DEGREE PROJECT, IN COMPUTER SCIENCE , SECOND LEVEL
*STOCKHOLM*, *SWEDEN 2015*

# Machine learning and statistical analysis in fuel consumption prediction for heavy vehicles

HENRIK ALMÉR

**KTH ROYAL INSTITUTE OF TECHNOLOGY**

**SCHOOL OF COMPUTER SCIENCE AND COMMUNICATION (CSC)**


KTH Computer Science
and Communication

# Machine learning and statistical analysis in fuel consumption prediction for heavy vehicles

HENRIK ALMÉR

Master's Thesis at CSC
Supervisor: Pawel Herman
Examiner: Anders Lansner

TRITA xxx yyyy-nn

# Abstract

I investigate how to use machine learning to predict fuel consumption in heavy vehicles. I examine data from several different sources describing road, vehicle, driver and weather characteristics and I find a regression to a fuel consumption measured in liters per distance. The thesis is done for Scania and uses data sources available to Scania.

I evaluate which machine learning methods are most successful, how data collection frequency affects the prediction and which features are most influential for fuel consumption.

I find that a lower collection frequency of 10 minutes is preferable to a higher collection frequency of 1 minute. I also find that the evaluated models are comparable in their performance and that the most important features for fuel consumption are related to the road slope, vehicle speed and vehicle weight.

# Referat

## Maskininlärning och statistisk analys för prediktion av bränsleförbrukning i tunga fordon

Jag undersöker hur maskininlärning kan användas för att förutsäga bränsleförbrukning i tunga fordon. Jag undersöker data från flera olika källor som beskriver väg-, fordons-, förar- och väderkaraktäristiker. Det insamlade datat används för att hitta en regression till en bränsleförbrukning mätt i liter per sträcka. Studien utförs på uppdrag av Scania och jag använder mig av datakällor som är tillgängliga för Scania.

Jag utvärderar vilka maskininlärningsmetoder som är bäst lämpade för problemet, hur insamlingsfrekvensen påverkar resultatet av förutsägelsen samt vilka attribut i datat som är mest inflytelserika för bränsleförbrukning.

Jag finner att en lägre insamlingsfrekvens av 10 minuter är att föredra framför en högre frekvens av 1 minut. Jag finner även att de utvärderade modellerna ger likvärdiga resultat samt att de viktigaste attributen har att göra med vägens lutning, fordonets hastighet och fordonets vikt.

# Contents

# Chapter 1

# Introduction

This study evaluates methods of machine learning (ML) and statistical analysis for predicting fuel consumption in heavy vehicles. The idea is to use historical data describing driving situations to predict a fuel consumption in liters per distance.

The general problem description is to examine a large number of attributes describing a fuel consumption situation and to employ ML methods to find a regression from such attributes to a fuel consumption. Attributes included could be environmental conditions, vehicle configuration, driver behavior and weather conditions. Research has been made into how to do such predictions for aircraft [1, 2], engines [3] and passenger cars [4] as well as heavy vehicles [5, 6, 7]. The previous research makes suggestions about which ML methods are most successful in fuel consumption prediction as well as what kind of attributes are most influential in fuel consumption for road vehicles. The specific problem investigated in this study is how to do fuel consumption prediction for Scania's heavy vehicles using the data sources available to Scania.

The study is part of the COMPANION project, which is a collaborative effort including Scania CV AB, Volkswagen Group Research, KTH, Oldenburger Institut für Informatik (OFFIS), IDIADA Automotive Technology, Science & Technology in the Netherlands and the Spanish haulage company Transportes Cerezuela [8]. The goal of the COMPANION project is to develop a real-time coordination system to dynamically create, maintain and dissolve platoons (road trains), according to a decision-making mechanism, taking into account historical and real-time information about the state of the infrastructure (traffic, weather, etc.) [8]. This study fits into the COMPANION project by researching ways to construct a fuel model and to predict fuel consumption using platooning as a factor.

One goal of the study is to evaluate different ML based approaches for regression from a set of descriptive attributes to a fuel consumption in liters per distance. Examples of the attributes considered in the study are the vehicle weight, engine strength, velocity of the vehicle, slope and speed limit of the road as well as weather data such as wind speed and direction. The data is collected from sources including Scania's Fleet Management (FM) system, a GPS routing system, weather observation

data from SMHI and vehicle configuration information. Several different ML methods are trained on this data to find a regression to a fuel consumption.

The available data from Scania's FM system is sent from active vehicles with a frequency that can vary between vehicles. The messages contain information about the vehicle's position, current odometer reading, fuel consumption as well as other descriptive features. The main goal of the study is to investigate how the data collection frequency affects the prediction, and if the current standard sampling rate of 10 minutes is sufficient.

Questions this study intends to answer are:

- How does the data collection frequency affect the quality of prediction, is it feasible to use a 10 minute sampling rate or is a higher frequency required?
- Which of the attributes in the available data are most relevant for fuel prediction?
- Which of the evaluated ML methods are best suited for this problem?

## 1.1 Contribution and expected impact

There is no existing reliable ML based model for predicting fuel consumption in heavy vehicles. A well functioning model for fuel prediction could be an important building block in a route planning system and could be used as a heuristic in finding routes that minimize fuel consumption. This is useful since reduced fuel consumption means reduced environmental impact as well as reduced fuel costs. The model could also be used for anomaly detection and identify vehicles with irregular fuel consumption. This is useful since it enables early identification and correction of possible faults in the vehicle.

The novelty of the approach lies in using collected statistical data for fuel consumption and connecting them not only to vehicle and engine characteristics but also to environmental parameters such as weather and road conditions as well as driver behavior.

Answering the questions posed in the study can give insight into how to best construct a predictive model for fuel consumption that can be used in planning applications or in anomaly detection. The study can also give insight into which parameters are of greatest importance and where Scania should direct their data collection and processing efforts. Investigating how the sampling rate affects the prediction results can give Scania insight into which collection frequency should be used as the default for their vehicles. Making a decision about the default collection frequency will impact Scania's data storage requirements and could potentially incur large costs for Scania. Finding a good trade-off between data resolution and storage requirements is essential to collect usable data while keeping storage costs low.

## 1.2 Scope

The study is limited to evaluating data from Scania's FM system, geographical data available in the DigitalReality 3.0 GPS Routing system, historical weather observations from SMHI and vehicle configuration data from the internal Scania system Product Individual Service (PIS). The collected data is limited to vehicles operated by Scania's Transport Laboratory that have been connected to the FM system and have been in operation between the 1st of June 2013 and 31st of October 2014. The data is also limited to observations within the Swedish borders.

The study does not attempt to answer which sampling rate is the optimal sampling rate for predicting fuel consumption. Rather it focuses on evaluating if Scania's default sampling rate of 10 minutes, which constitutes the majority of position messages in Scania's FM System, is sufficient or if a higher sampling rate is required.

# Chapter 2

# Background

In the following sections the study is put in context and important concepts are described. Section 2.1 describes the platooning concept and the presents the overarching COMPANION project which this study is a part of. Section 2.2 describes the FM system which is the single most important data source for the study and a prerequisite to be able to do statistical prediction of fuel consumption. Subsequent sections present previous work in the field and further motivate the contribution of the study.

## 2.1 Platooning and the COMPANION project

This study is part of the COMPANION project, which is a research project into the creation, coordination, and operation of vehicle platoons, or road trains [8]. Driving in a platoon has been shown to reduce air resistance and it has been proven to lead to reductions in fuel consumption [6, 9, 10].

The goal of the COMPANION project is to develop a real-time coordination system to dynamically create, maintain and dissolve platoons, according to a decision-making mechanism, taking into account historical and real-time information about the state of the infrastructure (traffic, weather, etc.) [8]. This study fits into the COMPANION project by researching ways to construct a fuel model and to predict fuel consumption using platooning as a factor.

COMPANION is a collaborative effort including Scania CV AB, Volkswagen Group Research, KTH, Oldenburger Institut für Informatik (OFFIS), IDIADA Automotive Technology, Science & Technology in the Netherlands and the Spanish haulage company Transportes Cerezuela [8].

## 2.2 Fleet Management

FM is the management of a company's transportation fleet. In the Scania case it is the tracking and management of all Scania trucks for which the customer has signed up for the Scania Fleet Management program. Scania's FM system includes a vehicle

tracking component in which the trucks send messages with their positions, current fuel level and other characteristics with a given frequency. The frequency can be set to any unit of time but most common is a frequency of 10 minutes. For some vehicles the frequency may be as high as 1 minute. The FM system also includes a component for analysis of driver behavior and tracks information about vehicle idling time, time spent in gears not suited for the current speed, frequency of hard breaks, etc. Section 4.1 describes the FM data in more detail.

The data for each vehicle is recorded with an onboard computer located on the truck, which is then sent to a backend system via a telecommunication link. The data is stored in a database system and made available both internally at Scania and externally for Scania's customers.

## 2.3 Related work

### 2.3.1 Simulation based approaches

Much of the previous work in fuel consumption prediction consists of simulation based approaches [11, 12] which perform physical calculations and are often slow to run as they simulate the internal components of the truck. One existing such model is the Scania Truck and Road Simulation (STARS) which is a simulation system that requires vehicle and driver specific configuration of the model to be able to perform prediction [11]. Simulation based approaches have the problem that they take a long time to run and require considerable manual configuration in order to perform prediction. Modifying a simulation based model to take more parameters into account would also increase the prediction complexity and potentially make it significantly slower. Further, a simulation based model can not generalize to become manufacturer independent, since they require vehicle specific configurations.

### 2.3.2 Machine learning based approaches

Scania has done research in the area of ML methods for fuel consumption prediction prior to this thesis and there is a lot of information to build on and learn from. There is also a significant amount of research done in related fields such as fuel consumption prediction for aircraft, passenger cars, and engines using statistical analysis and ML methods. This section will detail reports studied and considered in preparation for this study.

#### Fuel consumption prediction for heavy vehicles

Viswanathan [5], Lindberg [13] and Svärd [7] have done research for Scania on similar subjects using similar approaches as those in this study. However their studies are limited in scope and do not reach a clear conclusion about the usability of an ML model for fuel consumption. Nor do they investigate if it's possible to create an accurate fuel prediction model using observations with the default collection

frequency of 10 minutes. Hence there is still a need for Scania to do further research in the subject. There is also new data to train on which may improve the resulting model.

Viswanathan [5] did research into which features describing driver behavior in Scania's FM database were of greatest importance when predicting fuel consumption for Scania's vehicles. In order to investigate this she implemented a prediction model using random forests and gradient boosting and found that the best results were obtained with the random forest model. She concluded that the parameters speed, coasting, distance with trailer attached, distance with cruise control and maximum speed were the most significant with regards to fuel consumption [5]. She only examined driver behavior features and did not take into account road properties, vehicle properties or weather influence. Her focus was on parameter importance rating and there was only limited work put into building a predictive model. In addition, she did not investigate how to train a predictive model to be used for routing or anomaly detection.

Lindberg [13] attempted to realize a predictive model for fuel consumption using the FM data combined with road, vehicle and weather data. He focused on a small set of training data using observations from a route between Södertälje and Sälen. He trained a regression tree, random forest, boosted tree and support vector regression (SVR) model but made no conclusion about which method gave the best results. He concluded that the vehicle weight and slope of the road were the most influential variables for prediction. Due to the limited amount of data, data with low sampling rate and weather and road data of low quality, the results he published had low accuracy and underestimated fuel consumption by on average 26 % [13].

Even though Lindberg concluded that the altitude difference was the most significant variable for fuel consumption prediction, he made a great simplification when compiling his road data. He looked at two points next to each other in a sequence of observations and assumed that the slope of the road between those two points could be described by the difference in altitude. Since the distance between two points could be quite long, with a mean distance of 13 km [13], there is a possibility that the road segment had more uphill and downhill slopes than the altitudes of its endpoints would suggest. Svärd [7] made an improvement on this measure and divided the road into segments where slope and other road characteristics such as speed limit were constant.

Svärd [7] confirmed that a predictive model for fuel consumption should be possible to realize and he also confirmed Lindberg's conclusions that the vehicle weight and the slope of the road were the most important variables for the predictive model. He did his research using observations from the E4 motorway between Södertälje and Helsingborg collected between June 1st and December 31st 2014 [7]. Svärd's study did not take into account any vehicle characteristics except the weight. There are many other characteristics that could potentially have a large effect on fuel consumption, such as engine power and volume, wheel configuration or the rear axle gear ratio.

Svärd [7] trained several different models on his data set. He trained a linear

regression model, decision tree, artificial neural network (ANN), random forest and SVR. He found that of the different models the SVR model performed best. He discovered that he could improve his bias-variance tradeoff further by combining the different models into a weighted ensemble model. However, his results are questionable since he used different data for training the ensemble model than he had used while training the other models [7]. Svärd also gave insight into how to preprocess the data to calculate vehicle platooning relationships. Whether a vehicle is in a platoon or not was not accounted for in the available data but Svärd [7] showed a method for calculating it by comparing positions, headings and timestamps for messages from different vehicles.

Svärd [7] only examined observations with a 1 minute sampling rate and did not investigate how a lower collection frequency of 10 minutes between observations would influence the results. It is relevant to investigate this question since the majority of observations in Scania's FM system are collected with a 10 minute frequency. There are only a few vehicles in operation that send observations with a frequency of 1 minute and most of these vehicles are operated by the Scania Transport Laboratory. Only data with a 10 minute sampling rate is available for training more generalized models that can predict fuel consumption on other roads than those travelled by the Scania Transport Laboratory. In addition, Svärd did not evaluate the usefulness of his model in any experiment, he only examined the prediction error for his training, test and validation data sets but did not investigate how his model generalized to other data.

### Fuel consumption prediction in other applications

Togun & Baysec [3] were successful in using an ANN to predict torque and specific fuel consumption of a gasoline engine. They used only three input parameters in a set of observations that they compiled from experimental results. With this data they managed to obtain good results and achieve a high prediction accuracy for their test set. The application of ANNs to predict fuel consumption in an engine using only three input parameters is very different to predicting fuel consumptions for a heavy vehicle in a traffic environment. However, the problems are similar in their characteristics and the results of Togun & Baysec [3] motivate investigating how well an ANN may perform in this study.

Schilling [1] trained several neural networks for predicting fuel consumption of aircraft using input parameters describing the aircraft and weather conditions. He showed that ANNs can be equally accurate as models based on physical calculations while having significantly lower computational complexity. Trani et al. [2] continued the research into fuel consumption prediction for aircraft with ANNs and confirmed that they can be as accurate as analytical simulation models while being significantly faster. They concluded that ANNs can represent complex aircraft fuel consumption functions for climb, cruise and descent phases of flight. The findings of Schilling [1] and Trani et al. [2] together with the research of Togun & Baysec [3] suggest that ANNs may be a good fit for modelling fuel consumption.

Wang et al. [4] examined the influence of driving patterns on fuel consumption using a portable emissions measurement system on ten passenger cars. They concluded that vehicle fuel consumption is optimal at speeds between 50 and 70 km/h and that fuel consumption increases significantly during acceleration. These results indicate that both the speed limit of the road and driver behavior have large impact on fuel consumption.

# Chapter 3

# Method

The process for compiling the data set, training the models and evaluating the results can be broken down into a data mining phase and a training phase. The data mining phase is the collection, analysis and consolidation of data from the different sources. The training phase is using the consolidated data and fitting ML models to it. Once the training is finished the results will be evaluated to determine which features are most influential and which model performed best.

## 3.1 Machine learning

ML is a field in artificial intelligence that concerns construction of systems which can learn from examples in different ways. The concept can be described with the following definition by Tom M. Mitchell [14].

> 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 $E$.
>
> — *Tom M. Mitchell*

In the context of this study the experience $E$ is the information about fuel consumptions and the other data collected from the different data sources. The task $T$ is to estimate fuel consumption and the performance measure $P$ is related to the size of the error in the prediction.

## 3.2 Performance metrics for regression methods

To evaluate how well an ML method for regression describes the underlying relationship several metrics may be applied to the trained model. The metrics I intend to use are described below.

### 3.2.1 Bias and variance

In statistical ML applications for regression, the bias of a model is the difference between the estimated value and the true value of the parameter being estimated. This means that bias is a measure of the model's ability to give accurate estimations. High bias is related to underfitting [15].

Variance has to do with the stability of the model in response to new training examples. It can be described as the variation of estimations between different realizations of a model. For example if we have several different training sets describing the same underlying relations, and training a model on one of the sets produces a very different result than training the model on another set, then we have a high variance model [15]. Variance is small if the training set has a minor effect on the model's estimates. Variance does not measure if a model is correct or not, only if it is consistent. High variance is related to overfitting [15].

The total error of a model can be expressed as $Error = Bias + Variance$. The bias-variance tradeoff is the problem of simultaneously minimizing these two properties to achieve a low error [15]. Ideally, one wants to choose a model that both accurately captures the regularities in its training data, but also generalizes well to unseen data. The models ability to generalize can be evaluated by examining these two properties.

### 3.2.2 Mean squared error

The mean squared error (MSE) of a model is the average of the squares of the prediction errors. The error in this case is defined as the difference between the estimate and the true value. The MSE incorporates both the variance of the estimator and its bias and can be expressed by (3.1) [16].

$$MSE(Variance) = VarianceEstimate + Bias(Estimate, TrueValue)^2 \quad (3.1)$$

Thus the MSE assesses the quality of an estimator in terms of its variation and degree of bias. The root mean squared error (RMSE) is simply the square root of the MSE. Using the RMSE as a measure will give the same results as using the MSE, but the RMSE can be considered a more meaningful representation of the error. In this study RMSE will be used to evaluate the different models.

### 3.2.3 Percent error

The percent error is derived from the relative error and can be expressed by (3.2). In this study the relative error is used as a complement to the RMSE to describe the prediction error of the fitted models.

$$\delta = 100 \cdot \left| \frac{TrueValue - Estimate}{TrueValue} \right| \quad (3.2)$$

## 3.3 Linear regression

Linear regression is the fitting of a linear function of one or more inputs to an output. In the univariate case it is the fitting of a straight line with input $x$ and output $y$ on the form $y = w_1x + w_0$, where $w_0$ and $w_1$ are real-valued coefficients to be learned [17]. When finding the weights in a linear regression problem it is most common to minimize the squared loss function. The problem is then to find the weight vector $\mathbf{w}^*$ according to (3.3). Choosing the weights in this way guarantees that we'll find a unique global minimum [17].

$$\begin{aligned} \mathbf{w}^* &= \underset{w}{\operatorname{argmin}}\, Loss(h_w) \\ &\text{where} \\ h_w(x) &= w_1x + w_0 \end{aligned} \tag{3.3}$$

The multivariate case of linear regression is not much more complex than the univariate case. In multivariate linear regression each example $x_j$ is an $n$-element vector and the object is to find the hyperplane which best fits the outputs $y$ according to some loss function, most commonly the squared loss function. The hypothesis space is given by the set of functions of the form defined by (3.4) [17]. The vector of weights that minimizes the loss function is given by (3.5) [17].

$$h_{sw}(\mathbf{x_j}) = \mathbf{w}^\intercal \mathbf{x_j} = \sum_i w_i x_{j,i}. \tag{3.4}$$

$$\mathbf{w}^* = \underset{w}{\operatorname{argmin}} \sum_j Loss(y_j, h_{sw}) \tag{3.5}$$

Using either gradient descent or analytical solving we can reach the unique minimum and fit the weights to the outputs [17]. In this study linear regression will be used as a benchmark model to compare the more advanced regression models to. The loss function used is the least squares loss function.

### 3.3.1 Cook's distance

A common metric used to evaluate the influence of a single data point on a linear regression model is Cook's distance. Cook's distance, or Cook's D, is used to estimate the influence of a data point when performing least squares regression. The mathematical definition is given by (3.6) where $\hat{Y}_j$ is the prediction from the full regression model for observation $j$, $\hat{Y}_{j(i)}$ is the prediction for observation $j$ from a regression model trained on data where observation $i$ has been omitted, $p$ is the number of fitted parameters in the model and MSE is the mean squared error of the model [18].

$$D_i = \frac{\sum_{j=1}^{n} (\hat{Y}_j - \hat{Y}_{j(i)})^2}{p \text{ MSE}}, \tag{3.6}$$

In this study Cook's distance will be used to analyse the results of the linear regression fits.

## 3.4 Regression trees

Decision trees are a simple method of supervised learning in which the final model takes a vector of attributes as inputs and returns a single value, or decision, as output. In a decision tree, leaf nodes represent the decisions and branches represent conjunctions of attributes that lead to those decisions [17].

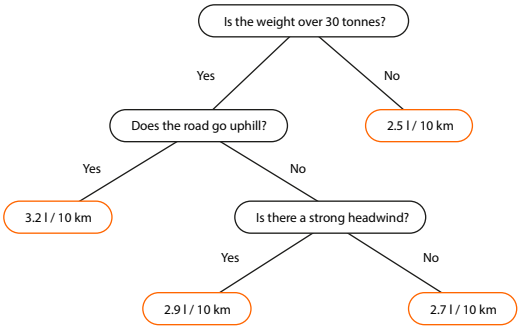


**Figure 3.1.** An example of a simple regression tree. This tree has 3 features that it evaluates. At each node a yes or no question is answered and depending on the answer the path down the tree is decided. Once a leaf node is reached the tree returns a response.

Regression trees are decision trees used for regression, that is their target variable can take continuous values. A regression tree is a tree of nodes where each leaf node has a linear function of some subset of numerical attributes, rather than a single value which is the case for classification trees [17]. For example a regression tree for fuel prediction may have leaf nodes that contain linear functions of vehicle weight, road slope and engine strength. The learning algorithm must decide when to stop splitting and start to apply linear regression over the attributes [17]. An example of a regression tree is illustrated in Figure 3.1.

The order in which to place the nodes and which node to choose as the root is decided by examining the entropy and information gain of the attributes. Information gain is the expected reduction of entropy achieved after eliminating an attribute from the equation [17].

## 3.5 Random forests

A random forest for regression is an ensemble learning method where several regression trees are trained and which outputs the mean prediction of the individual trees. Random forests use a modified tree learning algorithm that selects a random

subset of the attributes at each candidate split in the learning process. Random forests correct for the tendency of decision trees to overfit to training data [19].

Random forests uses two parameters for tuning a model fit. They are `mtry` and `ntrees`. `mtry` defines how many features to use in each tree and `ntrees` how many trees to train in total. The default `mtry` is usually set to the square root of the total number of features and `ntrees` is usually selected to be as high as possible while keeping training time reasonably short. In order to find optimal parameter settings I iterate values of `mtry`, fit a model to the data and evaluate the RMSE of the fitted model. I select a parameter value that keeps both the RMSE and model training time low.

## 3.6 Artificial neural networks

ANNs were first envisioned as a digital model of a brain, connecting many simple neurons into a network capable of solving complex problems. A neuron in ANN terms is a node in the neural network. Roughly speaking one can say that it "fires" when a linear combination of its inputs exceed some threshold [17]. The nodes have one or more inputs, each of which has an associated weight. The inputs multiplied by their corresponding weights are summed in the node and the sum is fed to an activation function which returns a binary response signalling if the sum exceeded the threshold or not [17]. Figure 3.2 illustrates how an ANN node is constructed.

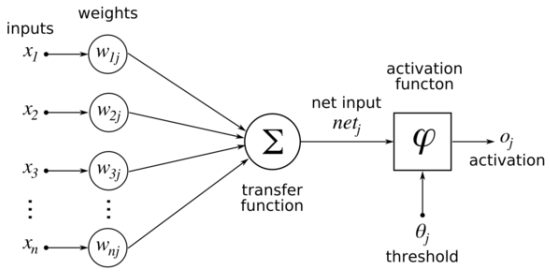


**Figure 3.2.** Illustration of an artificial neuron. Each neuron has a set of input links with associated weights, a transfer function and an activation function. The activation function outputs a binary response and the result is sent on the output link.

The activation function used in the node is most often either a hard threshold function, in which case the node is called a perceptron, or a logistic function [17]. The two types of activation functions are illustrated in Figure 3.3. In this study activation functions of the logistic kind will be used since they are continuous and thus possible to differentiate, which is a requirement for being able to update the weights in the training algorithm that I will use [17].

Note that the activation functions of the nodes are nonlinear, meaning that their output is not the sum of their inputs multiplied by some constant. This property of the individual nodes ensures that the entire network of nodes also can represent

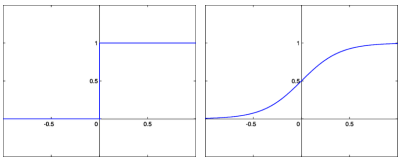

**Figure 3.3.** Activation functions commonly used in ANN nodes. The left image shows the hard threshold function associated with a perceptron. The right image shows a logistic function.

nonlinear functions [17].

To form a network, the nodes of an ANN are arranged in layers and connected by directed links where each link has an associated weight. The layers in between the input and output layers are referred to as hidden layers [17]. Figure 3.4 illustrates how an ANN may be constructed.

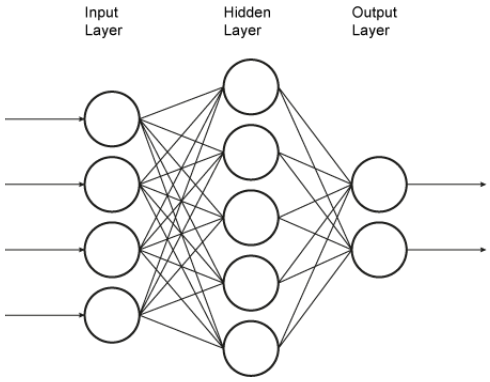


**Figure 3.4.** An example of an ANN with four inputs and a hidden layer. Each node is connected to all nodes in the succeeding layer by a directed link. Each of the links have an associated weight and it's the values of these weight that, together with the activation functions, define the behavior of the ANN.

ANNs can be constructed in two ways. The first possibility is the feed-forward network in which the connections are only in one direction. That is every node receives input from nodes in the previous layer and sends their output to the next layer, without any loops. Such an ANN where each layer is fully connected to the next one and where the nodes use logistic activation functions is called a multilayer perceptron (MLP) [20]. The other option is a recurrent network where nodes may feed their responses back to nodes in preceding layers. Recurrent networks form dynamic stateful systems that may exhibit oscillations and chaotic behaviour and can be difficult to understand [17]. This study will focus on feed-forward networks.

In an MLP you can back-propagate the error from the output layer to the hidden layers. Such backpropagation of error in a multilayer network implements gradient descent to update the weights in the network and minimize the output error [17]. The backpropagation learning algorithm used in this study is resilient backpropagation with weight tracking as defined by Riedmiller [21].

**Table 3.1.** ANN parameters.

| Variable name | Description | Default value |
|---|---|---|
| `hidden` | The number of nodes in the hidden layer | 2 |
| `rep` | The number of times training is repeated | 1 |
| `threshold` | A threshold used as stopping critera for convergence | 0.01 |
| `learningrate.factor` | a list containing the multiplication factors for the upper and lower learning rate | [0.5, 1.2] |

The model fit is dependent on several parameters defined by the `neuralnet` R package. They are described in table 3.1 [22]. During training all parameters except `hidden` are set to their default values and a search is performed to investigate which value of `hidden` yields the best results. The network is configured to have sigmoid activation function in the hidden layer and a linear output function.

## 3.7 Support vector regression

Support vector machines (SVM) is a very popular method for supervised learning and it is a good first try for problems where you do not have any specialized prior knowledge of the problem domain [17]. In its original formulation SVMs do classification of data points by a maximum margin decision boundary. For example an SVM might find the line between two clusters of data points that give the largest margin to the clusters [17]. To find such a decision boundary the SVM finds so called support vectors, which are the data points that lie on or inside the margin. Using the so called kernel trick and dual formulation of the SVM optimization problem different kernels may be used to embed the input data in a higher dimensional space, producing non-linear classifiers and greatly expanding the hypothesis space [17].

SVR is an extension of the SVM where the same principles are applied to do regression instead of classification. Instead of finding a decision boundary with maximum margin the SVR finds a function approximation that minimizes the error. Like SVMs, SVR optimizes the generalization properties of the model. They rely on defining a loss function that ignores errors which are situated within a certain distance of the true value, this distance is denoted by the variable $\epsilon$. Thus they depend only on a subset of the training data and ignore any training data close to the model prediction [23]. An example of an SVR decision boundary is illustrated in Figure 3.5.

In this study I will train an $\epsilon$-SVR model on the data and evaluate its performance. The optimization problem solved by the $\epsilon$-regression has the following def-

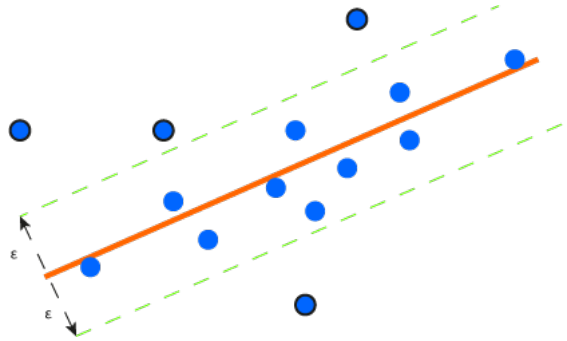


**Figure 3.5.** A 2D example of SVR estimation. The orange line represents the function approximation and the green lines the $\epsilon$-boundaries. The highlighted data points are the support vectors, in the regression case they are the data points outside of the $\epsilon$-boundaries.

inition [24]; Consider a set of training points, $\{(\mathbf{x_1}, z_1), \ldots, (\mathbf{x_l}, z_l)\}$, where $\mathbf{x_i} \in R^n$ is a feature vector and $z_i \in R^1$ is the target output. Under given parameters $C > 0$ and $\epsilon > 0$, the dual form of support vector regression is given by (3.7) [25].

$$\begin{aligned} \min_{\boldsymbol{\alpha},\boldsymbol{\alpha^*}} \frac{1}{2}(\boldsymbol{\alpha} - \boldsymbol{\alpha^*})^\intercal Q(\boldsymbol{\alpha} - \boldsymbol{\alpha^*}) + \epsilon \sum_{i=1}^{l}(\alpha_i + \alpha_i^*) + \epsilon \sum_{i=1}^{l} z_i(\alpha_i - \alpha_i^*) \\ \text{subject to} \\ \mathbf{e}^\intercal(\boldsymbol{\alpha} - \boldsymbol{\alpha^*}) = 0, \\ 0 \leq \alpha_i, \alpha_i^* \leq C, i = 1, \ldots, l, \\ \text{where} \\ Q_{ij} = K(\mathbf{x_i}, \mathbf{x_j}) \equiv \phi(\mathbf{x_i})^\intercal \phi(\mathbf{x_j}). \\ \text{and} \\ \mathbf{e} = [1, \ldots, 1]^\intercal \end{aligned} \tag{3.7}$$

The function $K(\mathbf{x_i}, \mathbf{x_j})$ is the kernel function that allows mapping of the problem into higher dimensional spaces. In this study both the linear and radial basis kernels are considered. The radial basis kernel is given by (3.8) and the linear kernel by (3.9) [24].

$$K(\mathbf{x_i}, \mathbf{x_j}) = e^{-\gamma|\mathbf{x_i} - \mathbf{x_j}|^2} \tag{3.8}$$

$$K(\mathbf{x_i}, \mathbf{x_j}) = \mathbf{x_i}^\intercal \cdot \mathbf{x_j} \tag{3.9}$$

After solving the optimization problem in (3.7) the regression function can be expressed as (3.10). The resulting model output is $\boldsymbol{\alpha^*} - \boldsymbol{\alpha}$ [24].

$$\sum_{i=1}^{l}(-\alpha_i + \alpha_i^*)K(\mathbf{x_i}, \mathbf{x}) + b. \tag{3.10}$$

The problem thus depends on the parameters $\epsilon$ and $C$. If the radial basis kernel is used the parameter $\gamma$ is also influential for the result of the training. To find a parameter setting that minimizes the RMSE a search over the parameter space is performed. The search is divided into two parts; first a suitable value for $\epsilon$ is searched for by varying its value over a large search space while keeping the values of $C$ and $\gamma$ constant; secondly, the found value for $\epsilon$ is used while performing a grid search over varying values of $C$ and $\gamma$. The reason for dividing the search into two steps and not doing a grid search in three dimensions is primarily due to the long execution time such a search would require. To evaluate which parameter settings are most succesful the RMSE of the resulting fits are compared.

## 3.8 Evaluation

To determine which models are best suited to fit the data error metrics are computed of the models prediction on test data. The error metrics used are RMSE and percent error (see Section 3.2). The metrics are analyzed using statistical tests to determine if there is a significant difference in performance between the models and between sampling rates. The tests used in the study are 2-way analysis of variance (ANOVA) and the Friedman test. They are described in further detail below.

In order to evaluate which features are most influential for fuel consumption I will use the linear regression model and the random forest model to extract feature importance ratings. The ratings will be analysed to determine which features are rated as most important by several different implementations of the models.

### 3.8.1 2-way analysis of variance

The 2-way ANOVA test is a statistical test for analysing the influence of two different indepent parameters on a single dependent variable [26]. 2-way ANOVA is a parametric test that makes strong assumptions about the distribution of the data [26]. In this study 2-way ANOVA is used to assess what effect the choice of model and the choice of sampling rate has on the prediction error.

### 3.8.2 Friedman test

The Friedman test is a non-parametric statistical test that can be used for the same purpose as 2-way ANOVA [27]. Unlike the 2-way ANOVA test, the Friedman test does not make any assumptions about the data distribution [27]. In this study the Friedman test is used to assess what effect the choice of model and the choice of sampling rate has on the prediction error. The test statistic for the Friedman rank test is described by (3.11) [27].

$$F_R = \frac{12}{rc(c+1)} \sum_{j=1}^{c} R_j^2 - 3r(c+1)$$

where

$R_j^2$ = square of the total of the ranks for group $j (j = 1, 2, \cdots, c)$

$r$ = number of blocks

$c$ = number of groups (3.11)

## 3.9 Practical implementation

The aspects of the study that concerns ML methods and statistical analysis will be carried out using the R programming language. In data collection and preprocessing the languages Java, C# and Python will be used.

# Chapter 4

# Data collection and processing

The data that will be used for training comes from four different sources. The first and arguably the most important data source is the FM data that is collected by Scania and stored in their own database system. To complement the FM data there is also map data and information about road characteristics which comes from Scania's parent company Volkswagen, as well as weather data which is accessed through SMHI's web based interface [28] and vehicle configuration data from an internal Scania system.

## 4.1 Fleet management data

The FM database contains information collected from vehicles operated by many different companies all over the world. This study however will be limited to a small subset of these vehicles, namely the ones operated by the Scania Transport Laboratory. This limitation is made since the Transport Laboratory's vehicles send messages to the FM database with a frequency of 1 minute, instead of the normal frequency of 10 minutes which is used by most production vehicles. With this higher collection frequency I have the opportunity to evaluate if such a high frequency is necessary or if it is possible to achieve good results using only a tenth of the data points.

The Transport Laboratory's vehicles send more detailed data than the production vehicles. For instance they send information about weight, a parameter which is missing in the data from many production vehicles. Svärd has shown that the vehicle weight is one of the most influential parameters in his predictive model [7] and it makes intuitive sense that the weight of the vehicle would have a large influence on the fuel consumption.

A large amount of data is sent from the vehicles and the parameters which are of interest for this study are summarized in Table 4.1.

The GPS positions received in the FM system are not always highly accurate. The precision of the GPS units mounted on the vehicles depends on many environmental factors such as whether or not the vehicle is driving in an urban area with many

houses close to the road, or close to high mountains. It can also depend on how many satellites are within range and a wide range of other factors. However, for most purposes the GPS units in question are considered by Scania employees to have an accuracy of a couple of meters.

**Table 4.1.** Variables of interest from the FM data.

| Variable name | Description |
|---|---|
| Heading | Vehicle heading |
| Latitude | Latitude of the vehicles position |
| Longitude | Longitude of the vehicles position |
| Speed | Vehicle speed |
| Time position | The time the message was recorded |
| Vehicle ID | Vehicle identifier |
| Odometer | Accumulated distance in total |
| Total fuel | Accumulated fuel in total |
| Total fuel idle | Accumulated fuel when idling |

The Transport Laboratory have had ca 70 different vehicles in operation which have sent data to the FM system. These vehicles have been driven in many different places in Europe. The most travelled route is between Södertälje in Sweden and Zwolle in Holland but they also travel routes in Central and Eastern Europe and the Balkans. Figure 4.1 visualizes the travelled routes for all considered vehicles. The scope of this project however is limited to weather readings from Swedish weather stations. Because of this limitation the study will be limited to fuel predictions in Sweden. One interesting question is if the model may generalize and make accurate predictions in other countries.

The FM data that is available for training comes from the same database that Svärd [7] and Lindberg [13] used in their research. With the key difference that another year's worth of data has been collected. Figure 4.2 shows a histogram of the number of messages collected per month since the data collection started in March 2013. The increase of messages in the summer of 2013 corresponds to the time which the Transport Laboratory's vehicles started sending messages with 1 minute frequency. For this study I choose June 1st 2013 as the lower bound of the date interval for selecting data points. This is the same lower bound used by Svärd [7].

The FM database also contains information about driver behavior characteristics such as the number of brake application during a time period or the time spent in gears not suited for the current speed. This driver behavior data has been aggregated in the database and is only available with a temporal resolution of ca 1 hour. To deal with this limitation I calculated averages over time of the driver behavior data. The variables of interest are detailed in Table 4.2.

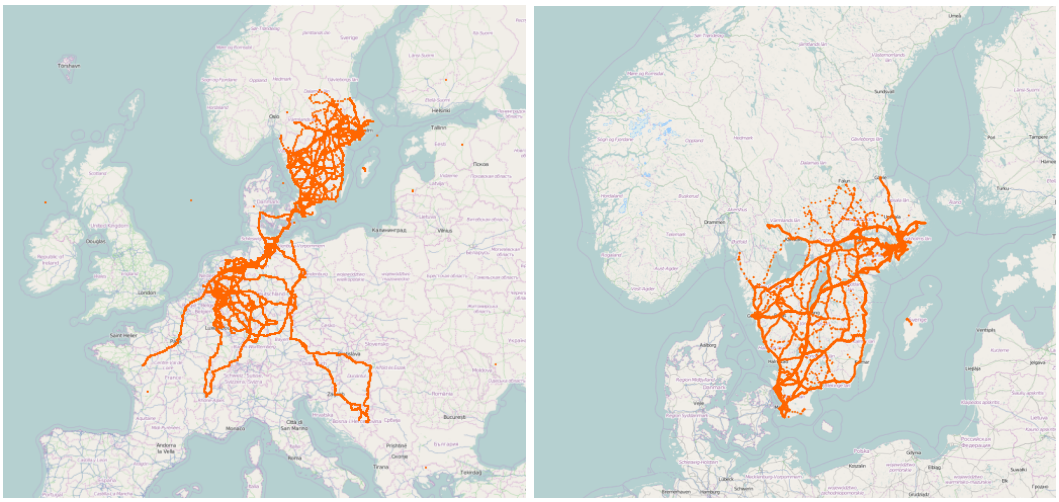

**Figure 4.1.** Maps overlayed with positions reported from the Transport Laboratory's vehicles. Each dot on the map represents a position message. In the right image one can see positions limited to inside of Sweden, it seems that no positions north of Gävle have been reported.

## 4.2 Vehicle data

The FM database does not contain any descriptive information about vehicle configuration. This information instead has to come from the separate PIS system. PIS is accessed using a C# wrapper for an underlying web based SOAP XML API. The service provides access to vehicle specific data such as what kind of engine, gearbox, cab, tyres, etc. the vehicles are equipped with.

In an interview with Scania employees it was concluded that the attributes described in Table 4.3 are the attributes available in the PIS system that will have greatest impact on the fuel consumption. When selecting the attributes it was taken into consideration that the vehicles are limited to those operated by the Transport Laboratory, which share some common characteristics that could consequently be eliminated from the study. Many of the attributes are qualitative in nature and are not possible to use as is in for example standard linear regression. They will therefore be pre-processed and converted into sets of boolean attributes for use in training. Figures 4.3, 4.4 and 4.5 describe how some interesting attributes are distributed over the vehicles included in the study.

It is reasonable to assume that there may be a correlation between some of these characteristics, for example the rear axle gear ratio and the fuel type. While extracting the information from the PIS system it turned out that data for some of the Transport Laboratory's vehicles was missing, subsequently these vehicles were removed from the study leaving 62 vehicles in total.

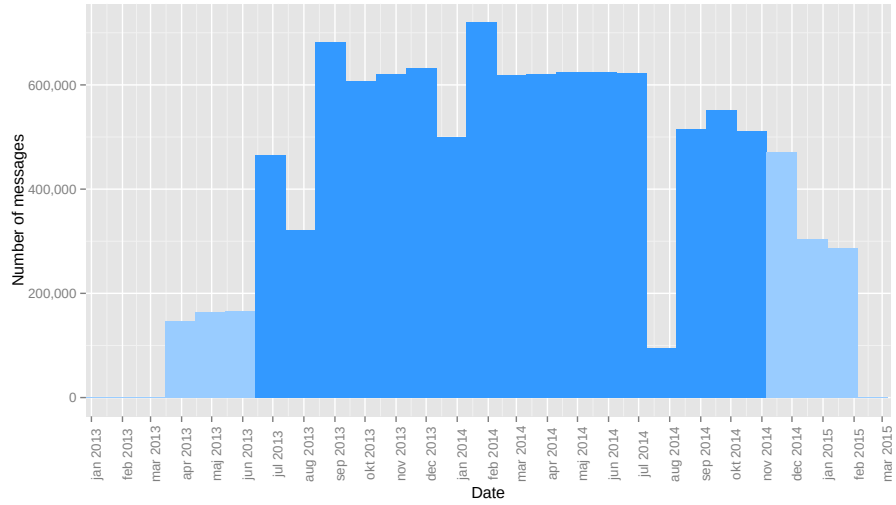


**Figure 4.2.** Histogram showing the number of messages sent from the Transport Laboratory's vehicle fleet each month since they started their FM data collection in 2013. The highlighted area shows the approximate date range that was used in the end. The dips in the histogram which occur during the holiday seasons and summers of 2013 and 2014 can be explained by the drivers taking holiday. In total almost 11 million valid messages have been received from the Transport Laboratory.

## 4.3 Road data

The road data comes from the system DigitalReality 3.0 which is provided by Scania's parent company Volkswagen. The DigitalReality system is developed by Volkswagen as part of the COMPANION project. It is a GPS routing system and the desktop installation consists of a frontend GUI, or workbench, which allows high level interactions with the underlying map data. Using the workbench one can visualize map data and routes and have access to methods for routing between two or more waypoints. The system also includes a low level Java API which provides methods for querying the map database using the Java programming language. The database itself is in the Navigation Data Standard (NDS) format and has been purchased from TomTom which is a navigation and mapping company.

All roads in the database are broken down into links. A link, in the terms of the GPS routing software, is the longest possible piece of road on which a vehicle can travel without any updated navigation instructions. For example if a road has an intersection or a roundabout the link will be broken and new links will be added. The intersection is itself not a link, but rather a link-connector element. In terms of the systems data model a new link will be created whenever a value of the fixed attribute set changes along a road, e.g. whether or not the road is a bridge or a tunnel or if it passes through an urban area. Each such link has a number of intrinsic properties (for instance the members of the fixed attribute set) as well as a number of computed properties such as the average slope and the average speed.

**Table 4.2.** Variables of interest in the driver behavior data.

| Variable name | Description |
|---|---|
| Distance with trailer | The total distance driven with a trailer attached during the period |
| Time overspeeding | Time spent over 80 km/h during the period |
| Time overrevving | Time spent in high revolutions during the period |
| Harsh brake applications | Number of harsh brakes during the period |
| Brake applications | Number of brakes during the period |
| Harsh accelerations | Number of harsh accelerations during the period |
| Time out of green band driving | Time spent in environmentally optimal revolutions during the period |
| Time coasting | Time spent coasting during the period |
| Distance with vehicle warnings | Distance driven with vehicle warnings during the period |
| Distance with CC active | Distance driven with cruise control during the period |
| Distance moving while out of gear | Distance travelled in neutral gear during the period |
| Calculated vehicle weight | An estimate of the vehicle weight, measured by the suspension |

The properties of the links that are likely to influence fuel consumption are detailed in Table 4.4.

For the purpose of fuel consumption prediction this division into links might not be optimal. One could argue that it would be better to divide the road into segments based on other properties such as slope or curvature, since Svärd [7] has shown that the slope is so important for fuel consumption. Dealing with this limitation will prove a challenge during the data consolidation process.

## 4.4 Weather data

The weather data for this study comes from SMHI and is accessed through their public web based interface [28]. This is an improvement on the study by Lindberg [13] as the data available over the API has higher temporal och spatial resolution than the data sets used in his study. SMHI [28] states that their historical observations can only be considered accurate if they are older than three months. SMHI has a correction and quality control process which cannot be guaranteed to have finished for observations that are less than three months old [28]. This puts a restriction on what time period of FM data should be used for training the model. Since the data

**Table 4.3.** Variables of interest in the PIS system.

| Variable name | Description |
|---|---|
| Product class | Whether the vehicle is a truck or a bus |
| Technical total weight | The weight of the vehicle |
| GTW technical | The maximum allowed gross trailer weight |
| Engine stroke volume | The volume of the engine |
| Horsepower | The power of the engine |
| Rear axle gear ratio | The rear axle gear ratio |
| Emission level | The emission level, one of 3 classes |
| Overdrive | Whether the vehicle has overdrive or not |
| Ecocruise | Whether the vehicle has ecocruise or not |

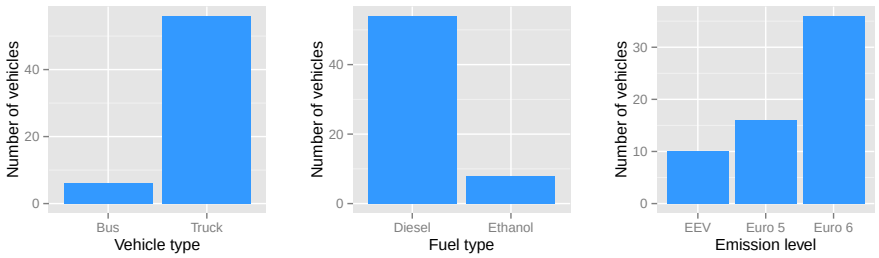


**Figure 4.3.** Description of the vehicles. The left diagram shows the distribution of trucks and buses. The middle diagram shows the fuel types used by the vehicles and the right diagram shows which emission level classes the vehicles have. It is evident that most of the vehicles in this study are diesel trucks with the Euro 6 emission level.

is extracted in February 2015 it makes October 31st 2014 the upper bound of the date interval for selecting data points.

The weather data consists of meteorological observations from SMHI's weather stations spread over the country. Each weather station measures different parameters, some measure several parameters and others measure only one. Figure 4.6 shows heatmaps of the distributions of weather stations in the country. It is evident that the southern part of Sweden has best weather station coverage. Comparing Figure 4.6 with Figure 4.1 promises good coverage of weather observations for the FM data in this study.

Similar patterns as in the left heatmap of Figure 4.6 emerge when examining the distribution of weather stations that track other parameters than wind. There are a large number of weather stations in the country that track one or a few of the parameters important for this study, but there are only a limited set of stations that track all of them. Dealing with this limitation will be a challenge when consolidating the data.

The parameters that are included in this study have been chosen in part based

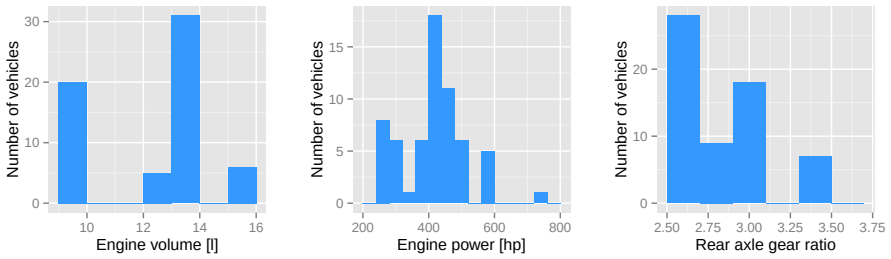


**Figure 4.4.** Descriptions of the engine characteristics.The left image shows engine volumes and the middle image horsepower. The right image shows rear axle gear ratios, which is a factor that may have large impact on the fuel consumption according to Scania employees.

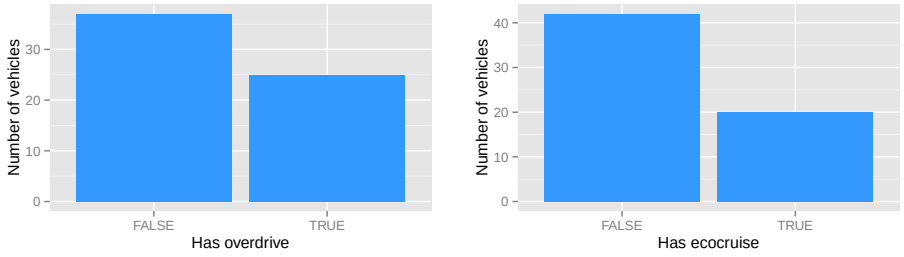


**Figure 4.5.** Descriptions of the vehicles' control systems. The left diagram shows how many of the vehicles have an overdrive gearbox, which may reduce fuel consumption if in place. The right image shows how many of the vehicles have an ecocruise system installed, which may also reduce fuel consumption.

on what parameters are available in the data from SMHI but also on the parameters that Svärd [7] has shown to have greatest influence in his research. The chosen parameters are described in Table 4.5. Each meteorological observation is coupled with a timestamp and a position. The spatial resolution is described by Figure 4.6. The temporal resolution varies between different stations and parameters, Figure 4.7 shows a bar chart describing the distribution of collection frequencies over station-parameter pairs.

## 4.5 Data selection and filtering

Due to the facts stated in the above sections the FM data is limited to:

- vehicles operated by Scania Transport Laboratory,
- vehicles that have their configurations documented in PIS,
- messages sent from inside of Sweden,
- messages sent between June 1st 2013 and October 31st 2014,

**Table 4.4.** Variables of interest in the road data.

| Variable name | Description |
|---|---|
| Average slope | The average slope of the link |
| Average speed | The estimated average speed while driving the link |
| Speed limit | The speed limit of the link |
| Administrative road class | The type of link, e.g. highway or local road |
| Bridge | Whether the link is a bridge or not |
| Tunnel | Whether the link is a tunnel or not |
| Urban | Whether or not the link is near an urban area |

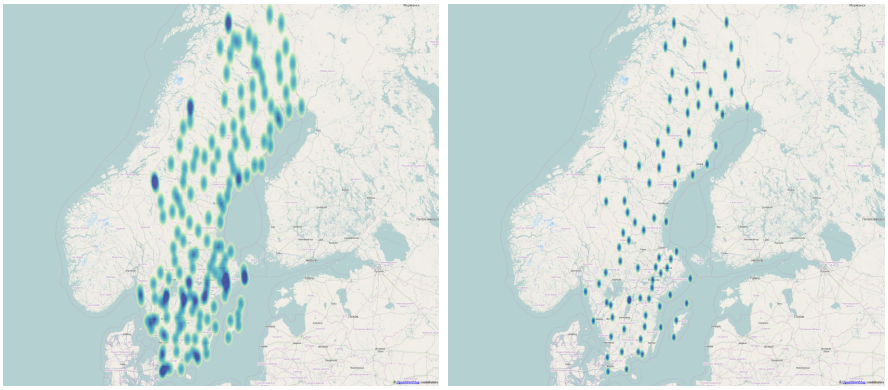

**Figure 4.6.** The map to the left shows a heatmap of all active weather stations in Sweden that measure wind speed and direction. The map to the right shows all active weather stations that measure all of the parameters wind speed, wind direction, temperature, air pressure, humidity and precipitation. The single parameter heatmap to the left displays much better coverage than the multiple parameter heatmap to the right.

- messages which include their vehicles fuel readings,
- messages which include their vehicles calculated weight.

The FM database also contains incorrect data. There are examples of messages that imply that the vehicles travel at speeds exceeding 300 km/h and that some do not lose fuel over several kilometer-long stretches. To deal with this faulty data Scania has developed a filtering routine. The filtering process selects all messages such that the vehicle is in motion and that the speed and fuel consumption lie within sensible bounds.

Using this filtering process and applying the selection criteria listed above results in a set of over 5 million messages. This data set is the raw data to be used for training but will be further reduced by other sanity checks and pre-processing steps during the data consolidation process.

**Table 4.5.** Variables of interest in the meteorological data.

| Variable name | Description |
|---|---|
| Temperature | The air temperature in °C |
| Wind speed | The wind speed in m/s |
| Wind direction | The wind direction in degrees |
| Humidity | The relative air humidity in % |
| Air pressure | The air pressure in hPa |
| Precipitation | The amount of rain or snow fall in mm/h |
| Current weather | A qualitative description of the weather |

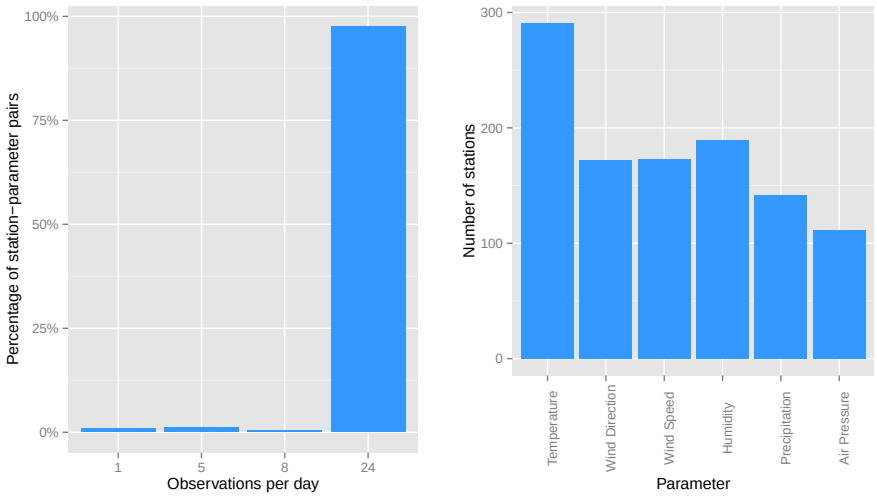


**Figure 4.7.** The left diagram shows the distribution of the number of observations per day. The different stations may have different collection frequencies for different parameters, for instance weather station A might collect 24 temperature observations per day but only 8 wind speed and wind direction observations. The diagram shows that most of the parameters are collected once per hour by most of the stations. The right diagram shows how many stations track the different parameters. It is evident that most parameters have comparable coverage, while temperature has better and air pressure worse than average.

## 4.6 Data consolidation

Figure 4.8 shows a schematic overview of the data consolidation process. The data from the different sources is downloaded locally and filtered to only include relevant data points as per the data selection section above. The data points are then matched with each other using matching criteria. The resulting consolidated data set is put through a series of pre-processing steps to compute features and prepare the data for use in training. The details of this process are described in the following sections.

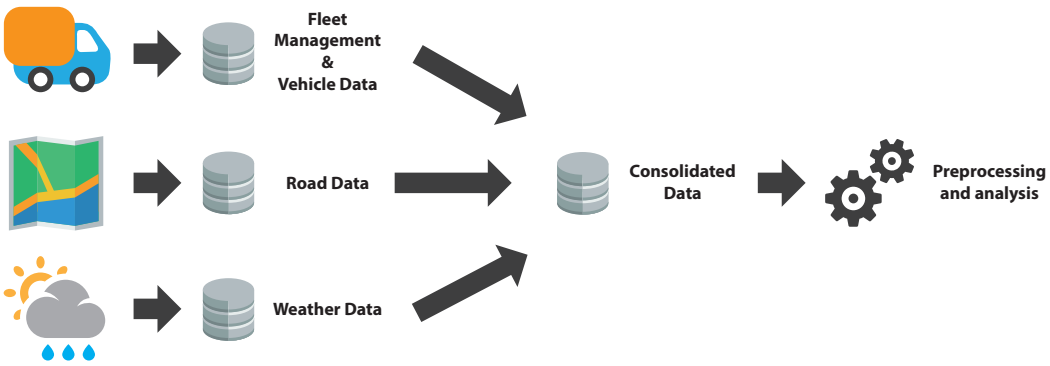


**Figure 4.8.** Schematic view of the data consolidation process.

### 4.6.1 Local database

In order to consolidate the data I need a database for storing the information from the different sources. For this purpose I chose to set up a local installation of the PostgreSQL object-relational database system. The reason for this is primarily the PostGIS extension which is available for PostgreSQL and allows spatial queries [29]. For example it makes it easier to extract all points that lie within a certain geometry (such as the Swedish borders) or to calculate the distance between two GPS coordinates. PostGIS is also possible to integrate with QGIS which is an open source GIS software. QGIS is useful for for visualizing and analyzing the geographical data.

### 4.6.2 Pairing the FM position messages to calculate fuel consumptions

In order to evaluate fuel consumption a single position message from the FM database is not enough. The messages must be examined in pairs in order to calculate the distance driven, speed and fuel consumption between the two messages. I constructed a program that iterated all position messages and paired them with the position message that was closest after in time and was sent from the same vehicle as the first message. These message pairs were then filtered to only include pairs with exactly 60 seconds between them. The fuel consumption and vehicle velocities described by these pairs are illustrated in Figure 4.9.

The message pairing routine was then repeated to compile a data set of pairs with exactly 10 minutes between the messages. The position messages were iterated in the same way as before, but ca 9 out of 10 messages were skipped over until pairs with 10 minutes between them were found. In this way a subsampled data set of approximately 10 % the size of the original set was created. This alternative data set will be used in training in the same way as the data set with 1 minute

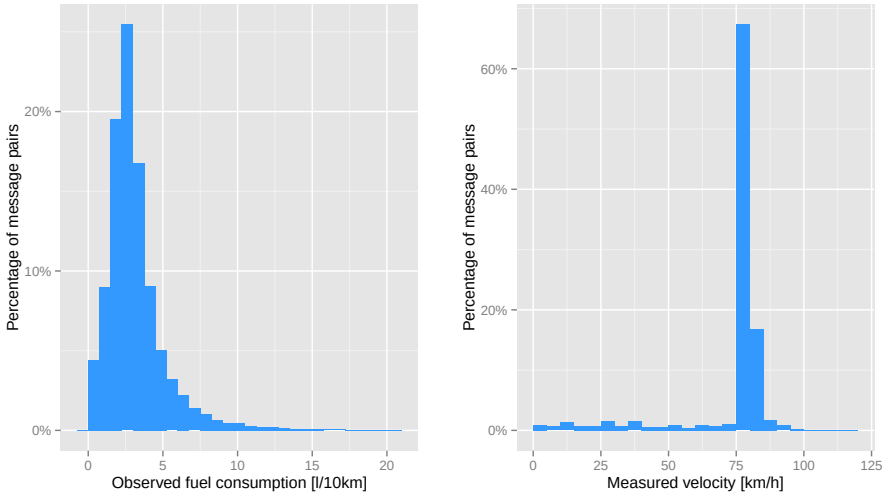


**Figure 4.9.** The left histogram shows the observed fuel consumption for pairs of messages sent with 1 minute frequency. The right histogram shows the measured velocities for the same message pairs. It's clear that most messages come from vehicles driving at speeds around 75-80 km/h and that their fuel consumption is distributed with a mean close to 31 per 10 km.

frequency. The resulting models will then be compared in order to evaluate if a 10 minute frequency is sufficient for building a usable fuel model. Drawing the fuel consumption observations from the same population in this way ensures a fair comparison between the models.

When examining the two different data sets side by side it is clear that the 1-minute data set has much larger variation in the observed fuel consumption values when compared to the 10-minute data set. Figure 4.10 illustrates this with an example of observed fuel consumptions from both data sets. Figure 4.10 shows two time series taken from the same truck driving on the same road during the same period. The data in the 10-minute data set can be seen as average values of the data in the 1-minute data set.

There were position messages in the original data set for which no pairing was possible or for which the frequency was wrong, these messages were discarded and not used in the final data sets. After discarding these messages ca 2.7 million out of the original 5 million messages remained.

### 4.6.3 Matching weather observations and FM position messages

To match the weather observations to position messages I use a nearest neighbour search of the weather stations using the GPS coordinates of the position message as starting point. The search algorithm first fetches the closest station and inspects which parameters are available there, it then continues with the second closest and inspects which parameters are available there. The algorithm continues in this

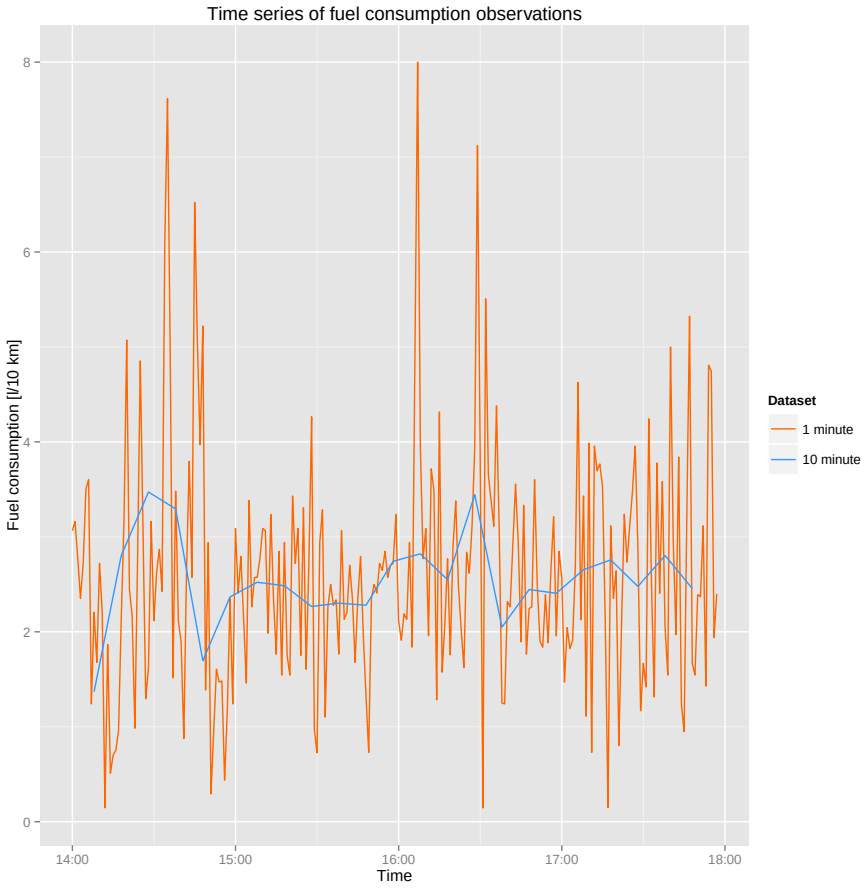


**Figure 4.10.** Comparison of a time series of fuel consumption observations extracted from the 1 minute and 10-minute data sets. The 1-minute data set shows a much higher variance with a maximum near 8 l/ 10 km and a minimum close to 0 l/10 km during the 4 hour time frame. The 10-minute data set has less variance and a smoother curve, with a maximum near 3.5 l/10 km and a minimum near 1.5 l/10 km.

manner until stations that track all parameters have been found, at which point the algorithm terminates. Once the set of stations has been compiled they are queried for the observations that are closest in time to when the position message was sent. The approach is illustrated in Figure 4.11 and ensures good locality of observations both in time and space to the position message.

### Quantifying the wind effect

The influence of wind on fuel consumption is given by the difference in direction between the vehicle and the wind as well as the wind strength. Driving against the wind increases fuel consumption while driving with the wind reduces fuel consumption. To quantify the effect of wind on fuel consumption a feature was computed

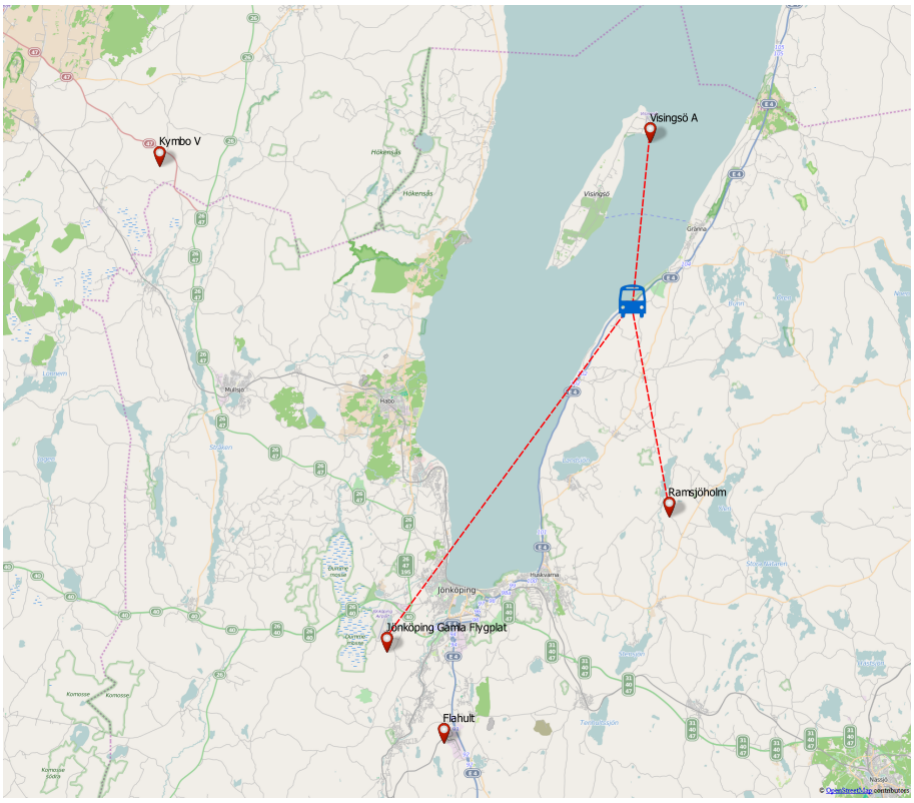


**Figure 4.11.** Illustration of how weather observations are matched to a position message. For the position just south of Gränna the closest stations are fetched and inspected in order based on their distance to the position. In this case 5 out of 7 parameters are obtained from the closest station at Visingsö. The other 2 parameters are not tracked by the second closest station in Ramsjöholm but they can be obtained from the third closest station in Jönköping. When all 7 parameters are found the search is terminated.

using (4.1), where the $Heading$ and $WindDirection$ are given in radians.

$$WindEffect = WindSpeed \cdot cos(Heading - (\pi + WindDirection)) \tag{4.1}$$

The addition of $\pi$ to the $WindDirection$ parameter has to do with the fact that the wind direction represents which direction it is blowing from, while the vehicle heading represents which direction the vehicle is driving towards. The value of the cosine function then becomes -1 if the vehicle has headwind and 1 if it has tailwind. The $WindSpeed$ parameter gives the amplitude of the wind effect.

### 4.6.4 Matching road data and FM position messages

The FM position messages each contain a position in the form of GPS coordinates, which are longitude and latitude values in the World Geodetic System 1984 (WGS

84). Using the DigitalReality Java API for accessing road data the FM position messages were evaluated in pairs and the route between the positions was found. The found routes were then verified so that the difference between the reported positions and the endpoints of the route were not too large and also so that the difference in length between the route and the reported odometer readings were not too large. An illustration of a position pair matched to a route can be seen in Figure 4.12. This routine was repeated both for the 1 minute pairs and the 10 minute pairs. In some cases the API could not find a route at all, or the found route did not meet the verification requirements. In these cases the position pair was discarded and not used in the final data set. This reduced the input from ca 2.7 million messages to ca 2 million.


**Figure 4.12.** Example of a route between two points found using the DigitalReality API. The API uses the two end points to find a set of line segments corresponding to the road between the points. The found route can then be inspected to determine its average slope as well as the total climb and descent when driving on the road. This particular route is on the E4 just south of Gränna in Småland.

Once a route was found a set of attributes was extracted using the DigitalReality Java API. The attributes chosen to describe the slope profile of the route was the average slope as well as the total climb and total descent of the route. In addition, attributes designating the type of road, average speed and other features described

in Table 4.4 were also collected and matched to the position pair.

### 4.6.5 Calculating platooning

To calculate whether a vehicle is in a platoon or not I used the fact that the vehicles' clocks are synchronized. They are synchronized in such a way that they send their position messages at the same time. This makes it possible to look at all position messages from a given moment and determine if any of the vehicles are close enough to each other at that moment to be considered to be part of a platoon. The platooning calculations were modelled on the algorithm described by Svärd [7].

To reduce the risk of including false data the messages were filtered using the following steps.

1. First the messages that seemed to not have synchronized clocks were removed. This was determined by investigating which messages were sent at times when no other messages had been sent, and removing those from the data set.

2. The second step was to separate all possible platooning candidates from the messages that decidedly did not come from a platoon. This was determined by examining the distance between messages. If a message was less than 100 meters from another message it was concluded that it could be part of a platoon.

These filtering steps divide the data set into one subset of position messages that are not part of a platoon and another subset of messages that could be part of a platoon. The second subset that contains platooning candidate messages was processed further to separate the highly probable platooning messages from improbable platooning messages using the following steps.

1. Look at the pair of position messages closest to one another.

2. Make sure that they are travelling in the same direction.

3. Make sure that the difference in bearings between the two positions and the travel direction is small enough. This ensures that the vehicles travel in the same lane

To calculate the bearing between two positions the Haversine formula for GPS coordinates was used. The expression is given in (4.2) where $Lat_1$, $Long_1$, $Lat_2$ and $Long_1$ are the latitude and longitude values of the two GPS coordinates. The calculated bearing is the compass heading between the two points. If the found bearing is close enough to the average heading of the vehicles it is assumed that the vehicles were platooning. Figure 4.13 illustrates how the bearings are compared to vehicle headings.

$$\arctan\left(\frac{\sin(Long_2 - Long_1)\cos(Long_2 - Long_1)}{\cos(Lat_1)\sin(Lat_2) - \sin(Lat_1)\cos(Lat_2)\cos(Long_2 - Long_1)}\right) \quad (4.2)$$

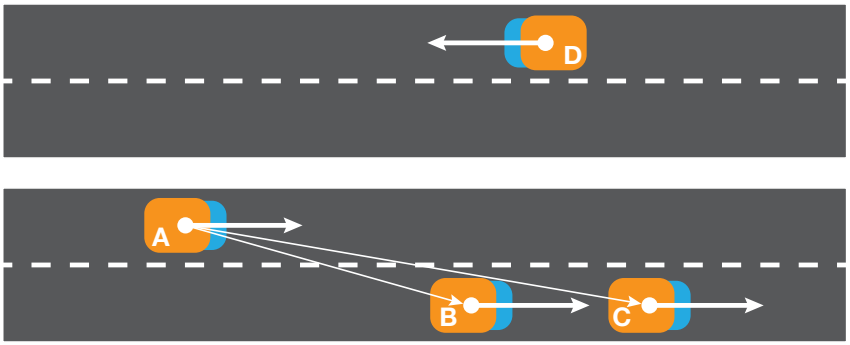


**Figure 4.13.** Illustration of how the platooning calculations were performed. To determine if the vehicles were travelling in the same lane the angles between their positions were examined. In this case we start with vehicle $A$ and can cross out vehicle $D$ from the candidate list since it is not travelling in the same direction. Vehicles $B$ and $C$ however could be part of a platoon with $A$ since they are travelling in the same direction and are close. By examining the angles between the heading of $A$ and the vectors $AB$ and $AC$ it can be concluded that the vehicles are not in the same lane and therefor not in a platoon with $A$.

All position messages that passed the steps above were determined to be positive platooning messages, all other messages were discarded due to uncertainty. This routine caused many position messages to be dropped from the data set, out of the 2.7 million input only 1.9 million messages remained. Since the GPS data has an accuracy of a few meters it is not reliable on the level of determining which lane a vehicle is in. Thus there is a risk that some of the messages that were included contain false information. Figure 4.14 illustrates this potential flaw in the method.

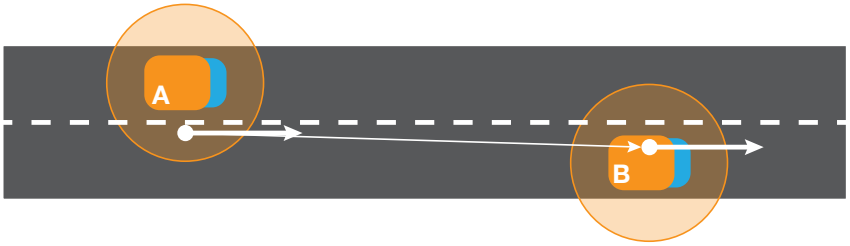


**Figure 4.14.** Illustration of a flaw in the method. Since the accuracy of the GPS positions is not guaranteed there is a risk that a method based on examining angles between vectors may produce both false positives and false negatives. In this case the true position of vehicle $A$ is illustrated by its position on the road and the reported GPS position is illustrated by the white dot and the heading vector. The same applies for vehicle $B$. In such a scenario the method would determine that the vehicles are platooning, when in fact they are not.

### 4.6.6 Matching driver behavior data and FM position messages

The data in the FM database describing driver behavior is only available with a resolution of ca 1 hour. The data is aggregated in a separate table in the FM database where properties such as the total number of harsh brakes and the other properties described in Table 4.2 have been calculated for a time span of approximately 1 hour. Since the position message pairs evaluated to determine fuel consumption have a time resolution of 1 minute there was no obvious way to match the driver behavior data to the individual messages. To solve this problem I calculated rates of e.g. harsh brakes instead of the total number by dividing the original attribute with the number of seconds during the aggregation period. This mean that the driver behavior data does not describe the driver behavior for the period of the fuel consumption observation, but an average of the driver behavior over a longer time span.

## 4.7 The resulting data set

The data sets for routes, weather, driver behavior, vehicle configuration and platooning information were compiled separately from the input of 2.7 million pairable messages. Once the data sets had been compiled they were stored in separate tables of the local database, with the observations matched to a unique position message ID that acted as a foreign key between the tables. To consolidate the data and compile a unified data set of matched observations a query was used to extract and join all data for the subset of position message IDs that were present in all tables. The resulting data set consisted of ca 1 million messages. This means that the preprocessing as described in the above sections reduced the data from 5 million to 1 million messages. Figure 4.15 illustrates how the resulting data sets are distributed geographically.

**Figure 4.15.** An illustration of the geographical distribution of the fuel observations that made it through all filtering and preprocessing steps. The left image shows observations from the 1-minute data set and the middle image shows observations from the 10-minute data set. The rightmost image shows a heatmap of the observation distribution, which is similar for both data sets. It is evident that the observations are focused on the E4 south of Södertälje and that almost all observations come from the major motorways.

After going through the preprocessing steps and filtering out the unusable data

the variation of vehicles included was reduced as well. The remaining observations came only from vehicles using diesel as fuel and having emission level Euro 6. Subsequently the features regarding fuel and emission level were removed from the data set. The final feature list used for training is described in Table 4.6.

The exact number of datapoints in the final 1-minute data set is 985 600, while the 10-minute data set has 33 677 observations. This means the 10-minute data set is only ca 3 % of the size of the 1-minute data set. Before the data consolidation process the 10-minute data set was approximately 10 % of the 1-minute data set. This reduction is primarily due to the fact that route matching was much less successful when the positions were far away from each other.

Table 4.6: All features included in the final training data.

| **Variable name** | **Description** |
|---|---|
| Day of year | The day of the year, works as a season indicator |
| Hour of day | The hour of day |
| Average heading | Average compass heading of the vehicle between the two positions |
| Calculated speed | Calculated speed of the vehicle |
| GTW technical | The maximum allowed gross trailer weight |
| Engine stroke volume | The volume of the engine |
| Engine Horsepower | The power of the engine |
| Rear axle gear ratio | The rear axle gear ratio |
| Overdrive | Whether the vehicle has an overdrive gearbox |
| Ecocruise | Whether the vehicle has an ecocruise system |
| Distance with trailer per second | Distance travelled with a trailer, divided by time |
| Overspeeding rate | Rate of overspeeding during an aggregation period |
| Overrevving rate | Rate of overrevving during an aggregation period |
| Harsh brakes per second | Number of harsh brakes divided by time |
| Brakes per second | Total number of brakes divided by time |
| Harsh accelerations per second | Number of harsh accelerations divided by time |
| Out of green band driving rate | Rate of environmentally unfriendly driving during an aggregation period |
| Coasting rate | Rate of coasting during an aggregation period |
| Distance with vehicle warnings per second | Distance travelled with warning light divided by time |
| Distance with CC active per second | Distance travelled using cruise control divided by time |
| Distance moving while out of gear per second | Distance travelled in neutral divided by time |
| Calculated vehicle weight | The estimated weight of the vehicle during an aggregation period |
| Platooning | Whether the vehicle is in a platoon |
| Platooning distance | The distance to the nearest vehicle in the platoon |
| Average slope | The average slope of the route |

Table 4.6: All features included in the final training data.

| **Variable name** | **Description** |
|---|---|
| Total climb | The calculated total climb of the route |
| Total descent | The calculated total descent of the route |
| Average speed | The estimated average speed of the route |
| Speed limit | The speed limit of the link |
| Max administrative road class | The maximum admin class of the route |
| Bridge | Whether the route has a bridge |
| Urban | Whether the route is near an urban area |
| Average temperature | The average temperature between the two messages |
| Wind factor | The average wind factor between the two messages |
| Average humidity | The average humidity between the two messages |
| Average air pressure | The average precipitation between the two messages |
| Average precipitation | The amount of rain or snow fall in mm/h |

# Chapter 5

# Results

## 5.1 Dividing and normalizing the data

Before building the different models the data is divided into training, validation and test data sets. The training set is used to train the different models and the validation test is used to verify and evaluate the models during iterative training in order to select the best meta parameters. Once the meta parameters have been chosen and a finished model selected the final model is tested on the test data to evaluate its performance.

The division is done by partitioning the data based on date information. The data used for training is the data collected during one year between 2013-06-01 and 2014-05-31. The data used for validation and testing is the rest of the data collected between 2014-06-01 and 2014-10-31. The validation and test data sets are also split based on date partitioning so that the two sets do not contain observations from the same dates. Partitioning based on date ensures that there is no strong correlation or dependencies between the data in the different sets. If a random sampling method had been applied there would be a risk that observations coming from the same vehicle and the same time period would appear in both training and test sets, giving a strong correlation between the data sets.

If samples that are close to each other in a time series (from the same truck, same road, same weather conditions and same driver behavior values) end up in each of the three data sets (training, validation and test) it could lead to high correlation between the data sets. This in turn could potentially lead to overfitting. Dividing the data by date minimizes the risk for very similar data points ending up in all three data sets, thereby reducing the risk for strong correlation between the data sets.

Partitioning in this way gives a split of ca 67 % training data, 11 % validation data and 22 % test data. Since the validation data can be thought of as part of the training data the major split into training and test sets is approximately 80 % to 20 %.

After splitting the data the mean and standard deviation of all features in the

training partition were computed to perform whitening transformation and then used to transform the features in the validation and test partitions.

## 5.2 Data analysis

The data is analyzed to find correlations between different features. A correlation plot is presented in Figure 5.1. Strong correlations are found between the features describing the engine characteristics. One interesting question is if the engine features could be reduced to one or two descriptive features using principal component analysis (PCA) or some other method for dimensionality reduction. This question is beyond the scope of this study but could be worth investigating in future research.

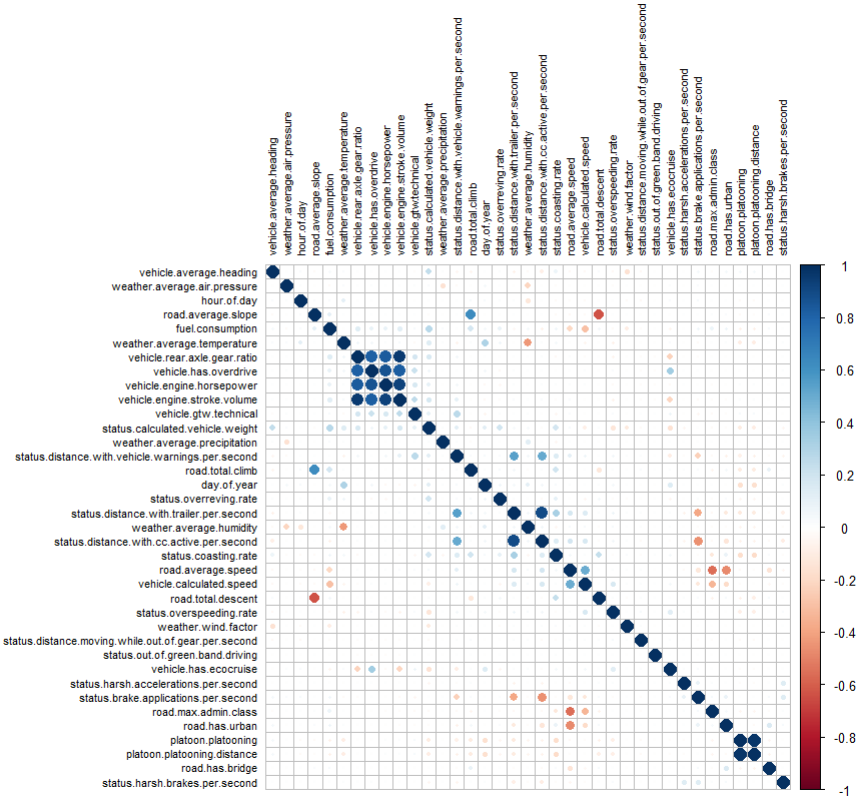


**Figure 5.1.** Correlation plot of the features in the 1-minute data set. Dark blue circles represent strong positive correlation and dark red circles represent strong negative correlation. The circles grow in size depending on the absolute value of correlation. Empty boxes indicate that there is no correlation between the parameters. The diagram shows that the engine parameters are strongly correlated with each other and that the three parameters describing the slope profile of a road are also correlated. Correlations between the road type and speed limit are found as well.

The two data sets, with 10 minute and 1 minute sampling rate respectively, were compared with regards to median value and variance. The results are shown in Figure 5.2. The 1-minute data set has higher variance and heavier tails compared to the 10-minute data set.

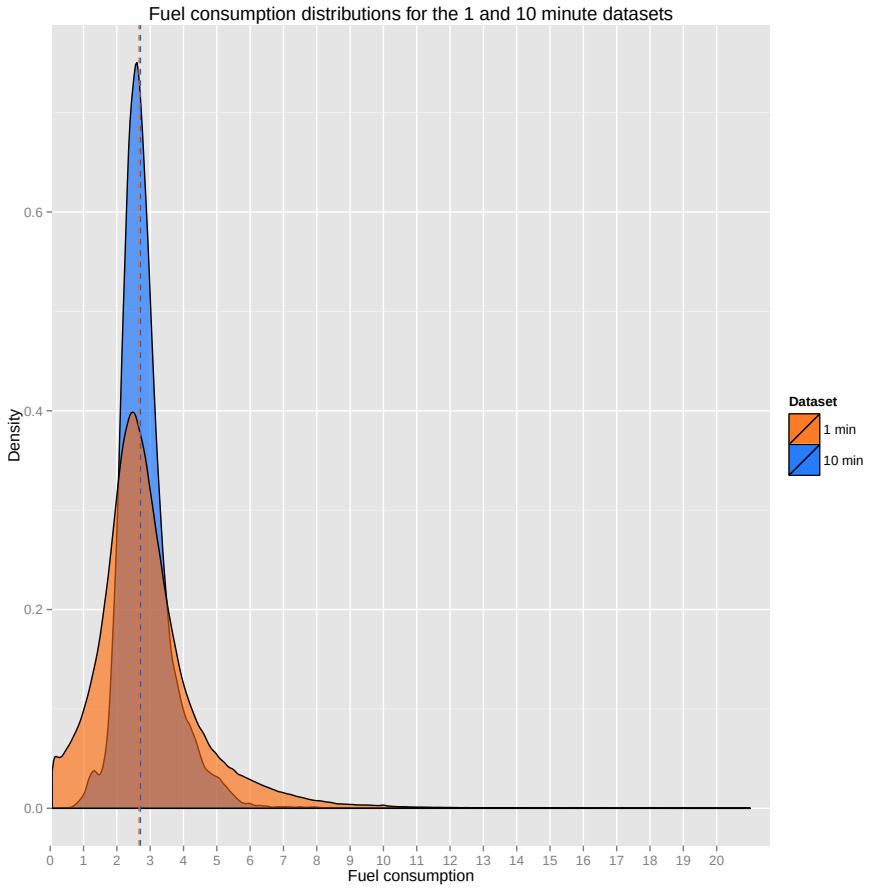


**Figure 5.2.** Gaussian estimate of the probability density function (PDF) of the fuel consumption for the different data sets. The 1-minute data set is plotted in orange and the 10-minute data set in blue. The dashed lines represent the median values of the different distributions. The diagram shows that the median values are very close to each other for both distributions but that the data set with a finer sampling rate has larger variance.

## 5.3 Linear regression

As a first step linear regression was applied to all the observations in the data in order to investigate the distributions properties and see how well the data could be fitted by a simple linear model. To do this I used the R programming language and its built-in functions for linear regression.

When training the linear model it became evident that 3 of the features in the data set were constant and thus could not have any influence in the training. These features indicate if the vehicle is a bus, if it is a truck and if the road has a tunnel or not. They were constant because only trucks and roads without tunnels made it through the filtering steps. The features were removed from the data set and not used in training.

To evaluate the fit of the linear model plots describing the residual vs fitted values as well as Cook's distance vs leverage were created. The diagrams are shown in Figure 5.3. The diagrams imply that a linear model cannot successfully model the underlying relations in the data. The Cook's distance plot also suggest that some points should be examined extra carefully and possibly removed from the data set before training the more advanced models.

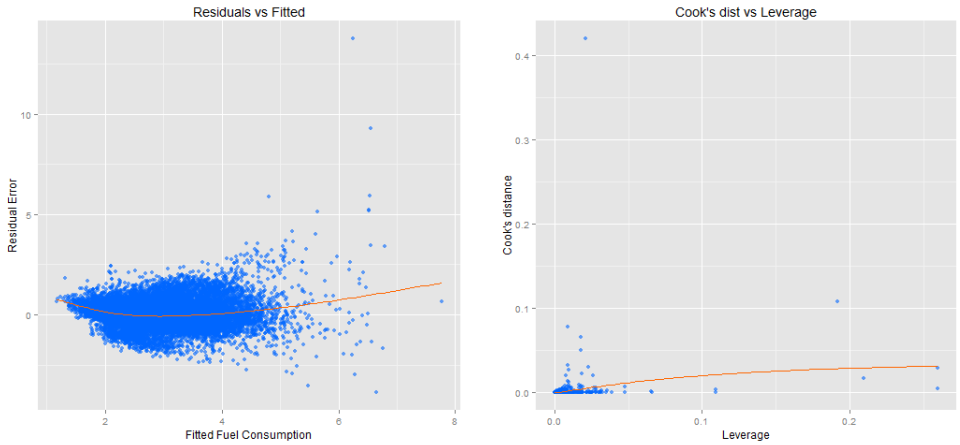


**Figure 5.3.** Plots describing how well the linear regression model fits the 10-minute data set. The orange lines shows the smoothed means of the blue points. The left image shows how the residual errors vary depending on the fitted value for fuel consumption. When the values are close to 3 l / 10 km the average residual is near 0 and the maximum absolute residual is around 2-2.5. But when the prediction is lower the variance is lower, and when the prediction is higher the variance is higher. The mean error also becomes greater for the extreme values. This implies that the distribution is heteroscedastic and nonlinear and cannot be accurately described by a linear model. The right image shows the Cook's distance vs leverage of each point in the data set. If a point has a high Cook's distance and/or high leverage it means that point is highly influential in the model result. These points should be checked extra carefully for validity and possibly be removed from the data set.

The points indicated by the Cook's distance vs leverage plot turned out to be faulty on closer inspection. The point with highest Cook's distance indicated by the plot contained an extreme value of fuel consumption of 20 l / 10 km. This data point was considered an outlier and removed from the data set. The 1-minute data set was also filtered to remove outliers using the same method. The residuals vs fitted and Cook's distance vs leverage plots for the 1-minute data set, after outlier removal, can be seen in Figure 5.4.

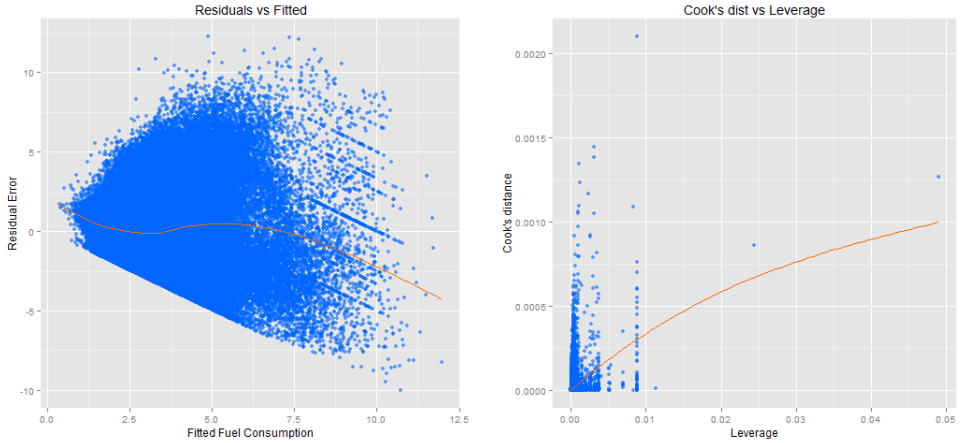


**Figure 5.4.** Residuals vs fitted and Cook's distance vs leverage plots for the 1-minute data set after outlier removal. The orange lines show the smoothed means of the blue points. The sharp lower boundary in the left image indicates the 0 value for fuel consumption. Both diagrams show similar properties as the diagrams for the 10-minute data set. The right plot looks different due to the compressed y-axis, but represents similar conditions. The heteroscedasticity in the left image is much more pronounced compared to the 10-minute data set. This implies that the variance in fuel consumption is greater when using a finer sampling rate. It also indicates that neither of the data sets can be accurately described by a linear model.

When trained on the training data and tested with the testing data the linear models for the 10 and 1-minute data sets achieved an RMSE of 0.80 and 1.65 respectively. The results are summarized in Table 5.1. It is interesting to note that the 10-minute data set was better suited for linear regression compared to the 1-minute data set. This is most likely because of the lower variation of fuel consumption observations in the 10-minute data set.

**Table 5.1.** Summary of the RMSE values for the linear regression model.

| | **Training set RMSE [l/10 km]** | **Test set RMSE [l/10 km]** |
|---|---|---|
| **1-minute data set** | 1.24 | 1.65 |
| **10-minute data set** | 0.46 | 0.80 |

The linear regression model is used to perform a variable importance rating in which the individual features are ranked based on their associated weight in the found linear equation. The results of the variable importance rating are illustrated in Figure 5.5.

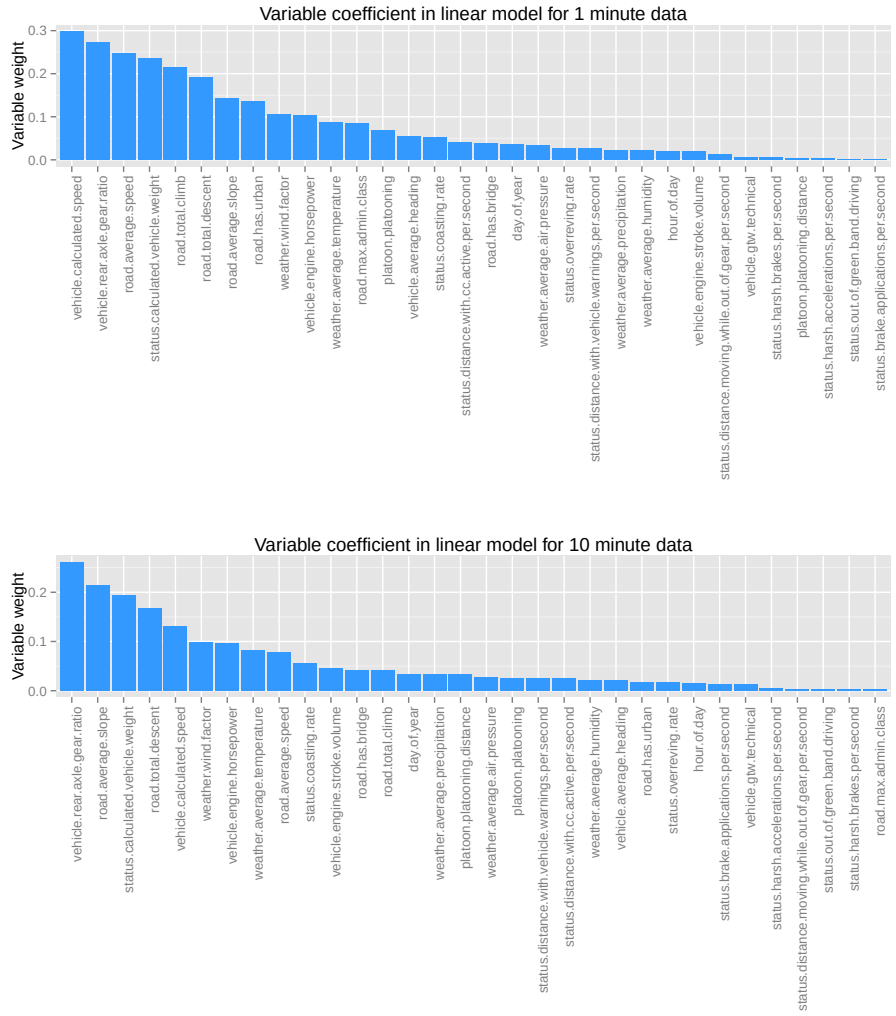


**Figure 5.5.** The weights of the linear models plotted in a bar chart. The top diagram shows the weights for the 1 minute fit. The bottom diagram shows the weights for the 10 minute fit. The y-axis of the diagrams are scaled by taking the base 2 logarithm, this is because there is a very large difference between the most important weight and the other weights. The diagrams show that vehicle weight, vehicle speed and road slope are very important in both models. The diagrams also show that the driver behavior features are more descriptive for the 10 minute set. This can be assumed to be because a 10 minute sampling rate is closer than a 1 minute sampling rate to the 1 hour sampling rate we have for driver behavior.

## 5.4 Random forest

For fitting a random forest to the training data I used the R package `randomForest`, which implements Breiman's random forest algorithm for classification and regression [30]. I used the data divided into training, test and validation sets as described

above.

As described in the method section a random forest works by building several regression trees using random subsamples of the features for each tree. The parameter `mtry` defines the number of features to be used in the trees. The default value of `mtry` is set to the square root of the total number of features, with 36 features this comes out to a default value of 6 features per tree. However, in order to find the optimal parameter setting I iterated values of `mtry` and for each value trained a model and evaluated the RMSE of both the training and validation sets. The `ntrees` parameter, which defines how many trees to train for each model, was held constant at 100 in order to speed up the calculations. This experiment was done on the 10-minute data set as it is significantly smaller and therefore faster to train compared to the 1-minute data set. Initial prototyping showed the RMSE decreases as `mtry` approaches the upper bound of 36. It also showed that the decrease in RMSE is very small after the value of `mtry` passes 10. The RMSE as a function of `mtry` is plotted in Figure 5.6 and compared to the RMSE of a standard regression tree. The final random forest fit for the 10-minute data set was trained using an `mtry` value of 10 and an `ntrees` value of 500, keeping the RMSE low while optimizing training speed.

For the 1-minute data set I reduced `ntrees` to 100 and subsampled the data to only use 150 000 data points out of the total ca 650 000. This was done due to the very large memory and time requirements of training a random forest for the amount of samples in the 1-minute data set. Reducing the data set in this way may give a worse fit than if the entire data set had been used. The results of the training are summarized in Table 5.2.

The final random forest fit was used to evaluate the relative importance of the different features, measured by the increase in RMSE caused by removing the features from the data set. The results of this rating is described by Figure 5.7.

## 5.5 Support vector regression

The SVM-implementation available through the R-package `e1071` is used to fit an SVR model to the data. The `e1071` package provides an interface to `libsvm`, the C++ implementation by Chih-Chung Chang and Chih-Jen Lin [31].

SVR performance depends on several parameters, namely the choice of kernel function, the value of $\epsilon$ and settings for $\gamma$ and $C$. To find suitable parameters for estimating fuel consumption the parameter space was searched in iterations. The first search was for a good value of $\epsilon$. $\epsilon$ has an effect on the smoothness of the SVR response and it affects the number of support vectors, so both the complexity and the generalization capability of the network depend on its value [32]. During this search the radial basis kernel function was used and the $\gamma$ and $C$ parameters were held constant. To make the search more time efficient the data used in training was subsampled to 4000 data points. The result of the search for the 10-minute data set is illustrated in Figure 5.8. Using the plots in Figure 5.8 as a guide an $\epsilon$ value of 0.25 and 1.75 is chosen for the 1 and 10-minute data sets respectively. A

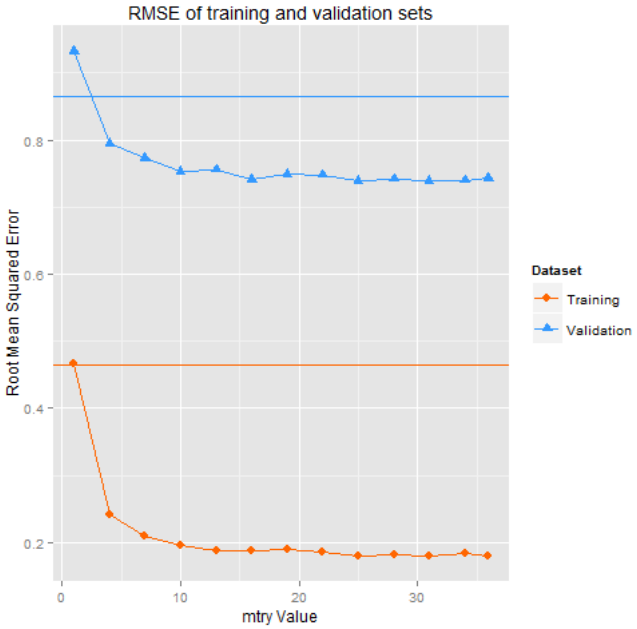


**Figure 5.6.** Plot of the RMSE of the training and validation sets for the random forest prototype. The horizontal lines represent the RMSE of a standard regression tree trained on the same data. The `mtry` parameter is incremented with 3 per step in the x direction. The diagram shows that the error for both the training and validation sets become smaller as `mtry` approaches the upper bound of 36. The jagged appearance of the lines can be attributed to the influence of randomness caused by the small number of trees in the forest. The diagram also shows that a random forest has better performance than a standard regression tree.

**Table 5.2.** Summary of the RMSE values for the random forest model.

| | **Training set RMSE [l/10 km]** | **Test set RMSE [l/10 km]** |
|---|---|---|
| **1-minute data set** | 0.68 | 1.29 |
| **10-minute data set** | 0.10 | 0.74 |

comparison between kernel function using the selected $\epsilon$ value showed that the radial basis function was best suited with regards to minimizing error on the validation set.

To find good values for $\gamma$ and $C$ a grid search over possible values is carried out using the tune framework of `e1071`. The $C$ value controls the trade-off between complexity of decision rule and frequency of error [25]. The $\gamma$ value defines how far the influence of a single training example reaches, with low values meaning "far" and high values meaning "close" [25]. Initial searches over a random subsample of 4000 data points from the 1-minute data set suggested that a $\gamma$ value near 0.0135 and a

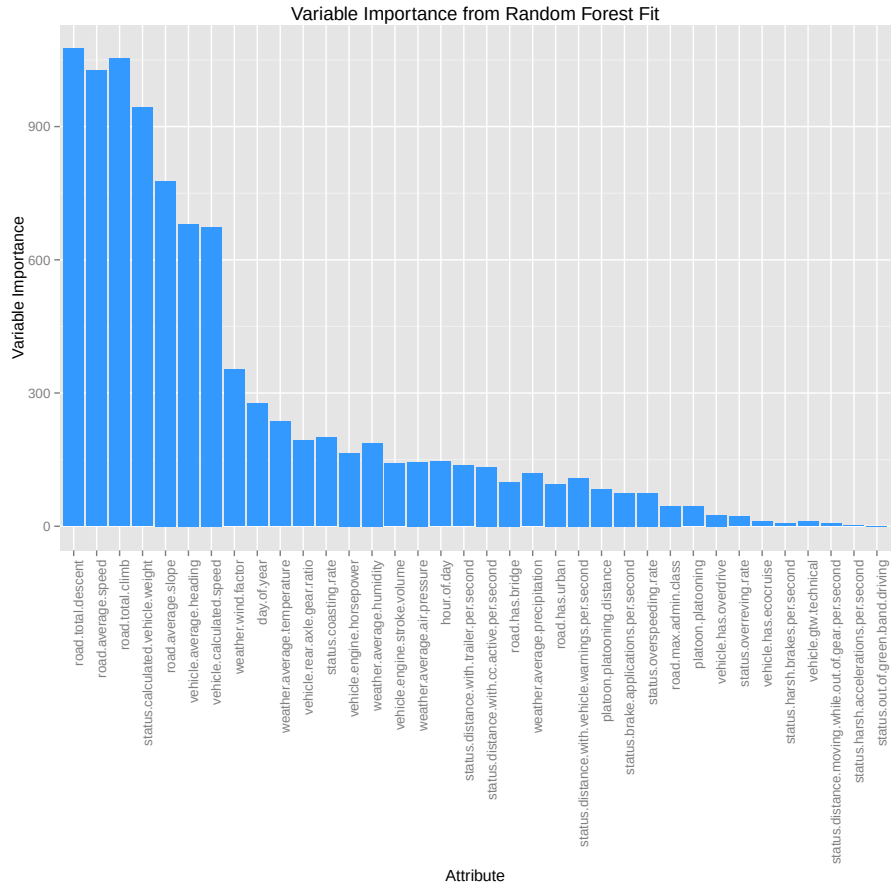


**Figure 5.7.** Chart of variable importance rating found by the random forest fit on the 10-minute data set. The results from the 1-minute data set are similar. The chart shows that the most important features for fuel consumption is the road slope, road speed and the vehicle weight. The next most important type of factors are weather and season related. Driver behavior, platooning and vehicle properties also have a large effect but do not come close in importance to road slope, road speed and vehicle weight.

$C$ value near 12 was a good starting point. Searching over an equal size subsample of the 10-minute data set produced a $\gamma$ value of 0.025 and a $C$ value of 20. The search space was selected by doing an initial coarse grid search over a larger space of values of $C$ and $\gamma$ ranging between $2^{-12}$ and $2^{12}$. Figure 5.9 illustrates the result of the grid search over both data sets. Using these values another more narrow search using a larger sample was performed. The results of the narrow search on the full 10-minute data set indicates that a $\gamma$ value of 0.001 and $C$ value of 18 gives the best results, however only by a very small margin compared to the values found by the search on the subsample. For the 1-minute data set a search over the full sample size is not feasible so the parameters found for the subsamples were used in

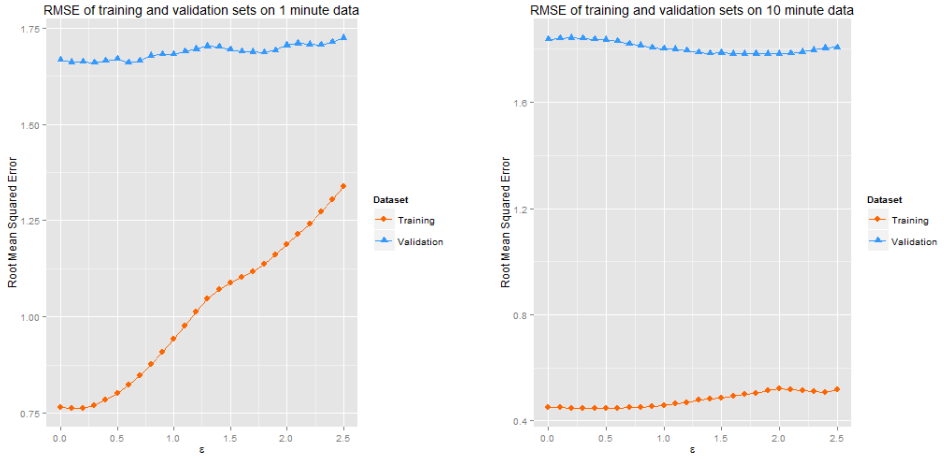


**Figure 5.8.** The influence of $\epsilon$ on the RMSE. The left plot for the 1-minute data set shows that an $\epsilon$ value between 0 and 0.5 is the best fit. The right plot for the 10-minute data set shows that a value between 1.5 and 2 seems to minimize the error on the validation set.

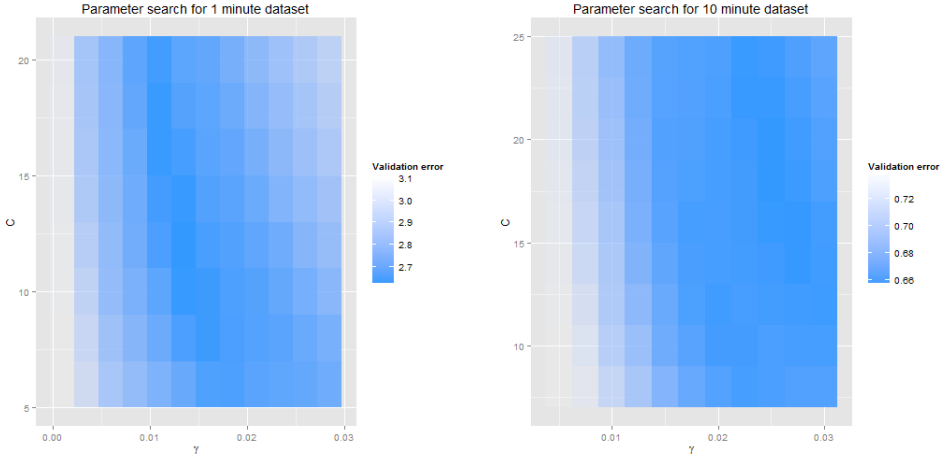


**Figure 5.9.** Validation errors found for different values of $\gamma$ and $C$ for subsamples of both the data sets. The 1-minute data set in the image to the left produced minimal error with $C = 16$ and $\gamma = 0.0125$. The 10-minute data set in the image to the right produced minimal error with $C = 20$ and $\gamma = 0.025$.

the final training.

When fitting the SVR model to the 1-minute data set I subsampled the data to 65 000 observations. This was done due to the very large time requirements for training the model on the entire training set of 650 000 observations. The performance of the final SVR fits on the training and test sets are summarized in Table 5.3. Although it is possible to use SVR to do variable importance rating, the `e1071` R library does

**Table 5.3.** Summary of the RMSE values for the SVR model.

| | **Training set RMSE [l/10 km]** | **Test set RMSE [l/10 km]** |
|---|---|---|
| **1-minute data set** | 0.87 | 1.41 |
| **10-minute data set** | 0.48 | 0.76 |

not include any such methods [31]. Due to this constraint in the software package no variable importance rating was performed using the SVR model.

## 5.6 Artificial neural network

To fit an ANN the R package `neuralnet` was used. `neuralnet` is a package for feed-forward neural networks that implements several variants of backpropagation algorithms that can be configured extensively [22]. I configured the `neuralnet` model as an MLP with a single hidden layer. Figure 5.10 describes how the performance of the ANN varies with the number of nodes in the hidden layer. It seems like the model overfits as the number of hidden nodes increases and that 2 nodes minimizes the error on the validation set.

The ANN was trained with a resilient backpropagation learning algorithm with weight backtracking and a linear function in the output node. The results are summarized in Table 5.4.

## 5.7 Training summary and model comparison

The results of fitting the models to the two data sets consistently shows that the 10-minute data set produces better results. This can be explained by the variance in the 1-minute data set being much higher than in the 10-minute data set, as described by Figure 4.10 and Figure 5.2. Another part of the explanation is that the sampling rate of the driver behavior is closer to the sampling rate of the 10-minute data set and thus more descriptive for the 10-minute data set than for the 1-minute data set.

Figure 5.11 shows the error distributions in percent for the different models fitted to the 1 and 10-minute data sets respectively. For both data sets the random forest model gives the best fit when evaluated on the test data.

Examining the found RMSE values using the 2-way ANOVA statistical test produces results summarized in Table 5.5. The results of a Friedman test applied to the same data is summarized in Table 5.6. The tests investigate how the choice of data set (1 and 10-minute) as well as the choice of model (linear regression, random forest, SVR and ANN) affects the median percent error.

The results of both analyses show that there is a statistically significant difference in error measurements between the two data sets (sampling rates), with a

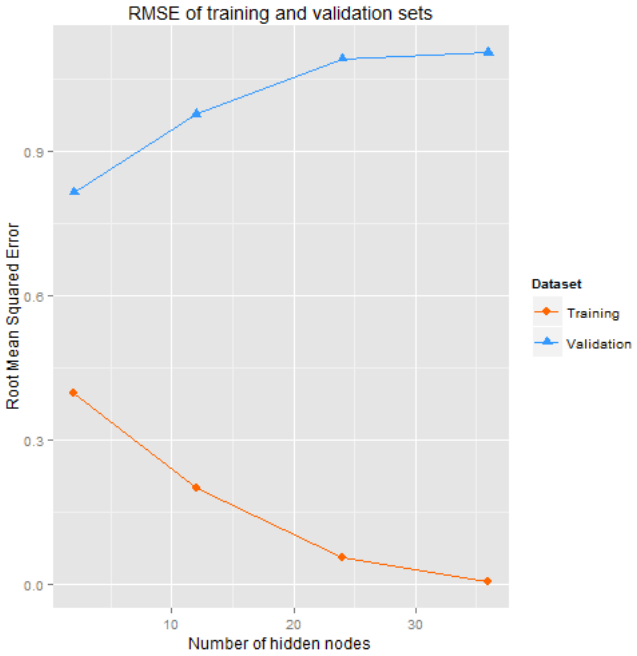


**Figure 5.10.** The error as a function of the number of hidden nodes. It is evident that the training error becomes smaller and that the validation error grows as the number of hidden nodes increases. The training error is nearly 0 when the number of hidden nodes is equal to the number of input nodes. This suggests that the model overfits when the number of hidden nodes is large.

**Table 5.4.** Summary of the RMSE values for the ANN model.

| | **Training set RMSE [l/10 km]** | **Test set RMSE [l/10 km]** |
|---|---|---|
| **1-minute data set** | 1.17 | 1.55 |
| **10-minute data set** | 0.43 | 0.77 |

significance level $\alpha$ of 0.05 (5 %). However, the results also show that there is no statistically significant difference between the different models. For these reasons I must conclude that the sampling rate has a significant effect on the error but that the models are comparable.

## 5.8 Variable importance

To estimate the influence of the different features in the data sets the random forest fitted to the 10-minute data set was trained using subsets of the features. The features were grouped into driver behavior, platooning, road, vehicle and weather

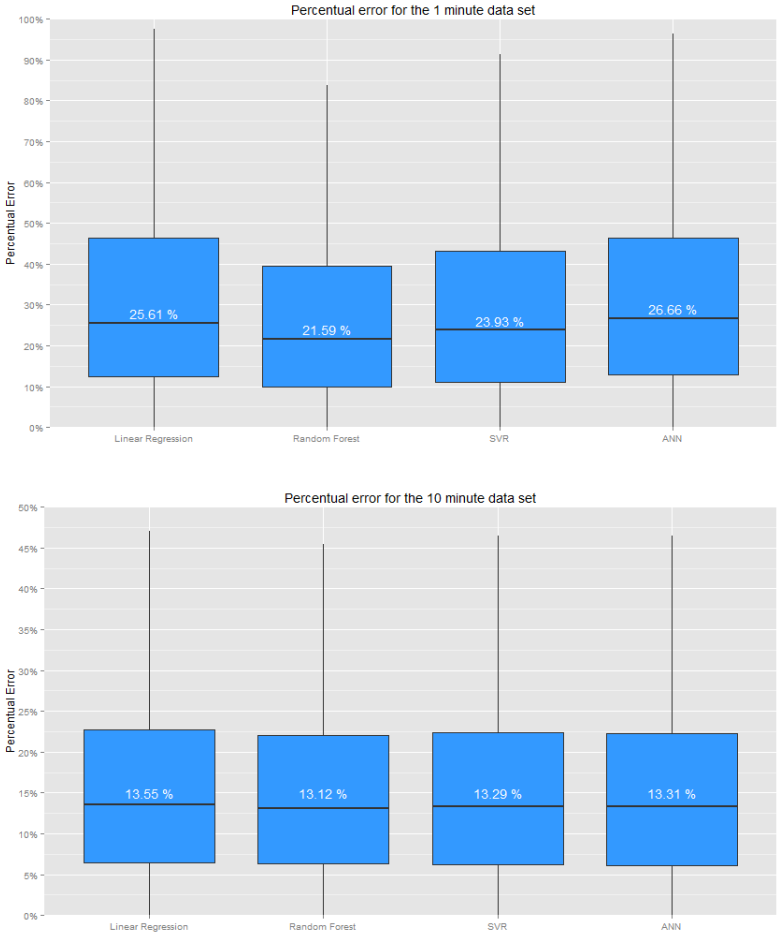


**Figure 5.11.** Comparison of error distributions between the different models fitted to the data. The top image shows the 1-minute data set and the bottom image the 10-minute data set. The diagrams show boxplots of the error measured in percent. The boxplots are divided into 4 quartiles and the horizontal lines and the annotated numbers represent the median of the distributions. The diagrams show that the random forest fit gives the best results on the test set in both cases.

features. Figure 5.12 describes how the prediction error is affected by eliminating each of the feature groups from the data set. It is evident that removing the driver behavior features reduces the error. This indicates that the driver behavior features are of low quality and contain a lot of noise.

To estimate the importance of the individual features the variable importance ratings from the linear model and from the random forest are analyzed. Sets of the ten most important features from the four model fits, one for each data set and model, are compiled and their intersection taken to determine which features are common to all four sets. Table 5.7 shows the importance ratings of the features gen-

**Table 5.5.** Results of a 2-way ANOVA test on the percent error measurements. The "model" row examines the influence of the model choice (linear, random forest, SVR or ANN) on the median percent error. The "data set" row examines the influence of the sampling rate (1 or 10-minute). The P-value is a measure of the statistical significance of the effect of the examined factor. It is clear that the choice of sampling rate has a significant influence while the choice of model does not.

| | DF | Sum of squares | Mean of squares | F-value | P-value |
|---|---|---|---|---|---|
| **Model** | 3 | 8.2 | 2.7 | 1.3 | 0.428 |
| **Data set** | 1 | 247.8 | 247.8 | 113.4 | 0.002 |

**Table 5.6.** Results of a Friedman test on the percent error measurements. The "model" row examines the influence of the model choice (linear, random forest, SVR or ANN) on the median percent error. The "data set" row examines the influence of the sampling rate (1 or 10-minute). The P-value is a measure of the statistical significance of the effect of the examined factor. It is clear that the choice of sampling rate has a significant influence while the choice of model does not.

| | Friedman $\chi^2$ | Degrees of freedom | P-value |
|---|---|---|---|
| **Model** | 5.3 | 3 | 0.145 |
| **Data set** | 4 | 1 | 0.046 |

erated by the different models. The results show that the most important features are related to road slope, vehicle speed and vehicle weight.

## 5.9 Modifying the 10-minute model

Having obtained the initial results it seems worth investigating how the different models will perform when the driver behavior data is excluded from the data set. In order to investigate this I trained new SVR and ANN fits using alternative configurations. The 10-minute data set was chosen for training since it produces the best results.

### 5.9.1 SVR modification

In order to simplify the SVR model training I use a linear kernel instead of a radial basis kernel, reducing the parameter search space and making it easier to find optimal parameters. A grid search over $\epsilon$ and the cost parameter $C$ suggest that an $\epsilon$ value of 1.25 and $C$ value of 11 gives the best results. The search is performed on a subsample of 4000 data points from the training set. In the first iteration a coarse grid search is performed over a large range of values, in the second iteration a finer grid search is performed on a smaller range of values. The results are illustrated in Figure 5.13.

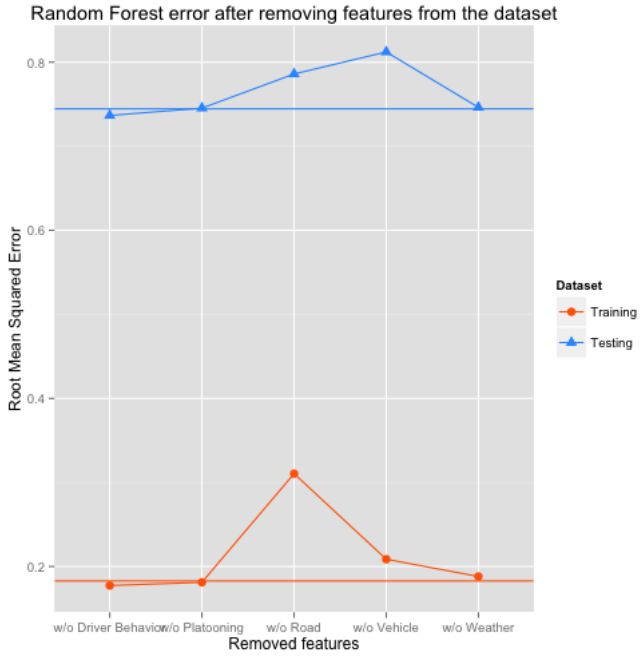


**Figure 5.12.** The change in error in the random forest model when groups of features are excluded from the training set. The horizontal lines indicate the error on the data sets when all features are included. It can be seen that eliminating the driver behavior data actually reduces the error. It can also be seen that eliminating the platooning data has virtually no effect on the error. The other three feature groups however have a larger influence when removed from the data set. This indicates that the quality of the driver behavior and platooning data is low and that they contain quite a bit of noise.

**Table 5.7.** Variable importance ratings extracted from the random forest and linear regression models.

| | **Linear regression 1-minute data set** | **Linear regression 10-minute data set** | **Random forest 1-minute data set [% error increase]** | **Random forest 10-minute data set [% error increase]** |
|---|---|---|---|---|
| road average slope | 0.23 | 0.22 | 0.75 | 0.13 |
| road average speed | 0.24 | 0.07 | 1.06 | 0.17 |
| road has bridge | 0.04 | 0.05 | 0.04 | 0.03 |
| road has urban | 0.13 | 0.03 | 0.12 | 0.02 |
| road max admin class | 0.11 | 0.01 | 0.03 | 0.00 |
| road total climb | 0.24 | 0.02 | 0.61 | 0.22 |
| road total descent | 0.22 | 0.19 | 0.62 | 0.17 |
| status calculated vehicle weight | 0.42 | 0.41 | 0.23 | 0.09 |
| vehicle average heading | 0.06 | 0.01 | 0.73 | 0.13 |
| vehicle calculated speed | 0.37 | 0.19 | 0.20 | 0.04 |

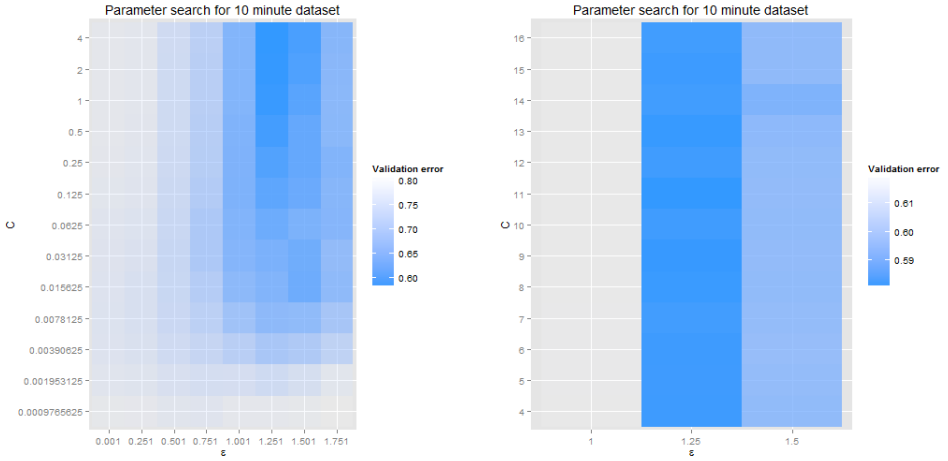


**Figure 5.13.** Illustration of a grid search over $\epsilon$ and $C$ for the SVR model trained on 10-minute data. The left diagram shows the search over a wide range of parameter settings. The right diagram shows the finer search over a smaller parameter range. It is evident that an $\epsilon$ value near 1.25 is ideal, while the $C$ parameter has a much larger range of acceptable settings.

**Table 5.8.** Training results for a linear SVR fit trained on data without the driver behavior features.

| | **Training set RMSE [l/10 km]** | **Test set RMSE [l/10 km]** |
|---|---|---|
| **10-minute data set without driver behavior features** | 0.48 | 0.75 |

Training an SVR model with a linear kernel and the selected parameters produces a better fit than the initial fit using a radial basis kernel, however only by a small margin. After removing the driver behavior parameters from the data set the results were further improved. The training results are summarized in Table 5.8.

### 5.9.2 ANN modification

The ANN model was also refitted without the driver behavior features. However, unlike the random forest and the SVR models the ANN only showed a very small improvement after removing the driver behavior features. The results are summarized in Table 5.9.

**Table 5.9.** Training results for an ANN fit trained on data without the driver behavior features.

| | Training set RMSE [l/10 km] | Test set RMSE [l/10 km] |
|---|---|---|
| **10-minute data set without driver behavior features** | 0.43 | 0.77 |

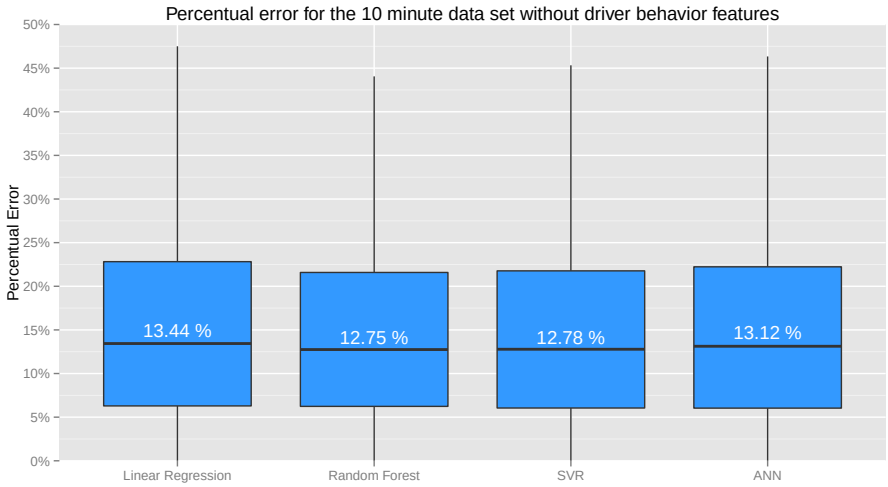


**Figure 5.14.** Error distribution of the final optimized models trained without the driver behavior data. The results show that removing the driver behavior features gives a significant improvement of the prediction. This can be attributed to the driver behavior data containing a lot of noise.

### 5.9.3 Final comparison

Figure 5.14 illustrates the error distributions of the final modified models. The lowest median percent error achieved by the random forest model is 12.75 %. Statistical analysis of how the results are affected by removing the driver behavior features is carried out using 2-way ANOVA and Friedman tests. The results are summarized in Tables 5.10 and 5.11. The statistical tests indicate with a significance level $\alpha = 0.05$ (5 %) that the resulting model fits are improved by removing the driver behavior features.

**Table 5.10.** Results of a 2-way ANOVA test on the percent error measurements. The test compares the regular 10-minute data set with an alternative 10-minute data set in which the features describing driver behavior have been removed. The "model" row examines the influence of the model choice (linear, random forest, SVR or ANN) on the median percent error. The "data set" row examines the influence of the choice of data set (with our without driver behavior features). The P-value is a measure of the statistical significance of the effect of the examined factor. It is clear that the effect of the choice of data set is more significant than the effect of the choice of model.

| | **DF** | **Sum of squares** | **Mean of squares** | **F-value** | **P-value** |
|---|---|---|---|---|---|
| **Model** | 3 | 0.36 | 0.12 | 7.5 | 0.067 |
| **Data set** | 1 | 0.17 | 0.17 | 10.8 | 0.046 |

**Table 5.11.** Results of a Friedman test on the percent error measurements. The test compares the regular 10-minute data set with an alternative 10-minute data set in which the features describing driver behavior have been removed. The "model" row examines the influence of the model choice (linear, random forest, SVR or ANN) on the median percent error. The "data set" row examines the influence of the choice of data set (with our without driver behavior features). The P-value is a measure of the statistical significance of the effect of the examined factor. It is clear that the effect of the choice of data set is more significant than the effect of the choice of model.

| | **Friedman $\chi^2$** | **Degrees of freedom** | **P-value** |
|---|---|---|---|
| **Model** | 6 | 3 | 0.112 |
| **Data set** | 4 | 1 | 0.046 |

# Chapter 6

# Conclusions and discussion

## 6.1 Comparison of sampling rates

The training results consistently shows that a sampling rate of 10 minutes is not only sufficient for prediction, it is actually better than the higher sampling rate of 1 minute. Even though the 1-minute data set contains orders of magnitude more data than the 10-minute data set the results indicate that prediction on the 10-minute data set produces less error. This can be attributed to the larger variance in the 1-minute data set as well as the higher influence of noise on the data with higher sampling rate.

This is however only true when evaluating the models ability to predict fuel consumption in liters per distance. Another problem, which is perhaps more interesting for practical applications, is how the models perform when predicting the total fuel consumed over a longer route. In order to evaluate which sampling rate gives the best results for predicting total fuel consumption an experiment would have to be carried out. Such an experiment could be constructed in the following way:

1. Collect new data from the different data sources describing individual trucks driving different routes, e.g. a route between Helsingborg and Södertälje.

2. For each route, calculate the total fuel consumed by the truck.

3. For each route, use the predictive model to estimate the trucks fuel consumption.

4. Compare the actual total consumption with the predicted total consumption.

This experiment should be carried out using models fitted to both the 1 and 10-minute data sets and the results could be compared using an error measurement in liters. This way the models fitted to the two data sets would be evaluated on their ability to predict actual fuel consumed, and not by how closely they predict a momentaneous value of fuel consumption in liters per distance. However, such an experiment is not within the scope of this thesis.

Another relevant issue when comparing the different sampling rates is that the models trained on the 1-minute data set only used a subsample of the entire data set while the models trained on the 10-minute data used the entire data set. This sub sampling was done in order to limit the time and memory requirements of training the models but may have resulted in a worse fit than if the entire data set had been used.

Further, the parameter search for the SVR model was carried out in two steps; first searching for the $\epsilon$ value and then doing a grid search to find $\gamma$ and $C$. Dividing the search in this way may have caused me to miss the optimal parameter setting as the three variables are dependent. Doing a full grid search in three dimensions would have been better for finding optimal parameter settings. However, it was determined that the time requirements for a three-dimensional parameter search were too large.

## 6.2 Quality of the data

Figure 5.12 shows that the results for the random forest are actually improved by removing the driver behavior data from the data set. Figure 5.12 also shows that the results are not affected by whether the platooning features are included or not. For the other three groups of features however the model shows greater sensitivity.

These results support the belief that the driver behavior and platooning data are of low quality. Viswanathan [5] and Wang [4] have shown that driver behavior has a large influence on fuel consumption and it makes intuitive sense that driver behavior should affect fuel consumption. Hansson [6], Alam et. al [9] and Lammert et. al [10] have shown that platooning has a significant influence on fuel consumption as well. However, my results indicate that the features describing driver behavior and platooning produce a worse or equal fit when they're included in the training data. This suggest that the features contain a lot of noise and are not descriptive of the actual fuel consumption situation. Having driver behavior data with the same sampling rate as the position message data and having reliably calculated platooning data would improve the quality of these features and thus the quality of predictions.

The quality of road and weather data could also be improved. For instance, the accuracy of slope measurements in the road data could be more closely investigated and the locality in both time and space of weather observations could be improved. However, the greatest improvement in prediction quality will be got from improving the quality of the driver behavior and platooning data.

One parameter not included in the data is the age of the vehicles. It would be interesting to investigate if a model trained on data from a brand new vehicle can predict the fuel consumption of an older vehicle with identical configuration.

## 6.3 Usefulness of the trained model

The trained model is most usable for offline purposes, working on historical data of both weather observations and driver behavior. Such a model can be useful for anomaly detection or simulation purposes that work on historical data, but has flaws when applied in a routing context which makes predictions about the future.

Since a routing application requires information about future weather conditions, it would make more sense to train the model using historical weather predictions rather than historical weather observations. To do this one could set up a system that downloads weather prediction data from SMHI or some other weather service. In order to estimate future driver behavior characteristics one could use the historical data to generate driver profiles and use the drivers average behavior as inputs to the predictive model.

As it is, the trained model could still be used as a heuristic in a routing application by using weather predictions in place of weather observations and by assigning mean values for the driver behavior characteristics. However, a model designed specifically to be used in a routing application could potentially give better results.

## 6.4 Relation to previous research

Table 6.1 shows a comparison of the RMSE of the different models trained in this study with corresponding models trained by Svärd in his study [7]. The values come from prediction results on the test sets and are measured in l/10 km. Svärd's results tend to be better on average when compared the results in this study. This can be attributed to his study having a more homogenous data set, with fuel observations being limited to the E4 between Södertälje and Helsingborg [7]. Svärd also considered a smaller number of vehicles and a smaller time span of observations [7]. Since Svärd used different data sources for road and weather data than I did in this study it is possible that he had data of higher quality [7]. Svärd divided his observations into training and test data sets using a random subsampling method instead of a date based subsampling method like I did in this study. Randomly subsampling time series data could lead to high correlation between observations in the different data sets, which in turn could lead to overfitting of the models.

The variable importance rating found in this study correlates well with the variable importance ratings found by Svärd [7] and Lindberg [13]. Both of their studies show that vehicle weight, road slope and vehicle speed rank among the most important features. This result is confirmed by this study.

The previous research into fuel consumption prediction in other applications than heavy vehicles, such as airplanes and engines, showed that ANNs are a suitable model which produces accurate predictions [1, 2, 3]. These results are confirmed by this study which also shows that an ANN model performs well when predicting fuel consumption.

Running of the predictive models is fast, finishing in a matter of milliseconds for

**Table 6.1.** Comparison of my results with those of Svärd. My RMSE values come from models fitted to the 1-minute data set and Svärd's RMSE values also come from models fitted to observations with a 1-minute sampling rate.

| | **RMSE of my models [l/10 km]** | **RMSE of Svärd's models [l/10 km]** |
|---|---|---|
| **Linear regression** | 1.65 | 1.43 |
| **Random forest** | 1.29 | 1.05 |
| **SVR** | 1.41 | 0.99 |
| **ANN** | 1.55 | 1.13 |

a single prediction. Thus, the ML models outperform the simulation based models, described by Sandberg [11], with respect to the speed of calculation.

## 6.5 Recommendations

The single most important action I have identified that could improve the quality of fuel consumption predictions is increasing the sampling rate of the driver behavior data. The driver behavior data in Scania's FM system is currently aggregated and stored with a sampling rate of ca 1 hour. The majority of position messages in Scania's FM system however are collected with a 10 minute sampling rate. If the driver behavior data had the same 10 minute sampling rate it would be more descriptive of the fuel consumption situations and would be of better use in a regression model.

The second most important action I have identified is to develop a reliable method for classifying whether a vehicle is in a platoon or not. Ideally, platooning information should be sent from the vehicles. However, the vehicles currently in production do not have the ability to send this information, thus it has to be calculated from the information we have access to such as the vehicle positions. The method described in this study and in the study by Svärd [7] are theoretically sound but are affected by the noise in GPS position data. This noise propagates into noise in the platooning data, causing the platooning features in this study to be of little use to the regression models. A more reliable system for classifying platooning could be developed with ML methods for classification, provided that labeled platooning data can be compiled for training.

Once the relevant data has been made available with sufficient quality a general fuel consumption model could be trained on the production data with a 10 minute sampling rate.

Another possible development project is to build a system for downloading weather prediction data which can be used in place of weather observation data. Using predictions instead of observations may cause more noise in the data, but it would more accurately describe how the model is supposed to be used in a routing or planning application.

Yet another possible project that could prove useful is to compile driver profiles

using the historical data, and to use driver behavior predictions extracted from such profiles as inputs to a predictive model. Whether to use such driver behavior predictions or to use historical driver behavior observations would depend on what application the model is designed to be used in. Using historical observations would enable building a more accurate model for use in e.g. anomaly detection. Using predictions would produce a model better suited for routing and planning problems that deal with future events.

One could also do analysis of vehicle characteristics, look for correlations and use statistical methods such as PCA to reduce them to a few key metrics for use in future models.

## 6.6 Final conclusions

The results of this study indicate that

- a sampling rate of 10 minutes is better than a sampling rate of 1 minute for predicting fuel consumption measured in liters per distance,
- road slope, vehicle speed and vehicle weight are the most influential parameters for predicting fuel consumption,
- the random forest, SVR and ANN models produce comparable results.

When comparing the models fitted to the 10-minute data set to the models fitted to the 1-minute data set it is clear that the lower sampling rate of 10 minutes produces a smaller error in prediction. The reduction of error is analysed with 2-way ANOVA and a Friedman test and proven to be statistically significant at a 0.05 level. This can be explained by the lower variance of the 10-minute data set and the smaller influence of noise in the data.

The most influential parameters for prediction are shown to be related to road slope, vehicle speed and vehicle weight. These results are confirmed by the studies of Svärd [7] and Lindberg [13].

Out of the random forest, SVR and ANN models no statistically significant differences could be found, so I must conclude that the models are comparable when applied to predicting fuel consumption for heavy vehicles in liters per distance.

# Bibliography


[1] Glenn D. Schilling. Modeling aircraft fuel consumption with a neural network. Master's thesis, Virginia Polytechnic Institute and State University, 1997.

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

[3] Necla Kara Togun and Sedat Baysec. Prediction of torque and specific fuel consumption of a gasoline engine by using artificial neural networks. *Applied Energy*, 87:349–355, 2010.

[4] Haikun Wang, Lixin Fu, Yu Zhou, and He Li. Modelling of the fuel consumption for passenger cars regarding driving characteristics. *Transportation Research Part D: Transport and Environment*, 13(7):479–482, October 2008. 00046.

[5] Aiswaryaa Viswanathan. Data driven analysis of usage and driving parameters that affect fuel consumption of heavy vehicles. Master's thesis, Linköping University, Statistics, The Institute of Technology, 2013.

[6] Karl Hansson. Data-driven analysis of the fuel saving potential of road vehicle platooning. Master's thesis, KTH, School of Computer Science and Communications (CSC), 2013.

[7] Carl Svärd. Predictive modelling of fuel consumption using machine learning techniques. Technical report, Scania CV AB, 2014.

[8] Companion - cooperative dynamic formation of platoons for safe and energy-optimized goods transportation, 2015. [Online; URL: http://www.companion-project.eu/; accessed: 2015-02-13].

[9] Asad Al Alam, Ather Gattami, and Karl Henrik Johansson. An experimental study on the fuel reduction potential of heavy duty vehicle platooning. In *Intelligent Transportation Systems (ITSC), 2010 13th International IEEE Conference on*, pages 306–311, Sept 2010.

[10] Michael P Lammert, Adam Duran, Jeremy Diez, Kevin Burton, and Alex Nicholson. Effect of platooning on fuel consumption of class 8 vehicles over

a range of speeds, following distances, and mass. Technical report, SAE Technical Paper, 2014.

[11] Tony Sandberg. Heavy truck modeling for fuel consumption simulations and measurements. Master's thesis, Linköping University, 2001.

[12] S. H. Nasser, V. WeiBermel, and J. Wiek. Computer simulation of vehicle's performance and fuel consumption under steady and dynamic driving conditions, 1998.

[13] 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.

[14] Thomas M. Mitchell. *Machine Learning.* McGraw-Hill, Inc., New York, NY, USA, 1 edition, 1997.

[15] Christopher D. Manning, Prabhakar Raghavan, and Hinrich Schütze. *Introduction to Information Retrieval.* Cambridge University Press, New York, NY, USA, 2008.

[16] Dennis Wackerly, William Mendenhall, and Richard L. Scheaffer. *Mathematical Statistics with Applications.* Cengage Learning, Boston, MA, USA, 7th edition, 2007.

[17] Stuart Russell and Peter Norvig. *Artificial Intelligence: A Modern Approach.* Prentice Hall Press, Upper Saddle River, NJ, USA, 3rd edition, 2009.

[18] William Mendenhall and Terry Sincich. *A Second Course in Statistics: Regression Analysis.* Prentice Hall, 2003.

[19] Leo Breiman. Random forests. *Machine Learning*, 45(1):5–32, oct 2001.

[20] Frank Rosenblatt. *Principles of Neurodynamics: Perceptrons and the Theory of Brain Mechanisms.* Spartan Books, Washington D.C., USA, 1962.

[21] Martin Riedmiller. Rprop - description and implementation details: Technical report. 1994.

[22] Stefan Fritsch, Frauke Guenther, and following earlier work by Marc Suling. *neuralnet: Training of neural networks*, 2012. R package version 1.32.

[23] Harris Drucker, Christopher J. C. Burges, Linda Kaufman, Alex J. Smola, and Vladimir Vapnik. Support vector regression machines. In *Advances in Neural Information Processing Systems 9, NIPS, Denver, CO, USA, December 2-5, 1996*, pages 155–161, 1996.

[24] Chih-Chung Chang and Chih-Jen Lin. Libsvm: A library for support vector machines. *ACM Trans. Intell. Syst. Technol.*, 2(3):27:1–27:27, May 2011.

[25] Vladimir N. Vapnik. *The Nature of Statistical Learning Theory*. Springer-Verlag New York, Inc., New York, NY, USA, 1995.

[26] Howard J. Seltman. Experimental design and analysis, 2010.

[27] William J. Conover. *Practical nonparametric statistics*. Wiley series in probability and statistics: Applied probability and statistics. Wiley, 1999.

[28] Smhi open data api docs - meteorological observations, 2015. [Online; URL: http://opendata.smhi.se/apidocs/metobs/; accessed: 2015-01-29].

[29] Postgis - spatial and geographic objects for postgresql, 2015. [Online; URL: http://postgis.net/; accessed: 2015-02-02].

[30] Andy Liaw and Matthew Wiener. Classification and regression by randomforest. *R News*, 2(3):18–22, 2002.

[31] David Meyer, Evgenia Dimitriadou, Kurt Hornik, Andreas Weingessel, and Friedrich Leisch. *e1071: Misc Functions of the Department of Statistics (e1071), TU Wien*, 2014. R package version 1.6-4.

[32] John A. K. Suykens. *Advances in Learning Theory: Methods, Models, and Applications*. NATO science series: Computer and Systems Sciences. IOS Press, 2003.