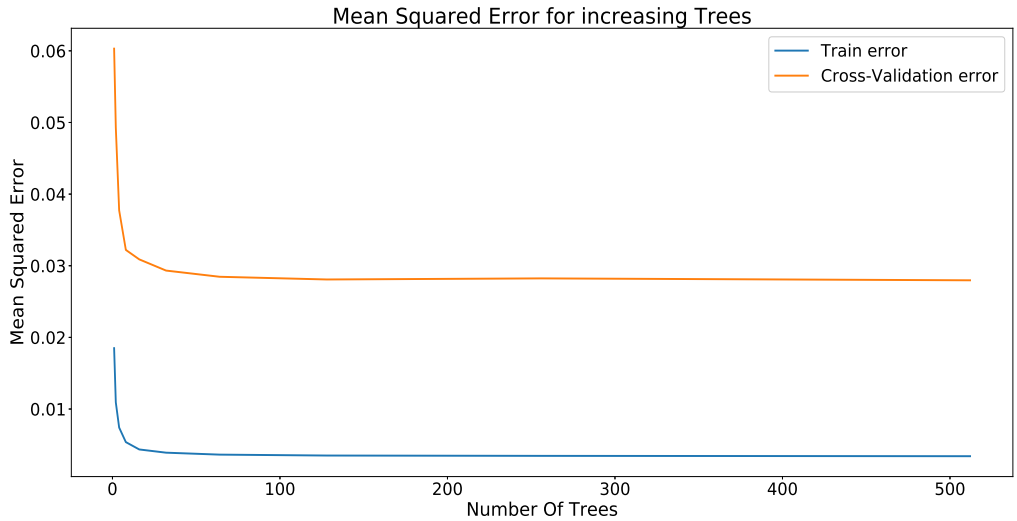


Figure 6.9: Mean Squared Error for Random Forests with increasing number of trees

Figure 6.9 shows that the number of trees does affect the performance of the algorithm. As stated earlier, higher the number of trees in the algorithm, the better are the chances of the algorithm to adapt to newer conditions. Hence, the number of trees is chosen as 200. This is not too high to affect the computational complexity but also not too low as to affect the performance of the algorithm. Hence, the number of trees considered for this study is 200.

Model 1: Results when model is trained using all the input variables, namely driver/drive influenced, vehicle parameters, road parameters and weather parameters

After the tuning of the algorithm, it can be used for the prediction of the fuel consumed by the tractor-semitrailer. As mentioned earlier, Model 1 is an online monitoring/prediction tool to predict the fuel consumption for the next time/distance step. The 60 kms of test-set is used to check the performance of the Random Forest Algorithm. The following figure shows how well the predictions from Random Forest algorithm compare to the Measurement data:

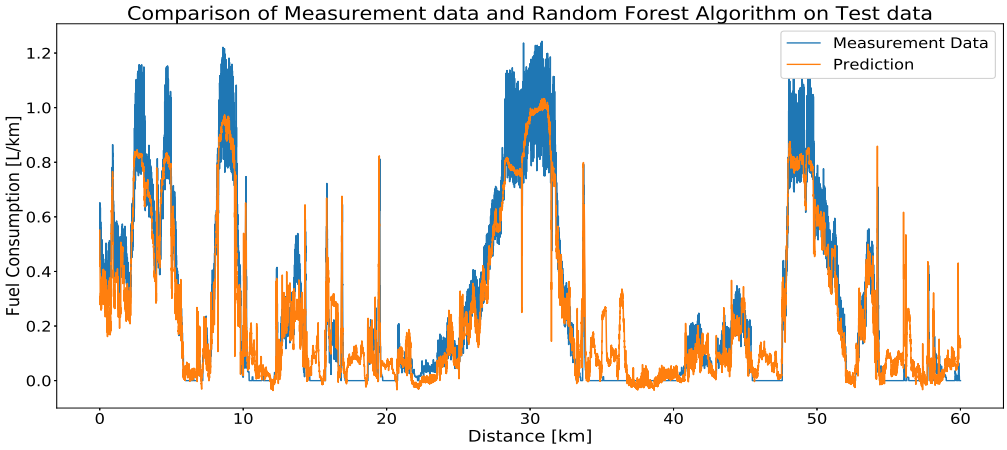


Figure 6.10: Comparison of fuel consumed as predicted by Random Forest for Model 1

The Figure 6.10 is able to capture the high variations of fuel consumption better than Linear Regression but struggles to perform at constant accelerations (the fuel consumption is almost zero as the truck would be cruising.) The following figure shows the cumulative fuel consumed as a function of distance as a function of the distance.

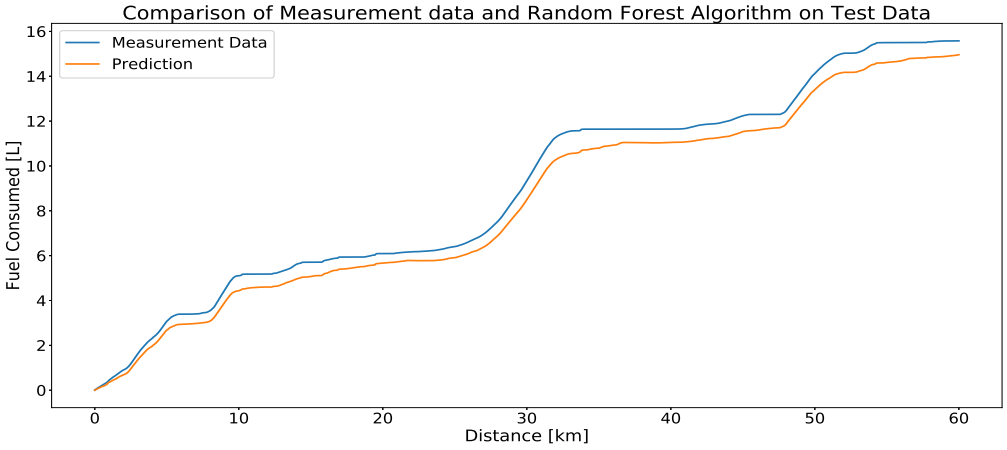


Figure 6.11: Comparison of the cumulative fuel consumed as predicted by Random Forest for Model 1

The Figure 6.11 works moderately well to estimate the amount of fuel consumed for the entire trip. It is able to capture the trend in the cumulative fuel consumption well. The following table further elaborates on the results obtained:

Table 6.5: Results from Random Forest for Model 1

Model	**RMSE** ±2 % **[L/km]**	**Total Fuel Consumed** ±2 % **[L/km]**	**Absolute Error in total Fuel Consumed** ±2 % **[L/km]**	**Total Fuel Consumed [L]**
Measurement Data	-	0.259	-	15.58
Random Forest	0.098	0.249	0.0102	14.96

It can be seen from the Table 6.5 that the Random Forest Algorithm performs comparably

to the Linear Regression Model and their RMSE values are almost the same with Linear Regression performing slightly better.

Model 2: Results when the model is trained using only vehicle and road parameters

Model 2 is where the learning ability of the model can be tested as the driver influenced inputs are left out. The following figure shows the cumulative fuel consumed as predicted by the model and compared with the measurement data.

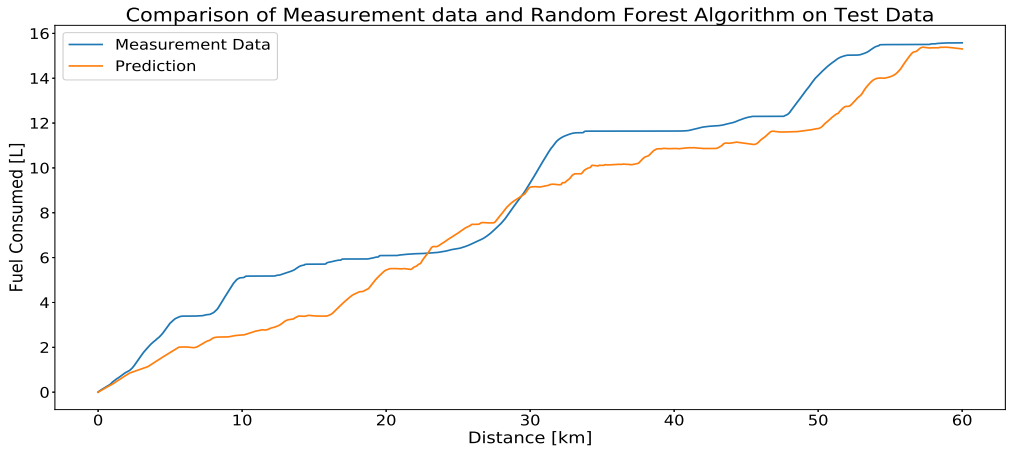


Figure 6.12: Comparison of the cumulative fuel consumed as predicted by Random Forest for Model 2

The Figure 6.12 shows that the Random Forest Algorithm works much better than the Linear Regression and the SVR models. It is able to somewhat capture the trend of the fuel consumed on the road and not just predict a linear approximation. The following table further elaborates the results obtained on the test set of 60 kms.:

Table 6.6: Results from Random Forest for Model 2

Model	**RMSE ±2 % [L/km]**	**Total Fuel Consumed ±2 % [L/km]**	**Absolute Error in total Fuel Consumed ±2 % [L/km]**	**Total Fuel Consumed [L]**
Measurement Data	-	0.259	-	15.58
Random Forest	0.42	0.255	0.004	15.3

It can be seen from Table 6.6 that the total fuel consumed as predicted by Random Forest is better than that predicted by Linear Regression. Although the RMSE value is higher, it is of a lesser significance for this kind of modelling as mentioned before. The error in the values predicted for the measurement value is quite small. Also, the algorithm is somewhat able to predict the trend of the fuel consumption which was seen missing in both Linear Regression and SVR. This goes onto say that the Random Forest algorithm is capable of performing well when the model becomes more complex and learning is more difficult.
Random Forest seems like a potent algorithm for the prediction of fuel consumption for Model 1 and performs moderately well for Model 2. It is capable of learning better than the Linear Regression and SVR algorithms.

6.3.4 Neural Networks

The tuning of a neural network model is more challenging than the other algorithms. But literature including Bratislav Predić et al. [14] and others have mentioned that it is highly potent and capable. In order to exploit that, the model needs to be tuned by using training and testing errors and checking for appropriate parameters. The first parameters to be tuned are the number of nodes in the hidden layer along with the dropout rate. The following figure gives some insight into it:

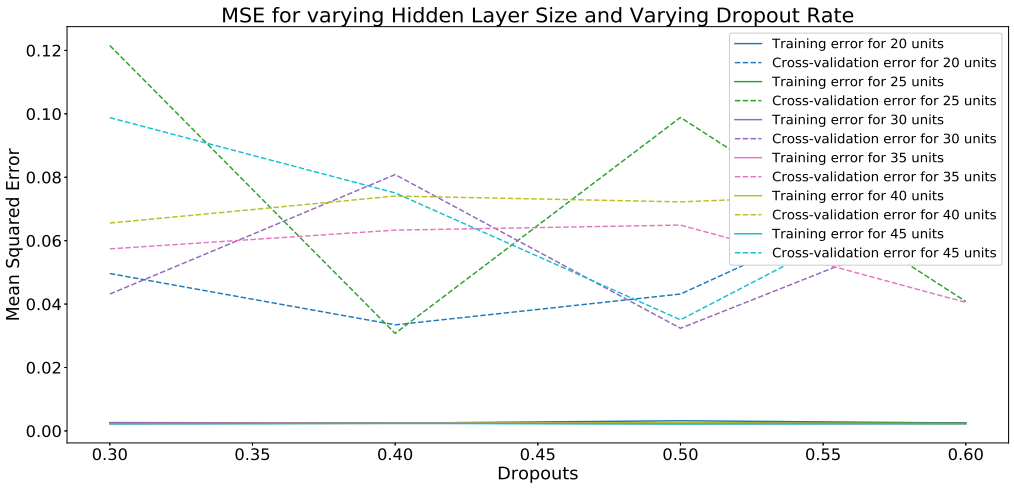


Figure 6.13: Mean Squared Error for the training and cross-validation set for Neural Network with varying nodes in hidden layer and dropout rate

It can be seen from Figure 6.13 that the best possible choice for the number of nodes in the hidden layer is 30 or 45 nodes with a dropout rate of 0.5. These values are a right mix where the model seems to be trained enough to avoid both underfitting and overfitting. All other alternatives seem to be facing one of the problems. The number of nodes is therefore chosen to be 30 instead of 45 as 45 nodes in the hidden layers would only make the model more computationally expensive which is something that is always avoided.
The third parameter to tune is the number of hidden layers. With an increasing number of layers, the model becomes more complex and more capable of learning more complex tasks but it also becomes more computationally expensive with iterations taking a longer time with increasing layers.

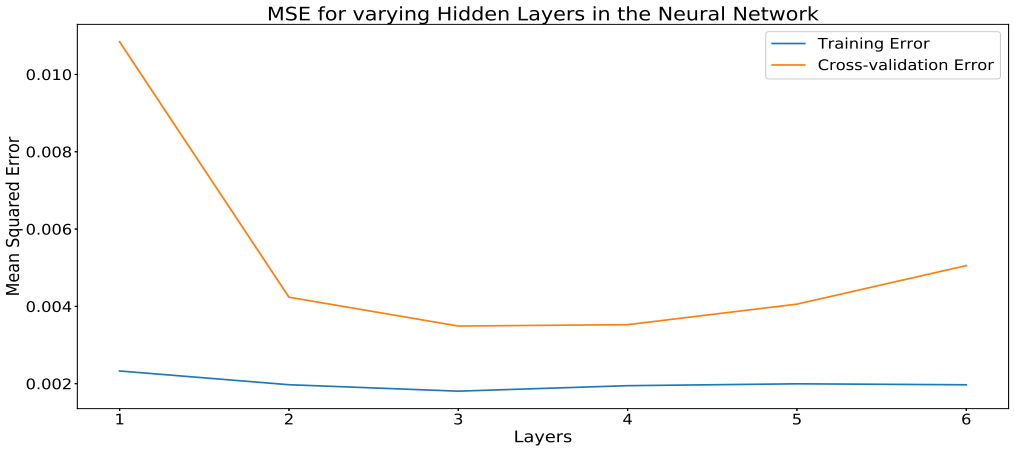


Figure 6.14: Mean Squared Error for the training and cross-validation set for Neural Network with increasing hidden layers

It can be concluded from Figure 6.14 that the cross-validation error increases with an increase in the number of hidden layers in the latter half of the figure. It also tends to increase the computation time. For this purpose, 4 hidden layers is chosen. From Figures 6.13 and 6.14, it can be inferred that a network with 4 hidden layers with 30 units in each at a dropout rate of 0.5 would be the most appropriate choice. The final parameter to tune is the number of epochs which is the number of times the algorithm is allowed to look at the training data. With increasing the number of epochs, there is a risk of the model overfitting. The error figure for the number of epochs is as follows:

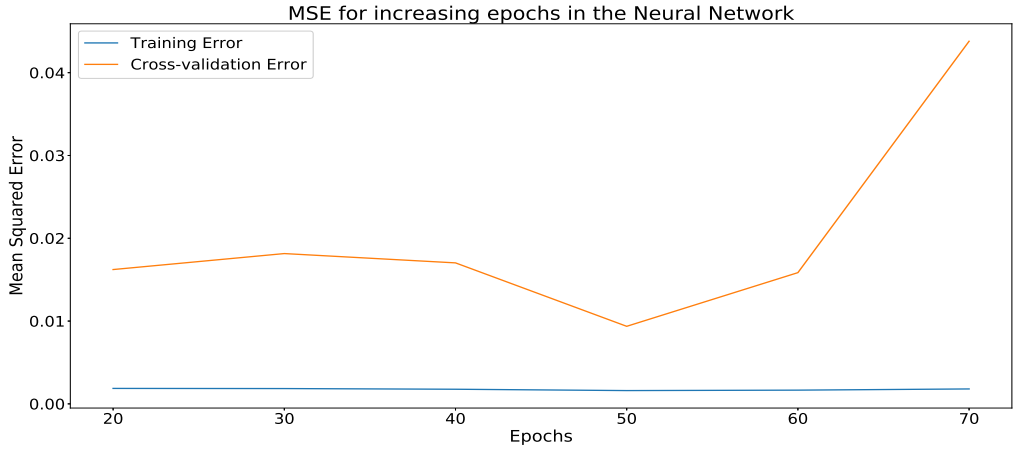


Figure 6.15: Mean Squared Error for the training and cross-validation set for Neural Network with increasing epochs

Figure 6.15 gives the value for the final parameter, the number of epochs that the learning algorithm needs to perform the best. This value was chosen as 50 as the cross-validation error increases after that which goes on to say that the model overfits when the number of epochs is more than 50.
The following table shows the values of chosen tuning parameters of the Neural Network:

Table 6.7: Final Tuning Parameters of the Neural Network algorithm

Parameter	Value
Number of hidden layers	4
Number of nodes in hidden layer	30
Dropout rate	0.5
Number of Epochs	50

After choosing suitable parameters from the tests mentioned above the estimates predicted by the Neural Network model can now be compared to the measurement data and its performance can be checked.

Model 1: Results when model is trained using all the input variables, namely driver/drive influenced, vehicle parameters, road parameters and weather parameters

As in the case with the other learning algorithms, the performance of the Neural Network is also checked on the 60 kms. of test-set. The following figures show how well the Neural Network is able to predict the fuel consumption when compared to the measurement data:

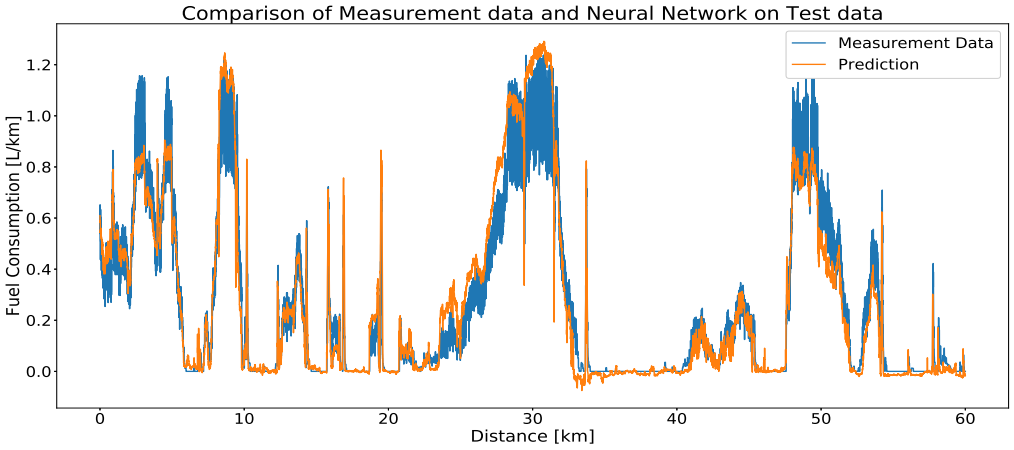


Figure 6.16: Comparison of fuel consumed as predicted by Neural Network for Model 1

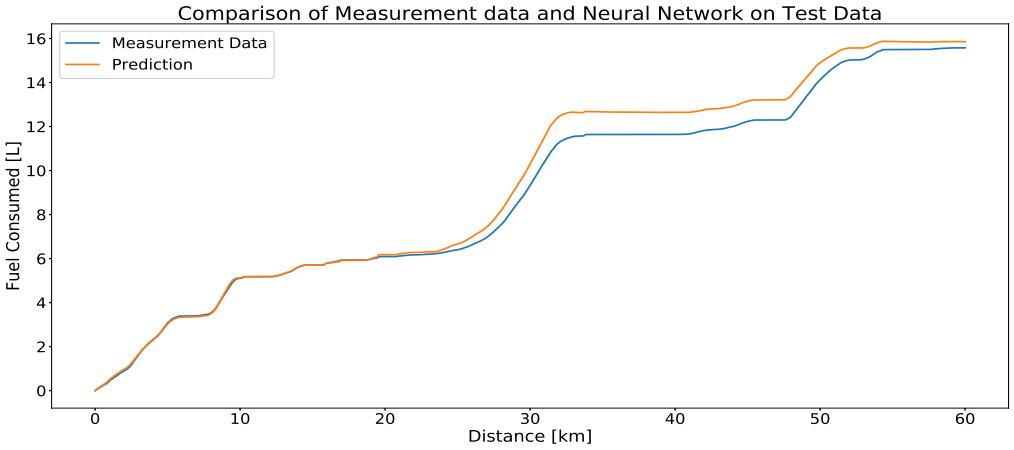


Figure 6.17: Comparison of the cumulative fuel consumed as predicted by Neural Network for Model 1

The Figures 6.16 and 6.17 show that a deep Neural Network could be used as a prediction technique too. The model slightly overestimates the actual measurements. The following table further elaborates on the results obtained:

Table 6.8: Results from Neural Network for Model 1

Model	**RMSE** ±2 % **[L/km]**	**Total Fuel Consumed** ±2 % **[L/km]**	**Absolute Error in total Fuel Consumed** ±2 % **[L/km]**	**Total Fuel Consumed [L]**
Measurement Data	-	0.259	-	15.58
Neural Network	0.092	0.264	0.005	15.85

It is clear from the results that Neural Networks could be used as a possible prediction technique as the fuel estimates predicted by the algorithm are close to the measured datapoints. The RMSE values obtained by this algorithm are the least when compared to the other learning algorithms. Along with that, the total fuel consumption is also close to the actual measurements. This goes on to say that Neural Networks would be an efficient algorithm to predict the fuel consumption of a truck.

Model 2: Results when the model is trained using only vehicle and road parameters

This section is dedicated to the performance of Neural Networks in cases where the driver related inputs are not considered in the model. The following figure shows how well the Neural Network predicts the total fuel consumed by the truck for the entire test-set:

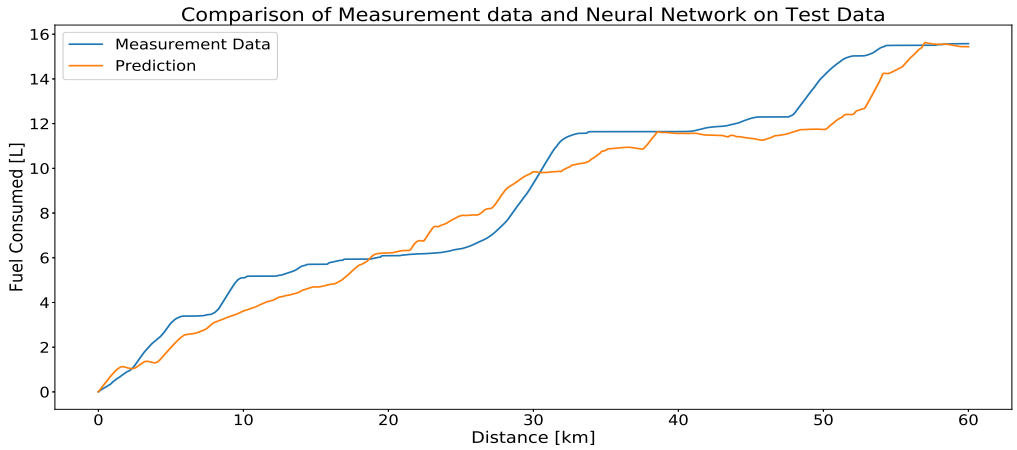


Figure 6.18: Comparison of the cumulative fuel consumed as predicted by Neural Network for Model 2

The Figure 6.18 shows that the Neural Network is able to predict the trend of the fuel consumption from the measurements. This already proves that the algorithm works better than the Linear Regression and the Support Vector Regression learning techniques since neither one of them could predict the trend in the fuel consumption. The following table further elaborates on the results obtained:

Table 6.9: Results from Neural Network for Model 2

Model	**RMSE ±2 % [L/km]**	**Total Fuel Consumed ±2 % [L/km]**	**Absolute Error in total Fuel Consumed ±2 % [L/km]**	**Total Fuel Consumed [L]**
Measurement Data	-	0.259	-	15.58
Neural Network	0.43	0.257	0.002	15.44

The Table 6.9 shows that the Neural Network is capable of predicting the total fuel consumed by the truck better than any other learning algorithm. It is also capable of performing better when the learning is more difficult.

6.4 Compilation and Comparison of Results

This section further elaborates on the results obtained from all the algorithms and compares their performances to find out the most potent one for each of the two models.

6.4.1 Comparison of results for Model 1

As mentioned earlier, Model 1 would act as an online prediction for the next step fuel consumption. This further stresses on the fact that the RMSE errors should be as low as

possible for the models to correctly predict the next step fuel prediction. The following table gives a comparison of all the learning algorithms with the measurement data:

Table 6.10: Comparison of all learning algorithms for Model 1

Model	**RMSE ±2 % [L/km]**	**Total Fuel Consumed ±2 % [L/km]**	**Absolute Error in total Fuel Consumed ±2 % [L/km]**	**Total Fuel Consumed [L]**
Measurement Data	-	0.259	-	15.58
Linear Regression	0.097	0.256	0.003	15.48
Support Vector Regression	0.13	0.259	0.0001	15.57
Random Forest	0.098	0.249	0.0102	14.96
Neural Network	0.092	0.264	0.005	15.85

The Table 6.10 confirms that the performance of 3 algorithms, namely, Linear Regression, Random Forest and Neural Network are similar with Neural Network performing slightly better than the other two with the RMSE value of 0.092. The error in total fuel consumed is also negligible which only goes in to say that the algorithms could be used for the online prediction of the fuel consumption for the next step. Although SVR has the lowest error in absolute fuel consumed, it has the highest RMSE value, which is important to know how well the model compares to the measurement data. Hence, SVR was seen to perform the worst among all the other algorithms with the highest RMSE.

6.4.2 Comparison of results for Model 2

Model 2 would act as the total fuel consumption prediction tool for the entire trip and not for the next step prediction. The following tables summarises the results obtained from the different algorithms for this model.

Table 6.11: Comparison of all learning algorithms for Model 2

Model	**RMSE ±2 % [L/km]**	**Total Fuel Consumed ±2 % [L/km]**	**Absolute Error in total Fuel Consumed ±2 % [L/km]**	**Total Fuel Consumed [L]**
Measurement Data	-	0.259	-	15.58
Linear Regression	0.35	0.349	0.089	20.95
Support Vector Regression	0.33	0.264	0.004	15.85
Random Forest	0.42	0.255	0.004	15.3
Neural Network	0.43	0.257	0.002	15.44

It is clear from Table 6.11 that the best algorithm for the prediction of total fuel consumed by the vehicle over the entire trip, without the vehicle actually driving on the road is Neural Network with an absolute error of 0.002. It was also able to predict the trend of the cumulative fuel consumption. Although SVR gives the least RMSE, it is only able

to predict a linear approximation of the fuel consumption as shown in Figure 6.8 while Neural Network is able to not just predict the total fuel consumed by the truck for the entire test-trip but also predict the trend of the fuel consumption.
The underlying problem with this is the fact that there is no source for any comparison. For the sake of checking the performance of the developed algorithms, they were compared to a simulation based model developed at TNO. The next chapter dives into the details of the results obtained from the simulation model and the ones obtained from the various machine learning algorithms developed in this study and a comparison is drawn between the two.

6.5 Variable Importance

An interesting aspect is to study the effect of the different variables in the measurement data on the fuel consumption of the vehicle. Linear Regression and Random Forests, both have a feature to find the importance of the variables in the process of prediction. The results from both will be analysed in this section.

6.5.1 Variable Importance for Model 1

Results from Linear Regression show the following dependence of the variables on the fuel consumption of the truck:

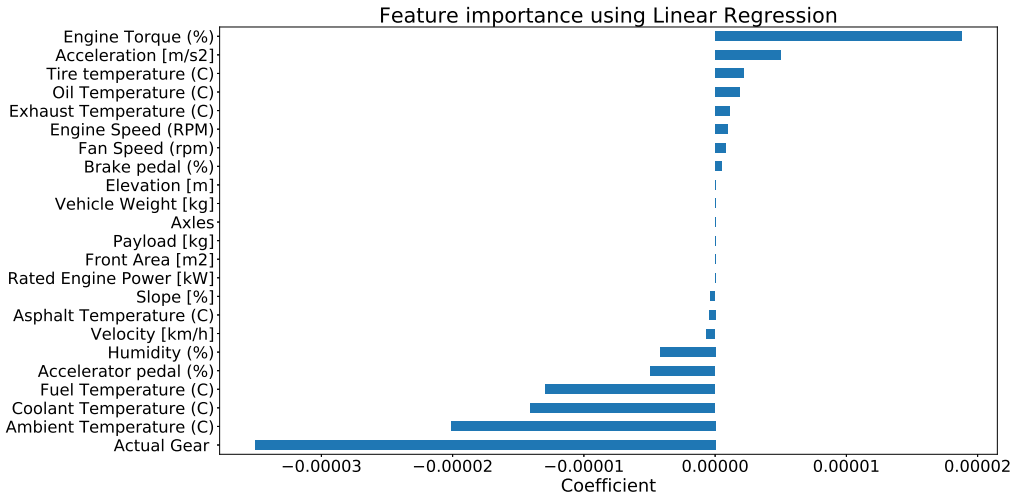


Figure 6.19: Feature Importance using Linear Regression for Model 1

It can be seen from Figure 6.19 that Linear Regression provides both positive and negative effects the variables (input variables) have on the target variable (fuel consumption, in this case). Positive effect is when the output variable increases with an increase in the input variable and vice versa. Negative effect is when the output variable increases with a decrease in the input variable. Investigating Figure 6.19 show that engine torque has the highest positive effect while the driving gear has the highest negative effect on the fuel consumption of the vehicle. Along with that, the features affected by the driver, i.e., acceleration and velocity too, have a high correlation to the amount of fuel consumed by the truck. From the knowledge of the system, it also makes sense that the engine torque affects the fuel consumption the most. The torque produced by the engine is the direct consequence of the fuel burning in the engine. Along with that, the actual driving gear, which affects the fuel consumption in the most negative way also makes good sense. A

vehicle would always consume more fuel at lower gears which is what is predicted by the models.
Results for the variable importance from Random Forests is as follows:

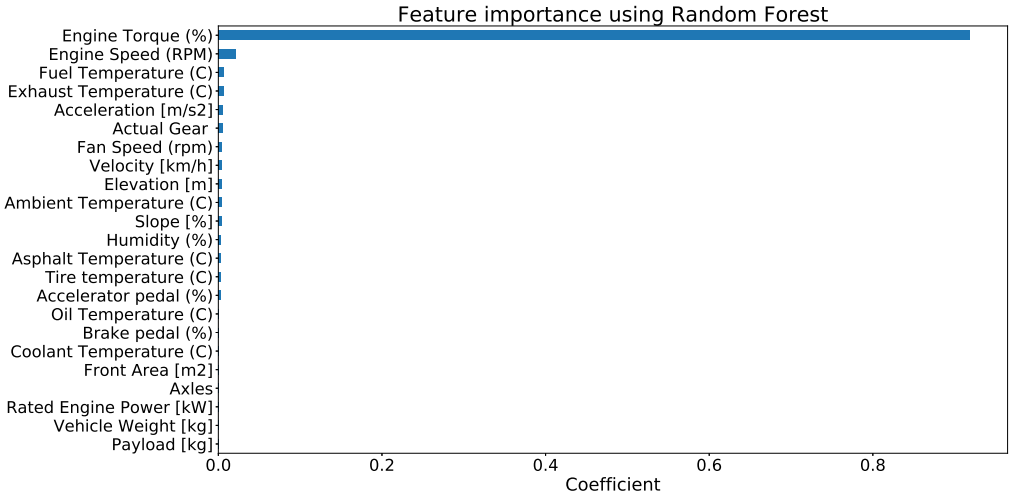


Figure 6.20: Feature Importance using Random Forests for Model 1

It is evident from Figure 6.20 that the Engine Torque is infact a very important feature in predicting the amount of fuel consumed. For a better visualisation on the importance of other variables, the following figure shows the dependence of the variables when the engine torque is not considered:

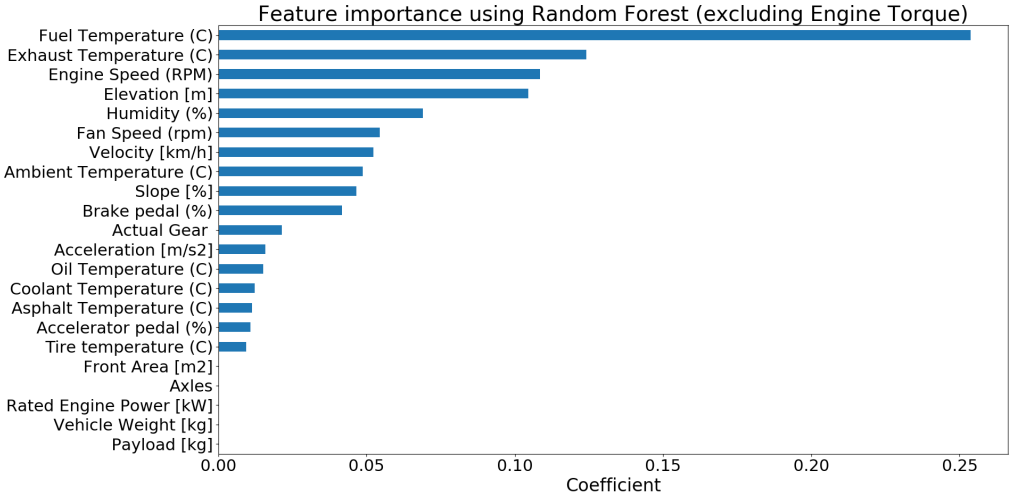


Figure 6.21: Feature Importance using Random Forests without considering the Engine Torque for Model 1

Random Forests give a better understanding of the dependence of fuel consumption on the different input variables. It can be concluded from Figure 6.20 and 6.21 that all the engine parameters have a high influence on the amount of fuel consumed. Along with that, the driver influenced inputs like acceleration and velocity of the vehicle along with the driving gear also affect the fuel consumption. Road attributes like elevation also have a strong correlation with the fuel consumption of the vehicle. Factors like fuel and exhaust temperature also have a high correlation to the amount of fuel consumed by the truck. The Figures 6.19 to 6.21 show that the weight of the vehicle and the payload have no effect on the fuel consumption of the vehicle. This might not be true in the real scenario but

the results show so as the training model consists of data only from one vehicle driven on one road. This voids any effect the vehicle has the amount of fuel consumed. This could be countered by conducting a similar study on more data about similar vehicles driven on different routes.

6.5.2 Variable Importance for Model 2

Results from Linear Regression show the following dependence of the variables on the fuel consumption of the truck:

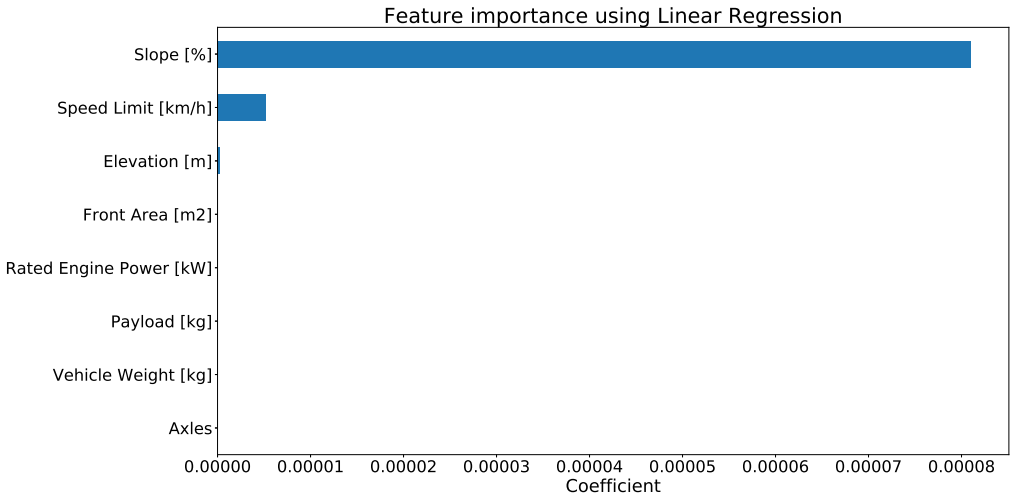


Figure 6.22: Feature Importance using Linear Regression for Model 2

Although it is known from previous results that Linear Regression is not the right algorithm for predictions from Model 2, it suggests that slope of the road acts as the most important variable to predict the fuel consumption of the vehicle.
The following figure gives an insight into the results obtained from Random Forest:

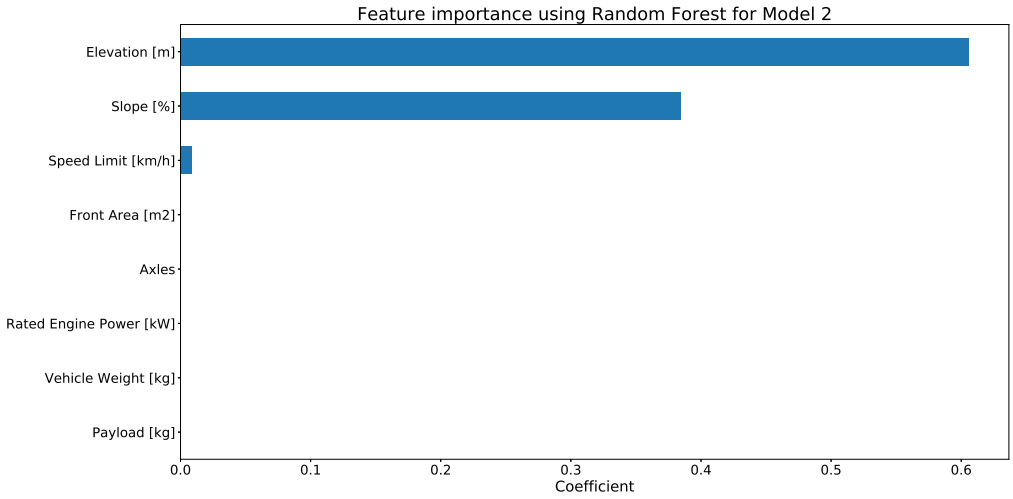


Figure 6.23: Feature Importance using Random Forests for Model 2

Figure 6.23 shows that the road elevation and slope are the most important factors affecting the fuel consumption of the truck. The speed-limit of the road seems to be affecting the fuel consumption the least. Along with that, as also mentioned earlier, it is unlikely that the vehicle attributes - inputs like weight of the vehicle, payload, engine power and the

others - do not affect the fuel consumption of the truck. It is a result of the fact that there was only one vehicle used for the measurement of the data. This voids the effect of any changes caused by the vehicle attributes on the fuel consumption of the vehicle. This could be rectified if a similar study is conducted with data from more than one vehicle.

6.6 Summary

This chapter elaborated on the results obtained in this study for both the models and all the learning techniques. It also dived into the method followed for the tuning of each learning algorithm and the analysis of the results obtained. It was found that all the learning techniques performed similar for Model 1 with Neural Network slightly out-performing the rest. As for Model 2, Neural Network and Random Forests performed similar with Neural Network out-performing in this case as well.
This chapter also established which variables affected the fuel consumption of the truck the most. It was found that the engine torque has the highest positive effect on the fuel consumption (i.e., high fuel consumption for high value of engine torque) while the actual driving gear had the highest negative effect on the fuel consumption (i.e., high fuel consumption for lower driving gears). Factors like vehicle acceleration, the temperature of the fuel and exhaust also seem to have a high correlation to the amount of fuel consumed. Road elevation and vehicle velocity too, have a strong correlation to the amount of fuel consumed.

Chapter 7

COMPARISON WITH A SIMULATION MODEL

This chapter looks further into the application of the developed learning algorithms. The algorithms in this section are checked for their performance with information only about the vehicle and the road. This is of course only possible by applying Model 2 since the other model requires actual measurements from the roads and the response of the driver, which is not available without the vehicle actually driving on the road.
To check for the performance of the learning algorithms, it is not only compared with the measurement data but also with a simulation model used at TNO. This model, the advanced vehicle model, as mentioned before, is a physics based simulation model. The simulation model is also given the same inputs as the learning algorithms to check for their performance. The following figures show the comparison of the performance of the different learning techniques with the simulation tool:

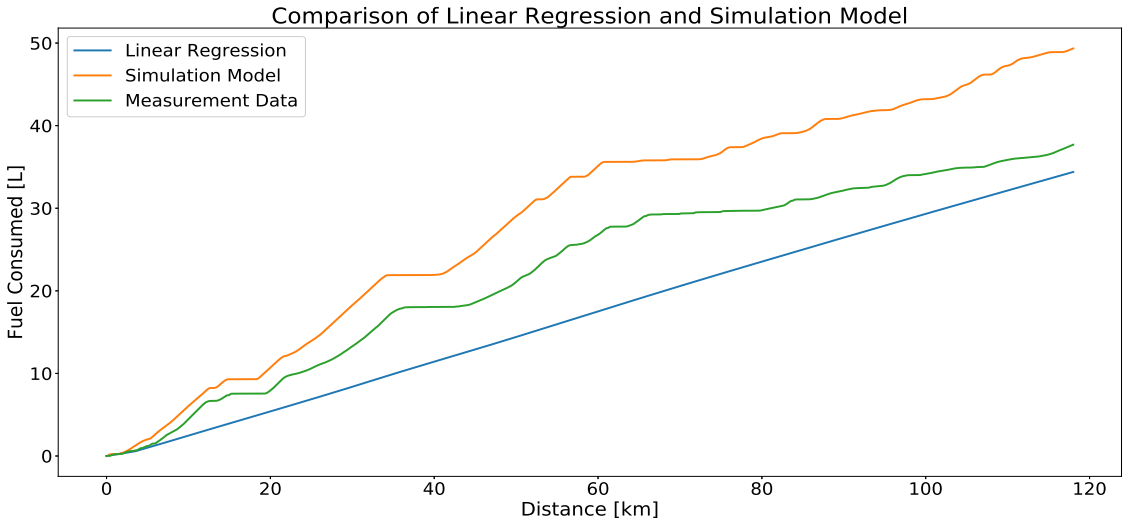


Figure 7.1: Comparison of Simulation Model and Linear Regression

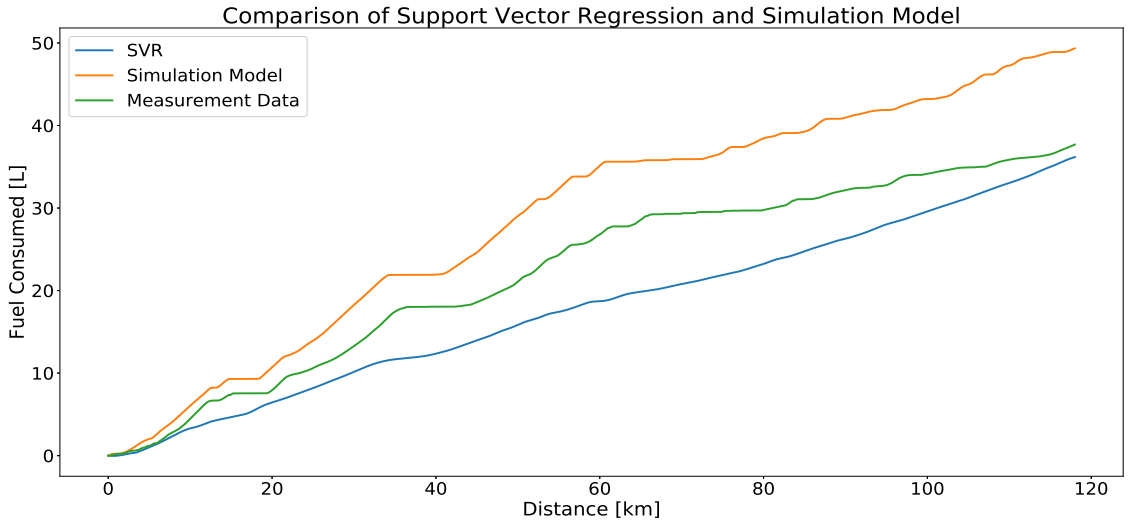


Figure 7.2: Comparison of Simulation Model and Support Vector Regression Algorithm

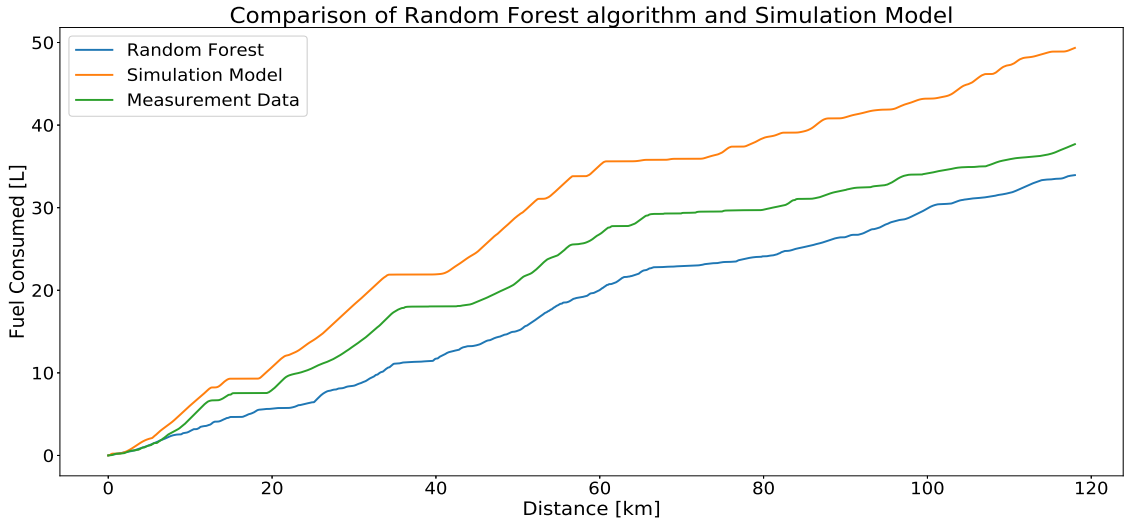


Figure 7.3: Comparison of Simulation Model and Random Forest Algorithm

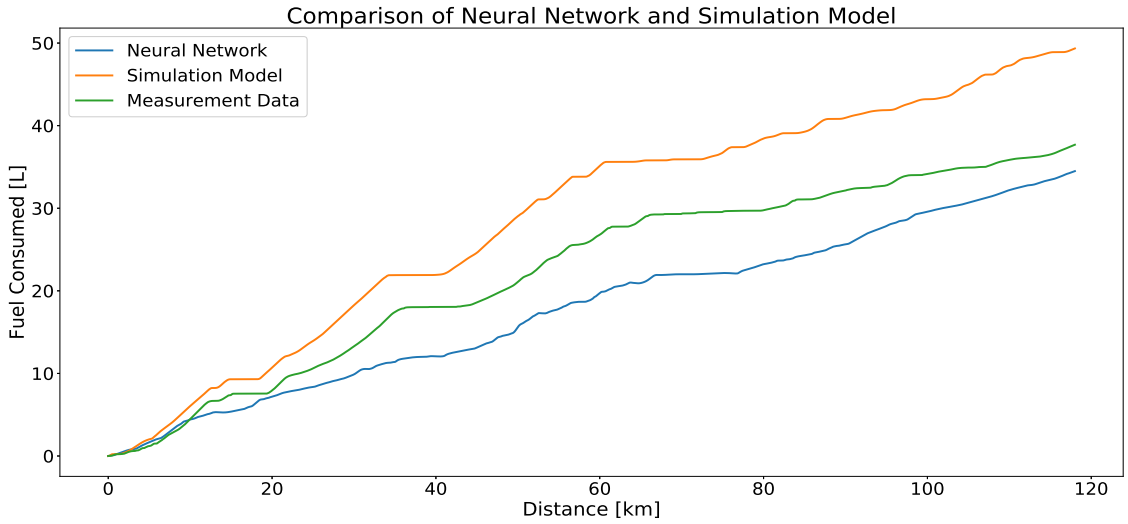


Figure 7.4: Comparison of Simulation Model and Neural Network

All the Figures from 7.1 to 7.4 show the estimation of total fuel consumed by each of the algorithms and their comparison to the simulation tool. Linear Regression and SVR, both give a linear estimation along the entire route. This was also seen in the results for Model 2. Hence, these techniques are not so good to predict the fuel consumption when the driver related inputs are not considered. Random Forest and Neural Network, on the other hand seem to be capable of performing this learning task. They are both capable of predicting the trend in the total fuel consumption along the entire route. They also seem to perform better than the Simulation tool. The table below further elaborates on these results.

Table 7.1: Comparison of the learning algorithms with Simulation model

Model	**Absolute error in fuel consumed [L/km]**	**Fuel Consumed [L/km]**	**Total Fuel Consumed for the trip [L]**
Measurement Data	-	0.319	37.69
Advanced Model (Simulation Model)	0.098	0.417	49.34
Linear Regression	0.028	0.291	34.31
SVR	0.013	0.306	36.17
Random Forest	0.032	0.287	33.94
Neural Network	0.026	0.293	34.57

The results obtained from Linear Regression and Support Vector Regression will be ignored as they are not able to predict the trend of the total fuel consumption. The other two are quite comparable with Neural Network outperforming Random Forests with an absolute error of 0.026 L/km for Neural Network as compared to 0.032 for Random Forest. Both these techniques also fare better than the Simulation tool with the error in the Simulation tool being higher than the machine learning algorithms by a factor of about 3 for Random Forest and about 3.8 for Neural Network. This goes on to say that the algorithms could be used for such a prediction.

7.1 Summary

This chapter compared the results obtained from the learning algorithms to a simulation tool used at TNO for Model 2. It was necessary to compare the results obtained from Model 2 to the ones obtained from the simulation tool as there was no existing literature that could act as a source of comparison. Random Forests and Neural Networks performed comparably while Linear Regression and SVR did not perform well with Neural Networks performing the best. It was also found to perform better than the Simulation Model by a factor of 3.8. But, it also incorporates a few limitations that will be addressed in the following chapters.

Chapter 8

CONCLUSION AND DISCUSSION

After analysing the results obtained from all the learning algorithms for both the models, it is necessary to validate these results. This study compares its results to a similar study conducted on trucks by Henrik [8] for Model 1, where the machine learning algorithms predict the next step fuel consumption of the truck. The following table shows the comparison of the results obtained in his study to the ones obtained in this work:

Table 8.1: Comparison of all learning algorithms to literature for Model 1

Model	**RMSE from this study [L/km]**	**RMSE from Henrik's study at 1 minute sampling time [L/km]**	**RMSE from Henrik's study at 10 minute sampling time [L/km]**
Linear Regression	0.097	0.165	0.08
Support Vector Regression	0.13	0.141	0.076
Random Forest	0.098	0.129	0.074
Neural Network	0.092	0.155	0.077

The Table 8.1 shows that the values obtained are comparable in both the studies with Henrik's RMSE values at a 10 minute (i.e., 0.0016 Hz.) sampling rate to be a little less while the RMSE values at a 1 minute (or 0.016 Hz.) sampling rate were quite high. This difference in the RMSE values of study was due to the fact that the variance in the measurement of fuel consumption is less when the sampling rate is low. It is interesting to note that the results obtained in Henrik's study were done at a data sampling time of either 1 minute (i.e., 0.016 Hz.) or 10 minutes (i.e., 0.0016 Hz.) while in this study it was done at 0.1 seconds (i.e., 10 Hz). As mentioned earlier, the sampling frequency in Henrik's work is quite low and the input variables have the potential to change over a large range of values. This is not favourable. Neural Network performed the best in this study with a RMSE value of 0.092 L/km at a 10 Hz. sampling frequency while Random Forest performed the best in Henrik's study with a RMSE value of 0.074 L/km at 0.0016 Hz. This study goes on to conclude that the machine learning algorithms are also quite potent when the sampling frequency is higher and when there is a high variance in the fuel consumption measurements as otherwise stated by Henrik. It could also be concluded that all the learning algorithms perform similar with Neural Network performing slightly

better than the rest.

As for Model 2, since there were no bench-marks for the performance metrics, the simulation tool was used as the source of comparison. The results obtained from Model 2 showed that learning algorithms like Linear Regression and SVR do not perform so well when the learning task is more complex while Random Forest and Neural Networks fare better. The exclusion of driver affected inputs from the training data-set affects the performance of the algorithms but Neural Networks and Random Forests still manage to provide a decent estimation of the total fuel consumed. These algorithms were also found to out-perform the simulation tool with Random Forests performing better by a factor of 3, while Neural Networks performing better by a factor of almost 3.8. Both these algorithms also could capture the trend of the fuel consumption and not just the total fuel consumed. It is important to note that although the learning algorithms provide better results than the simulation tool, thee are certain limitations in the algorithms which will be elaborated in the next chapter.
The research questions targeted in this work can be answered as follows:

1. *Is machine learning a possible option to predict fuel consumption of heavy duty vehicles?*

   (a) *If yes, which learning technique is the most accurate one?*

   (b) *Which variables in the measurement data-set affect the fuel consumption of the vehicle the most?*

   - Machine learning could definitely be used for the next-step online prediction of the fuel consumption of heavy duty vehicles. The results acquired from this study go on to say that machine learning is a viable technique and with the availability of data, would only get more robust. It was also observed that machine learning algorithms are a good solution for prediction even at high sampling frequencies. This study was conducted at 10 Hz. and the results obtained were still comparable to the one obtained by a comparative source of literature [8] which was performed at a sampling frequency of 0.016 Hz. (1 minute) and 0.00166 Hz. (10 minutes).
   - Although all the techniques used in this study perform quite similar, neural network performs slightly better than the rest. With the option to fine-tune neural networks, the algorithm becomes capable of performing complex learning and is evident from the results obtained.
   - From the measurement data-set, engine torque was found to be the variable which affected the fuel consumption the most. Features influenced by the driver, namely velocity, acceleration and the driving gear of the vehicle also affect the fuel consumption of the vehicle. Engine parameters are found to have a high correlation to the amount of fuel consumed by the truck. The next important set of features is the road attributes (road slope, elevation). They are also found to have an effect on the amount of fuel consumed. Weather conditions affect the fuel consumption the least.

2. *Is excluding driver affected input variables from the training data a viable option for the prediction of total fuel consumption? Also, is this approach adaptive to new roads and new truck configurations? Does it compare to the simulation tool developed at TNO?*
   - Excluding the driver affected inputs for the prediction of fuel consumption of the truck is potentially possible as can be observed from the results obtained from Model 2 of this work. The results obtained also compare well to the simulation model and even perform better than the simulation tool. However, a concrete conclusion about

this could be drawn only with access to more data. This study was conducted with only one long haul vehicle on one route. With access to more data of similar vehicles on different routes, this study could be used to predict the fuel consumption of a truck without the need of driving it on a road.

This work concludes with the prediction of fuel consumption of long haul heavy duty vehicles. It analyses two different models for the prediction of fuel consumption by using four different learning algorithms. Model 1 shows that the results obtained from this work at a sampling frequency of 10 Hz. (0.1 seconds) are comparable to the ones reported by other sources at a sampling rate of 0.016 Hz. (1 minute) or 0.0016 Hz. (10 minutes). This goes on to say that the machine learning algorithms are also potent at higher sampling frequencies. The work is a stepping stone towards the use of machine learning for the prediction of fuel consumption and thereby emissions without the need for driving on the road. This work concludes with the limitations which will be elaborated in the following chapter.

Chapter 9

LIMITATIONS, RECOMMENDATIONS FOR FUTURE WORK

9.1 Limitations of this work

As mentioned earlier, this study is derived from data obtained from one long haul vehicle driving on one route (one tractor-semitrailer driving on a motorway). It is necessary that the algorithms be trained with more data from similar vehicles on different routes for the algorithms to provide results that are more robust. The results obtained in that case would be more reliable as the algorithms would be subjected to not just one route but also different routes and different vehicles.
Another important factor to consider is the driver. This study also considers only one driver driving the vehicle. With data from more than one driver, the algorithms would be able to adapt to not just the driving style of one driver but also incorporate different driving styles. This would also help make the Model 2 of this work more reliable. The Model 2 developed in this study was an attempt to establish the fact that a prediction of estimating the total fuel consumed by a vehicle without the need for it driving on the road is a possibility to be explored. With more labelled data of similar vehicles driven on different roads by different driver would help translate this study into one with robust results.
Although the model does provide good results against the simulation model, it is imperative to understand that the models were trained along the same route. For the comparison to be fair, it is necessary that both the simulation tool and the learning algorithms be tested on newer/ unseen routes. This could not be done due to lack of data about the fuel consumption of the vehicle on different routes which could lead to a possible case of over-fitting.

9.2 Recommendations for Future Work

The author of this work puts forward the following recommendations for possible future work:

1. Acquiring more labelled data of similar long haul vehicles driving would help make the study translate to a larger variety of long haul heavy duty vehicles. The issue with acquiring data for long haul heavy duty trucks is the capital cost needed for such measurements. But, having data for more vehicles would help in making the machine learning models more robust and adaptable to newer variations/ newer innovations in the trucks. It would also help in determining the correlation between the vehicle parameters and fuel consumption.

2. Acquiring data of the trucks driving on different roads would also make this study adapt better to new roads. Driving different trucks on different motorways would help the learning algorithms to get a better understanding of the correlation between the road attributes and the fuel consumption.

3. With labelled data on the driving styles of different drivers, this study could help estimate the amount of fuel consumed by the vehicle for different driving styles. The techniques used in this work could be tuned for different driving styles thereby giving different estimates of fuel consumption for different driving styles, imitating the real world where the kind of driver strongly affects the amount of fuel consumed. The studies conducted by Chris Bingham et al. [20] and João C Ferreira wt al. [21] could be used to characterise different driving techniques, provided one has access to labelled data for different drivers driving on different routes.

4. This study is a starting step towards the prediction of emissions produced by the vehicle. Many studies including the ones conducted by the European Automobile Manufacturers' Association [36] and G Mellios et al. [46] relate the amount of fuel consumed to the amount of CO_2 produced. Since there was no data on the amount of CO_2 emitted by the vehicle, this study did not analyse the amount of CO_2 produced. This study could also translate into the prediction of other gases emitted by the vehicle with the amount of fuel consumed. This would help in not just the prediction of the amount of fuel consumed during a trip by a particular heavy duty vehicle but also the amount of emissions it produces during that trip.

5. The need for cleaner vehicles has lead to the development of hybrid vehicles. This also holds true for the long haul heavy duty vehicles. With data about hybrid vehicles driving on different roads, this study could also be translated to the hybrid vehicles.

6. Another major scope of advancement to this study can be made with data about traffic. Traffic is a variable which affects the fuel consumption of the vehicle. The primary source of estimating traffic in this study was the velocity of the vehicle. Velocity would reflect the traffic conditions on the motorway but it does not help in correlating the effects of traffic on fuel consumption. Adding traffic as a variable in the training data would help in estimating the fuel consumption better.

Bibliography

[1] Paris Agreement. United nations. *United Nations Treaty Collect*, pages 1–27, 2015.

[2] Ben Kraaijenhagen, Cor van der Zweep, Andreas Lischke, Julius Engasser, Per Elofsson, Magnus Ölback, Alessio Sarcoli, Alex Freixas, Marta Tobar, and GertjanKoornneef. Aerodynamic and flexible trucks for next generaion of long distance road transport (aeroflex). Internal Report, 2017.

[3] Aeroflex. `https://aeroflex-project.eu/`. Accessed: 03-10-2018.

[4] Tony Sandberg. *Heavy truck modeling for fuel consumption simulations and measurements*. Master's Thesis, Linkoping University, 2001.

[5] Kanit Wattanavichien, Phan Minh Duc, Chatchai Hongsa-Utain, and Therdsak Chaisuriyaphun. Computer simulation of light duty truck's performance and fuel consumption under steady driving conditions. *SAE Technical Paper 981089*, 1998.

[6] CM Silva, TL Farias, H Christopher Frey, and Nagui M Rouphail. Evaluation of numerical models for simulation of real-world hot-stabilized fuel consumption and emissions of gasoline light-duty vehicles. *Transportation Research Part D: Transport and Environment*, 11(5):377–385, 2006.

[7] Kyoungho Ahn, Hesham Rakha, Antonio Trani, and Michel Van Aerde. Estimating vehicle fuel consumption and emissions based on instantaneous speed and acceleration levels. *Journal of transportation engineering*, 128(2):182–190, 2002.

[8] Henrik Almér. Machine learning and statistical analysis in fuel consumption prediction for heavy vehicles. Master's thesis, KTH Royal Institute of Technology, 2015.

[9] Tin Kam Ho. Random decision forests. In *Proceedings of 3rd international conference on document analysis and recognition*, volume 1, pages 278–282. IEEE, 1995.

[10] Corinna Cortes and Vladimir Vapnik. Support-vector networks. *Machine learning*, 20(3):273–297, 1995.

[11] Zheyuan Cheng, Mo-Yuen Chow, Daebong Jung, and Jinyong Jeon. A big data based deep learning approach for vehicle speed prediction. In *2017 IEEE 26th International Symposium on Industrial Electronics (ISIE)*, pages 389–394. IEEE, 2017.

[12] J-SR Jang. Anfis: adaptive-network-based fuzzy inference system. *IEEE transactions on systems, man, and cybernetics*, 23(3):665–685, 1993.

[13] Abril Galang. Predicting hybrid vehicle fuel economy and emissions with neural network models trained with real world data. Master's thesis, Colorado State University, 2017.

[14] Bratislav Predić, Miloš Madić, Miloš Roganović, Marko Kovačević, and Dragan Stojanović. Prediction of passenger car fuel consumption using artificial neural network: a case study in the city of niš. *Facta Universitatis, Series: Automatic Control and Robotics*, 1(2):105–116, 2016.

[15] Federico Perrotta, Tony Parry, and Luis C Neves. Application of machine learning for fuel consumption modelling of trucks. In *2017 IEEE International Conference on Big Data (Big Data)*, pages 3810–3815. IEEE, 2017.

[16] Sandareka Wickramanayake and HMN Dilum Bandara. Fuel consumption prediction of fleet vehicles using machine learning: A comparative study. In *2016 Moratuwa Engineering Research Conference (MERCon)*, pages 90–95. IEEE, 2016.

[17] Lev Ertuna. Prediction of vehicle fuel consumption using feed-forward artificial neural network model. Working Paper, 2016.

[18] Thomas Robert Waters. Adaptive driver modeling using machine learning algorithms for the energy optimal planning of velocity trajectories for electric vehicles and realizing simultaneous lane keeping and adaptive speed regulation on accessible mobile robot testbeds. Master's thesis, Georgia Institute of Technology, 2018.

[19] Na Lin, Changfu Zong, Masayoshi Tomizuka, Pan Song, Zexing Zhang, and Gang Li. An overview on study of identification of driver behavior characteristics for automotive control. *Mathematical Problems in Engineering*, 2014, 2014.

[20] Chris Bingham, Chris Walsh, and Steve Carroll. Impact of driving characteristics on electric vehicle energy consumption and range. *IET Intelligent Transport Systems*, 6(1):29–35, 2012.

[21] João C Ferreira, José de Almeida, and Alberto Rodrigues da Silva. The impact of driving styles on fuel consumption: A data-warehouse-and-data-mining-based discovery process. *IEEE Transactions on Intelligent Transportation Systems*, 16(5):2653–2662, 2015.

[22] Bo Gu and Giorgio Rizzoni. An adaptive algorithm for hybrid electric vehicle energy management based on driving pattern recognition. In *ASME 2006 International Mechanical Engineering Congress and Exposition*, pages 249–258. American Society of Mechanical Engineers, 2006.

[23] Jianqiang Wang, Lei Zhang, Dezhao Zhang, and Keqiang Li. An adaptive longitudinal driving assistance system based on driver characteristics. *IEEE Transactions on Intelligent Transportation Systems*, 14(1):1–12, 2013.

[24] Glenn D Schilling. *Modeling aircraft fuel consumption with a neural network.* PhD thesis, Virginia Tech, 1997.

[25] Antonio Trani, F Wing-Ho, Glen Schilling, Hojong Baik, and Anand Seshadri. A neural network model to estimate aircraft fuel consumption. In *AIAA 4th Aviation Technology, Integration and Operations (ATIO) Forum*, page 6401, 2004.

[26] European Commission. *Proposal for a Regulation of the European Parliament and the Council setting CO2 emission performance standards for new heavy-duty vehicles.* Brussels, 2018.

[27] Ramadoni Syahputra. Application of neuro-fuzzy method for prediction of vehicle fuel consumption. *Journal of Theoretical & Applied Information Technology*, 86(1), 2016.

[28] Wawrzyniec Gołebiewski and Tomasz Stoeck. Prediction of the mileage fuel consumption of passenger car in the urban driving cycle. *TEKA Commision of Motorization and Energetics in Agriculture*, 14(1), 2014.

[29] Ahmet Gürcan Çapraz, Pınar Özel, Mehmet Şevkli, and Ömer Faruk Beyca. Fuel consumption models applied to automobiles using real-time data: A comparison of statistical models. *Procedia Computer Science*, 83:774–781, 2016.

[30] Michael Ben-Chaim, Efraim Shmerling, and Alon Kuperman. Analytic modeling of vehicle fuel consumption. *Energies*, 6(1):117–127, 2013.

[31] Heidelberg Institute for Geoinformation Technology. Openrouteservice. `https://openrouteservice.org`. Accessed: 2018-09-10.

[32] S Coast. Openstreetmap. `https://openrouteservice.org`. Accessed: 2018-09-10.

[33] Jonas Lindberg. Fuel consumption prediction for heavy vehicles using machine learning on log data. Master's thesis, KTH, School of Computer Science and Communications (CSC), 2014.

[34] European Road Transport Research Advisory Council. *European Roadmap Electrification of Road Transport.* Belgium, 2017.

[35] European Environment Agency. Greenhouse gas emissions from transport. `https://www.eea.europa.eu/data-and-maps/indicators/transport-emissions-of-greenhouse-gases/transport-emissions-of-greenhouse-gases-11`. Accessed: 2019-03-12.

[36] Acea - european automobile manufacturers' association. `https://www.acea.be/`. Accessed: 2019-02-20.

[37] Hilde Huismans. Electric trucks: wishful thinking or the real deal? Master's thesis, Delft University of Technology, 2018.

[38] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

[39] Kevin Gurney. *An introduction to neural networks.* CRC press, 2014.

[40] François Chollet et al. Keras. `https://github.com/fchollet/keras`, 2015.

[41] Martín Abadi, Ashish Agarwal, Paul Barham, Eugene Brevdo, Zhifeng Chen, Craig Citro, Greg S. Corrado, Andy Davis, Jeffrey Dean, Matthieu Devin, Sanjay Ghemawat, Ian Goodfellow, Andrew Harp, Geoffrey Irving, Michael Isard, Yangqing Jia, Rafal Jozefowicz, Lukasz Kaiser, Manjunath Kudlur, Josh Levenberg, Dandelion Mané, Rajat Monga, Sherry Moore, Derek Murray, Chris Olah, Mike Schuster, Jonathon Shlens, Benoit Steiner, Ilya Sutskever, Kunal Talwar, Paul Tucker, Vincent Vanhoucke, Vijay Vasudevan, Fernanda Viégas, Oriol Vinyals, Pete Warden, Martin Wattenberg, Martin Wicke, Yuan Yu, and Xiaoqiang Zheng. TensorFlow: Large-scale machine learning on heterogeneous systems, 2015. Software available from tensorflow.org.

[42] Guozhong An. The effects of adding noise during backpropagation training on a generalization performance. *Neural computation*, 8(3):643–674, 1996.

[43] Chuan Wang and Jose C Principe. Training neural networks with additive noise in the desired signal. *IEEE Transactions on Neural Networks*, 10(6):1511–1517, 1999.

[44] American Association of State Highway and Transportation Officials. A policy on geometric design of highways and streets, 2011.

[45] Guangchuan Yang, Hao Xu, Zhongren Wang, and Zong Tian. Truck acceleration behavior study and acceleration lane length recommendations for metered on-ramps. *International journal of transportation science and technology*, 5(2):93–102, 2016.

[46] G Mellios, Stefan Hausberger, Mario Keller, C Samaras, Leonidas Ntziachristos, P Dilara, and G Fontaras. Parameterisation of fuel consumption and co2 emissions of passenger cars and light commercial vehicles for modelling purposes. *Publications Office of the European Union, EUR*, 24927, 2011.

[47] Walid Gani, Hassen Taleb, and Mohamed Limam. Support vector regression based residual control charts. *Journal of Applied Statistics*, 37(2):309–324, 2010.

Appendix A

Interpolation for OpenStreetMap

The elevation profile w.r.t the distance as extracted from the OpenStreetMap does not have constant increments of distance. An example of this is as follows:

Table A.1: Distance, Elevation values obtained from OpenStreetMap

Distance (km)	Elevation (m)
0	149
0.129	149
0.293	147.7
0.452	144.1
0.464	143.8
0.471	143.5
0.476	143
0.488	142.8

As seen from Table A.1, the distance does not increment in constant steps. This creates certain issues when calculating slope as some data points might be too far away from its preceding point leading to errors in slope calculation. Also it creates issues as there is no constant sampling rate. This makes it difficult to incorporate OpenStreetMap data into measurement data. It would also not be useful to predict the fuel consumption from Model 2 as discussed in Section 3.2.2 as Model 2 was trained at a sampling interval of 2 meters. This brings the need to interpolate the data from the OpenStreetMap to match the training data sampling rate of 2 meters per datapoint. For the sake of illustration, the data obtained from OpenStreetMap has been interpolated for 50 meters of sampling rate (since the elevation does not change by a lot every 2 meters). This is shown in the table below:

Table A.2: Interpolated values of elevation for constant increments of distance

Distance (km)	Elevation (m)
0	149
0.05	149
0.1	149
0.15	148.835
0.2	148.437
0.25	148.04
0.3	147.535
0.35	146.405

Continued on next page

Distance (km)	**Elevation (m)**
0.4	145.275
0.45	144.145
0.5	142.683

As seen from Table A.2, the distance is now incremented constantly in steps of 50 meters. This helps to predict the slope of the route correctly and not like in the previous case where a unrealistic slope could be a possibility. It would also help in the route profiling of new routes.

Values shown in the tables above come from real data obtained from OpenStreetMap

Appendix B

Support Vector Regression

The derivation for the equations of SVR can be done in the following way [47]:
Given a training data-set $(x_1, y_1), (x_2, y_2) \dots (x_n, y_n)$, with X as the input space. A linear function $f(x)$ can be defined as the Equation 3.12:

$$f(x) = \langle \omega, x \rangle + b \tag{B.1}$$

Where, $\omega \in$ X and b $\in \mathbb{R}$.
The $\langle \omega, x \rangle$ denotes the dot product in the input space X, ω is the weight vector and b is the bias.
The objective is to find the minimum value of ω, so the optimisation problem can be written as:

$$\begin{aligned} & & min \quad & \frac{1}{2} \parallel \omega \parallel^2 \\ & Subject\ to: & & y_i - \langle \omega, x_i \rangle - b \leq \epsilon, \\ & & & \langle \omega, x_i \rangle + b - y_i \leq \epsilon \end{aligned} \tag{B.2}$$

Equation B.2 assumes that the all the pairs of (x_i, y_i) are approximated by the function f, which may not be the case. This is the reason why the slack variables ξ and ξ^* are introduced in the optimisation problem. The problem can then be formulated as Equation 3.13:

$$\begin{aligned} min \quad & \left(\frac{1}{2} \|w\|^2 + C \sum_{i=1}^{M} (\xi_i + \xi_i^*) \right) \\ such\ that\ & y_i - \langle w^T, x^i \rangle - b \leq \epsilon + \xi_i^* \\ & \langle w^T, x^i \rangle + b - y^i \leq \epsilon + \xi_i^* \\ & \xi_i, \xi_i^* \geq 0, \forall i = 1 \dots M \end{aligned} \tag{B.3}$$

The optimisation problem can now be solved with the help of the Lagrange Method:

$$\begin{aligned} L = & \frac{1}{2} \parallel \omega \parallel^2 + C \sum_{i=1}^{M} (\xi_i + \xi_i^*) - C \sum_{i=1}^{M} (\eta_i \xi_i + \eta_i^* \xi_i^*) \\ & - \sum_{i=1}^{M} \alpha_i (\epsilon + \eta_i - y_i + \langle \omega, x_i \rangle + b) \\ & - \sum_{i=1}^{M} \alpha_i^* (\epsilon + \eta_i^* + y_i - \langle \omega, x_i \rangle - b) \end{aligned} \tag{B.4}$$

Where, L is the Lagrangian and η, η^*, ξ and ξ^* are Lagrangian multipliers. The partial derivatives with respect to b, ω and ξ^* would look as follows:

$$\begin{aligned} \frac{dL}{db} &= \sum_{i=1}^{M}(\alpha_i^* - \alpha_i) = 0 \\ \frac{dL}{d\omega} &= \omega - \sum_{i=1}^{M}(\alpha_i^* - \alpha_i)x_i = 0 \\ \frac{dL}{d\xi^*} &= (C - \alpha_i^* - \eta_i^*) = 0 \end{aligned} \tag{B.5}$$

Substituting Equation B.5 into B.4 would result in the following optimisation problem:

$$\begin{aligned} min \quad & \frac{1}{2}\sum_{i,j=1}^{M}(\alpha_i - \alpha_i^*)(\alpha_j - \alpha_j^*)\langle x_i, x_j\rangle + \epsilon\sum_{i=1}^{M}(\alpha_i - \alpha_i^*) - \sum_{i=1}^{M} y_i(\alpha_i - \alpha_i^*) \\ \textit{Subject to:} \quad & \sum_{i=1}^{M} y_i(\alpha_i - \alpha_i^*) = 0 \\ & \alpha_i, \alpha_i^* \in [0, C] \end{aligned} \tag{B.6}$$

The Equation B.6 can finally be written as:

$$\begin{aligned} \omega &= \sum_{i=1}^{M}(\alpha_i - \alpha_i^*)x_i \\ f(x) &= \sum_{i=1}^{M}(\alpha_i - \alpha_i^*)\langle x_i, x\rangle + b \end{aligned} \tag{B.7}$$

The Equation B.7 is the same as the solution to the optimisation problem as mentioned earlier in Section 3.3.2.

Appendix C

Butterworth Low-Pass Filter

A low-pass butterworth filter is one which filters out the higher frequencies in the signal and lets the lower frequencies pass. The range of lower frequencies that are passed is called the bandwidth of the filter. The maximum frequency that the filter transmits is called the cutoff frequency.
A normalized butterworth filter with cutoff frequency of 1 rad/sec can be defined as:

$$|H(j\omega)|^2 = \frac{1}{1+\omega^{2n}} \tag{C.1}$$

Where,
$H(j\omega)$ is the Frequency Response (gain) of the filter
ω is the frequency, and;
n is the order of the filter A simple example of the working of the butterworth filter is as follows:

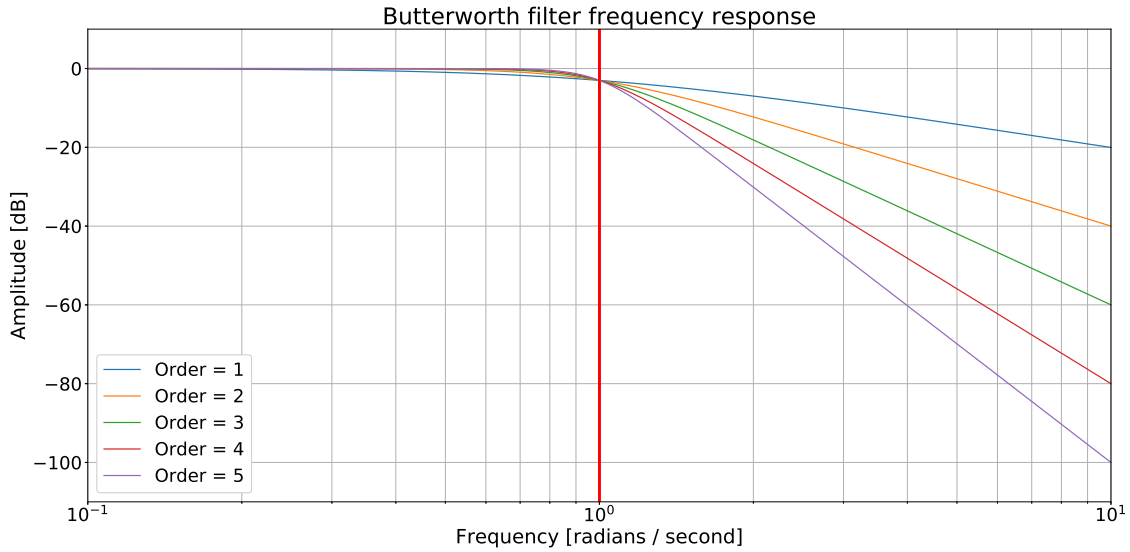


Figure C.1: An illustration of the Butterworth Low-Pass Filter

In the Figure C.1, the range of frequency below 1 Hz. is called the bandwidth of this filter.With a cutoff frequency of 1 Hz., the filter does not let frequencies above 100 Hz. pass through it. The order of the filter is responsible for the steepness of the curve after the cutoff-frequency. The higher the order, the steeper is the slope of the frequency response. This is useful as noise in the signal is high frequency. If the high frequencies are somehow filtered, one can get rid of the noise. This is the principle used in this study to filter out the noise in the signal.