Predictive Modeling of Emissions Heavy Duty Vehicles

Scania CV AB

Max af Klintberg

Student
Master's thesis, 30 hp,
Master of Science in Industrial Engineering
and Management, 300 Credits

Department of Mathematics and Mathematical Statistics
Umeå University
SE-901 87 Umeå, Sweden

Supervisors:
Konrad Abramowicz, Umeå University
Henrik Wentzel, Scania CV AB

Examiner:
Xijia Liu, Umeå University

Abstract

A coming law is approaching all heavy duty vehicle manufacturers operating in the EU area. Apart from meeting all standards of this law Scania wants to be ahead when it comes to every aspect of the market. A big part of this law is the Vehicle Energy Calculation Tool (VECTO) that simulates the emissions of a heavy duty vehicle. Scania wants to investigate if it is possible to implement a prediction model, based on a pre-simulated sub-set of sold Scania vehicles to predict the results from VECTO.

The aim of this thesis is to create a predictive model that can estimate grams of CO_2 per ton and kilometre as simulated in VECTO. The thesis is limited to simulations from the European Automobile Manufacturers Association predefined Long Haulage mission profile. Any explicit performance thresholds are not stated, but the aim is to minimize the estimated test error.

In this thesis statistical learning methods in the form of both parametric and non-parametric regression modelling are implemented. Furthermore, the data given and generated is dealt with and modified pursuing optimal predictive power. It is important to be able to perform predictions as accurately as possible for all vehicles. To minimize the largest prediction errors is the main focus while constructing the model.

The construction of a prediction model seem to be a success, depending on the accuracy requirements set by Scania. The final model predicts grams of CO_2 per ton and kilometre as simulated in VECTO for any given new vehicle in less than a quarter of a second with a prediction error being less than 0.85% for 95% of all vehicles tested.

Sammanfattning

En kommande lagstiftning gällande alla tunga fordon som är verksamma inom EU närmar sig. Bortsett från att uppfylla de krav som lagen kommer att medföra vill Scania ligga i framkant när det gäller alla aspekter av marknaden. En stor del av denna lag är simulerings programmet Vehicle Energy Calculation Tool (VECTO) som simulerar utsläpp för tunga fordon. Det Scania vill undersöka är huruvida det är möjligt att konstruera en prediktionsmodell, baserad på en i förväg simulerad delmängd av Scanias sålda fordon för att förutsäga resultaten från VECTO.

Syftet med denna rapport är att skapa denna prediktionsmodell som ska kunna uppskatta gram CO_2 per ton och kilometer som simulerat i VECTO. Rapporten är begränsad till simuleringar från European Automobile Manufacturers Association's fördefinierade Long Haulage transportuppdragsprofil. Inga förutbestämda prestationskrav finns på modellen men målet är att minimera det största uppskattade testfelet.

De statistiska metoder som tillämpas är i form av både parametrisk och icke-parametrisk regression. All data som samlas in och/eller genereras behandlas och modifieras i syfte att uppnå optimal prediktionsförmåga. Det är viktigt att utföra så exakta prediktionersom möjligt för alla fordon. Att minimera det största prediktionsfelet är då en stor del av huvudfokus när prediktionsmodellen konstrueras.

Av resultatet att dömma är den konstruerade prognosmodellen en framgång, beroende på de noggrannhetskrav som kommer ställas av Scania. Den slutliga modellen predikterar gram CO_2 per ton och kilometer så som simulerat i VECTO för varje givet nytt fordon på mindre än en fjärdedel av en sekund med ett prediktionsfel mindre än 0,85 % för 95 % av alla fordon som testats.

Acknowledgements

I want to express my deepest gratitude to my supervisor Dr. Konrad Abramowicz at Umeå University, you have guided me through setbacks and difficulties without question just encouragement. I also want to thank Allan who has been a supporting friend whenever needed, you are the man.

Contents

Abbreviations

ACEA	European Automobile Manufacturers Association
AUX	Auxiliaries
CD	Drag resistance coefficient
EC	The European Commission
FC	Fuel Consumption
FZ	Tire test load according to ISO 28580 (85% of maxload)
HDV	Heavy Duty Vehicle
IDW	Inverse Distance Weighting
IDW-KNN	Weighted k-Nearest Neighbors Interpolation
KNN	k-Nearest Neighbors
LOOCV	Leave-One-Out Cross-Validation
MLR	Multiple Linear Regression
MT/AMT/AT	Manual Transmission/Automatic Manual Transmission/ Automatic Transmission
RRC	Rolling Resistance Coefficient
VECTO	Vehicle Energy Calculation Tool
WHTC	World Harmonized Transient Cycle

1 Introduction

This thesis is made in collaboration with supervisors and employees on YDMC, technical centre, Scania AB in Södertälje and Umeå University. YDMC is a sub group of Scania's full vehicle testing department at the technical centre in Södertälje. They specialize in full-vehicle analysis regarding fuel consumption and all that it entails. Currently one of their main objectives is to prepare Scania for a CO_2 legislation that will put new requirements on all Heavy Duty Vehicles (HDV) in commercial use in the European Union.

1.1 Background

Fuel efficiency is one of the most important competitive factors in developing and selling HDVs. Therefore, one could say that the same market force encourages continuous progress regarding the reduction of fuel consumption and carbon dioxide (CO_2) emissions. To improve the performance of HDVs, European Automobile Manufacturers Association (ACEA) consider a manufacturer declaration of fuel efficiency. It is seen as the most appropriate way to enforce continuous development of fuel consumption and CO_2 emissions efficiency. The European Commission (EC) together with ACEA are working on a legislation concerning CO_2 certification for HDVs. This will cover a full HDV declaration, most likely grams of fuel and CO_2 emissions per ton and kilometer. Additionally it will cover different ways to correlate the certified CO_2 values to actual CO_2 emissions that are being explored.

Due to the diversity of all HDVs it would be inappropriate to carry out CO_2 testing on HDVs in the same homogeneous way as done for cars and vans. To solve that, the EC in cooperation with industry stakeholders has since 2009 been developing a simulation tool, VECTO, to measure the whole vehicle's CO_2 emissions. VECTO is expected to be the first industry-wide methodology in estimating an entire vehicle's CO_2 emissions, taking not only the engine but also transmission, aerodynamics, rolling resistance, and auxiliaries etc. To run VECTO is time consuming and takes (on an average computer) around one minute to simulate one vehicle. Scania thinks that VECTO will be an important tool seen from a sales perspective, and it is important that it can be incorporated in the sales process efficiently. In a sales situation it must be possible to communicate the CO_2 for a HDV specification faster than VECTO currently manages to deliver.

Scania sees two different solutions. First, the one being investigated, is to use historical sales records and simulate a subset of previously sold HDVs in VECTO. These vehicles will be used to create a discrete data set which is used to build a prediction model. This model would then replace VECTO in this specific matter. The second alternative is to setup a computer park which sole purpose would be to make the VECTO simulation faster, this is the not the desirable alternative.

1.2 VECTO

VECTO is a simulation tool used to approximate both fuel consumption and CO2 emissions from a whole vehicle, based on vehicle specifications and mission profile. The vehicle specifications that VECTO takes as in-data is a job-file that consists of a vehicle-file, an engine-file and a gearbox-file. These contain all the specific in-data parameters

that VECTO requires. A thorough explanation of VECTO can be found in CLIMA (2014).

VECTO takes the mission profile/ driving cycle and divides it in discrete time steps, 1Hz. Acceleration and deceleration are added to the cycle. The next step is to add driving characteristics as eco-roll, over-speed and look-ahead coasting which also are ways to alter acceleration and deceleration behaviors through the simulation for it to be as realistic as possible. After this the power calculation, which can be seen as the core of the VECTO simulation, is initiated. In this stage VECTO calculates a required engine speed and torque. Finally the Fuel Consumption (FC) calculation is initiated, from the engine speed and torque that was determined in the power calculation. The fuel consumption is calculated in three steps:

1. Interpolation of fuel consumption from the Fuel consumption map, this is done by triangulation.
2. Start/stop corrections are made for standstills and starts due to the fact that the consumption of the auxiliaries are not calculated when the vehicle is not moving.
3. World Harmonized Transient Cycle (WHTC) corrections are done due to the fact that the fuel consumption map is measured stationary and the FC is different when exposed to transient engine speed and torque.

After calculating the fuel consumption the CO_2 emissions for the cycle is calculated directly from the fuel consumption through a geometrical factor. The output from VECTO is presented in a number of files. Among these are one illustrative .pdf file presenting the whole simulation and a .vsum file which is more specific and that contains all summarized data from all vehicles simulated.

1.2.1 Job file

The job file collects the data from all the other files run by VECTO. In the Job file the vehicle Auxiliaries (AUX) are directly included. Also the desired mission profile/ driving cycle can be found in the Job file. More specifically the following are found:

1. Vehicle file
2. Engine file
3. Gearbox file
4. AUX
5. Driving Cycle

1.2.2 Vehicle file

The vehicle file contains information about the general vehicle parameters: Vehicle category e.g. rigid, tractor or buss, axle configuration, HDV class and Weight/Loading. Furthermore the following parameters are included in the vehicle file

1.2.3 Air resistance (CdxA)

The air resistance is defined by the product of drag resistance (Cd) and cross sectional area and air density together with the vehicle speed.

1.2.4 Axles/Wheels

For each axle the parameters; relative axle load, Rolling Recistance Coefficient (RRC) and FZ have to be defined in order to calculate the total rolling resistance coefficient. Furthermore the Wheels Inertia has to be set per wheel for each axle, but this is set automatically according to the type of tires selected. The FZ is the tire test load according to ISO 28580 (85% of maxload). The FZ is kept constant in this thesis as the documentation on this value is suffering from qualitative shortcomings. The weight load share (wls) is another required input in VECTO. This information is not included in the vehicle specification and have little effect on the final result in VECTO (Petren, 2014). Rolling resistance on the other hand represents about a third of the energy demand when running an HDV. The RRC for a whole vehicle is calculated as

$$RRC_{vehicle} = \sum_{i=1}^{A} wls \cdot RRC_i \cdot (w_{loading} + w_{vehicle} + w_{massextra}) \cdot wls \cdot (16.64 \cdot FZ)^{-0.1} \quad (1)$$

where A is the number of axles fitted on a vehicle, and $w_{massextra}$ is extra weight on the vehicle. The parameter $w_{massextra}$ is not considered in the legislation and is zero for all vehicles simulated in this thesis.

1.2.5 Engine file

The engine file defines all the engine related parameters and input files, like the Full load curve, a drag curve and the FC map. The Full load curve illustrates the maximum torque the engine can produce at a given engine speed. The drag curve illustrates the minimum torque of the engine. The FC map contains the fuel consumption for the engine in a number of points, given a torque and an engine speed. These points are measured stationary without any transient behaviour incorporated. The WHTC correction factors are included to correct for different types of driving e.g. rural or motorway. The WHTC is an important element of the simulation process. The WHTC correction coefficient is given by a quota between FC from both actual driving in representable cycles and the engine testing. The engine file also contain the main engine parameters:

1. Manufacturer and Model
2. Idling engine speed
3. Displacement
4. Inertia including flywheel, Inertia for rotating parts including engine flywheel.

1.2.6 Gearbox file

The gearbox file defines all the gearbox-related input parameters like gear ratios and transmission loss maps. The main gearbox parameters are

1. Make and model
2. Transmission type, MT/AMT/AT/custom, (this project is limited to focus on MT and AMT).
3. Inertia, rotational inertia of the gearbox (is set constant for all gears)
4. Traction interruption, interruption during gear shift event

1.2.7 Fuel Consumption Map

The FC map is the heart of the VECTO simulation in a sense. This is a discrete data set with recorded FC for a given number of torque/rpm requirements (points). When used in VECTO points in-between these points are interpolated to estimate a corresponding consumption.

1.2.8 Driving Cycles/Mission Profile

The ACEA predefined mission profile that in focus in this thesis is Long Haulage. This cycle is defined for HDVs over 7.5 tone. The Long Haulage cycle is described as delivery to national and international sites. Mainly highway operation and a small share of regional roads. For each cycle run, VECTO performs the simulation three times. Each time with a different loading;

- Empty
- Reference load
- Full load

The reference load is calculated depending on driving cycle, wheel configuration, chassis adaptation and weight. The full load is an input parameter in VECTO for the user to select.

1.3 Previous work

A parameter sensitivity study was done by Petren (2014) to get an answer regarding what affects the FC and the CO_2 emissions the most out of the main features of a HDV. The features analyzed was mainly; air drag, drag loss in the power train, and the FC of the engine. The analysis was done through steady state simulations in VECTO where as the vehicle had the constant speed of 85km/h driving on the highest gear. The Cd-coefficient, the rolling resistance and the FC map of the engine had the most significant effect on the result in VECTO. The Cd-value seemed to behave linear in VECTO with constant change in FC. The Cd-value is assumed to stand for a third of the resistance an HDV has to overcome driving in highest gear in 85km/h. The rolling resistance indicates a 30% impact on the FC, also the RRC seem to have a linear relation to the FC. Regardless of other parameters it was found that the FC map had the greatest influence on the FC. From which it can be concluded that engine FC efficiency has the biggest effect on the simulated FC in VECTO. The study suggested further studies regarding the simulated cycles. Since the mission profile is fixed to the ACEA Long Haulage cycle, changes in-between cycles does not affect the thesis and are overlooked. This Study gives an indication on what parameters in a HDV specification that can be of interest when initiating further analysis.

1.4 Problem specification

As the coming legislation is approaching, Scania along with every other HDV manufacturer operating in the EU area, has to adapt to future circumstances to stay competitive. Apart from meeting all standards of this coming legislation Scania wants to be ahead when it comes to every aspect of the market. So, Scania wishes to investigate if it is

possible to implement a tool, based on a pre-simulated sub-set of sold Scania HDVs to predict results from VECTO.

1.5 Aim

The aim of this thesis is to create a predictive model that can estimate grams of CO_2 per ton and kilometre (gCO_2/tkm) as simulated in VECTO. The thesis is limited to simulations from the ACEA predefined long haulage mission profile. Any explicit performance thresholds are not stated, but the aim is to minimize the estimated test error.

1.6 Approach

In this thesis statistical learning methods in the form of both parametric and non-parametric regression modelling are implemented. Furthermore, the data given and generated is dealt with and modified pursuing optimal predictive power. It is important to be able to perform predictions as accurately as possible for all HDVs. To minimize the largest prediction errors is the main focus while constructing the model. More specifically the largest prediction error in the empirical 95% quantile.

This thesis is divided in three more or less distinct parts. The first one is a learning and preparatory phase where deeper knowledge and understanding about Scania and the forthcoming legislation is acquired. This is followed by data collection. The collected data form the base for subsequent predictive modelling. Whereas each individual HDV in the data-set is simulated in VECTO to generate desired response variable. The third part is the predictive modeling.

The set of vehicles simulated in VECTO are used to build a prediction model. Exploratory models are analyzed and improved continuously. From the exploratory phase the most promising models are chosen for parameter tuning and testing. The best performing model from the tuning phase is choosen as final. The final model is tested to establish the prediction accuracy.

The parametric statistical learning methods used are all types of regression in a sense. Parametric multiple linear regression models are tested, followed by some non-parametric approaches such as Local Regression and KNN regression. The non-parametric models implemented holds favourable characteristics when it comes to fitting statistical learning model to large samples, such as the one approached in this thesis. What is favorable with the non-parametric models is in this case that they do not expect any predetermined form, but is constructed according to information derived from the data.

Finally the results are evaluated and discussed. The different models are evaluated regarding chosen method and results. From this follows final model recommendations and conclusions.

1.6.1 Limitations

In this thesis there are a number of constrains surrounding the data from Scania in regards to what is needed to simulate correct values from VECTO. Furthermore, the version of VECTO used for all simulations is VECTO 2.1.4 which is not the latest update, but due to practical restraints it is decided to use that version.

The final validity and applicability is not certain since the legislators are not finished defining proper guidelines for how the HDVs specification are to be certified and how the final law will appear in practice. This limits the predictive model in how it is set up, the lack of certainty also limits the forming of a fully representative data-set. Hence does this thesis consist of a fair amount of Scania's best guesses.

No designs for a graphical user interface is considered for the model as this is not justified by the aim of the thesis. Also there are so many constraints limiting the applicability of a model that before all the in-data is correct it is not be suited for practical use, yet. Hence, if this type of modeling proves applicable the data set must be run with correct certified data as mentioned, not Scania best guess.

2 Theory

This chapter aims to introduce relevant theory that implemented methods are based upon. section 2.1. Fundamental Theory contains the basic notations that are being used in the thesis. Further, basic theory regarding distance metrics and how categorical variables are handled. A definition of a neighborhood is presented and the theory of Inverse Distance Weighting is described. This is followed by Regression in section 2.2. Regression techniques used in this thesis are Multiple Linear Regression, K Nearest Neighbor Regression (KNN), also modified using Inverse Distance weighing (IDW) and lastly, Local Regression. These modeling techniques are followed by two different cross validation methods, k-fold Cross Validation and Leave One Out cross validation which are used model validation.

2.1 Fundamental Theory

Consider random vector $\mathbb{X} = (X_1, X_2, \ldots, X_n)^T$, where $X_1, \ldots, X_n$ are random variables. The mean vector $\boldsymbol{\mu}$ is defined as

$$\boldsymbol{\mu} := \begin{bmatrix} \mu_1 \\ \mu_2 \\ \vdots \\ \mu_n \end{bmatrix} = \begin{bmatrix} \mathbb{E}(X_1) \\ \mathbb{E}(X_2) \\ \vdots \\ \mathbb{E}(X_n) \end{bmatrix}$$

Similarily, we define covariance $n \times n$ matrix Σ with element in i-th row and j-th vector being

$$\Sigma_{ij} = \mathrm{Cov}(X_i, X_j) = \mathbb{E}((X_i - \mu_i)(X_j - \mu_j)), \qquad i = 1, \ldots, n,\ j = 1, \ldots, n$$

The corresponding definition in the matrix form can be expressed as

$$\Sigma = \mathbb{E}((\mathbb{X} - \boldsymbol{\mu})(\mathbb{X} - \boldsymbol{\mu})^T)$$

with expectation applied element-wise to each of the element of the matrix. The diagonal elements of matrix Σ describe the variances of corresponding random variables and are being denoted by $\sigma_1^2, \ldots, \sigma_n^2$, respectively. As usual, the standard deviation of i-th variable is then denoted by σ_i, $i = 1, \ldots, n$.

2.1.1 Point estimates of mean and covariance

Assume now that we have a sample $\mathscr{X}$ of m observations $\mathbb{x}_i = (x_{1i}, \ldots, x_{ni})$, $i = 1, \ldots, m$ of random vector $\mathbb{X}$. The point estimate of the mean value vector is now

$$\hat{\boldsymbol{\mu}} := \begin{bmatrix} \hat{\mu}_1 \\ \hat{\mu}_2 \\ \vdots \\ \hat{\mu}_n \end{bmatrix} = \begin{bmatrix} \bar{x}_1 \\ \bar{x}_2 \\ \vdots \\ \bar{x}_n \end{bmatrix}$$

with $\bar{x}_i$ being the mean of all the observations along the i-th dimension, i.e.,

$$\bar{x}_i = \frac{1}{m}\sum_{j=1}^{m} x_{ij}, \qquad i = 1, \ldots, n$$

Similarly, the elements of covariance matrix can be estimated using the corresponding sample covariance, i.e.,

$$\hat{\Sigma}_{ij} = \frac{1}{m-1}\sum_{k=1}^{m}(x_{ik} - \hat{\mu}_i)(x_{jk} - \hat{\mu}_j), \qquad i = 1, \ldots, n,\ j = 1, \ldots, n$$

and the corresponding estimated matrix is often denoted by S. Further, the estimated standard deviation and variance for i-th variable are denoted by $\hat{\sigma}$ and $\hat{\sigma}^2$, respectively, for $i = 1, \ldots, n$.

2.1.2 Measures of Distance

If $\mathscr{X} = \{\mathbb{x}_i,\ i = 1, \ldots, m\} \subset \mathbb{R}^n$ is a set of observations, then a function $d(\cdot, \cdot)$ is called a distance on $\mathscr{X}$ if for all $\mathbb{x}_a, \mathbb{x}_b \in \mathscr{X}$, the following holds

1. $d(\mathbb{x}_a, \mathbb{x}_b) \geq 0$ (non-negativity).
2. $d(\mathbb{x}_a, \mathbb{x}_b) = d(\mathbb{x}_b, \mathbb{x}_a)$ (symmetry).
3. $d(\mathbb{x}_a, \mathbb{x}_a) = 0$ (reflexivity).
4. $d(\mathbb{x}_a, \mathbb{x}_c) \leq d(\mathbb{x}_a, \mathbb{x}_b) + d(\mathbb{x}_b, \mathbb{x}_c)$ (triangle inequality)

In Section 2.1.2-2.1.2 we now describe four metrics used in this thesis.

Manhattan distance metric. The Manhattan distance metric uses a form of geometry in which $d_{MH}(\cdot, \cdot)$ is the sum of the absolute differences of their coordinates. The Manhattan distance between $\mathbb{x}_a$ and $\mathbb{x}_b$ is defined as

$$d_{MH}(\mathbb{x}_a, \mathbb{x}_b) = \sum_{j=1}^{n} |x_{a,j} - x_{b,j}|. \tag{2}$$

Euclidean distance metric. The euclidean distance $d_E(\cdot, \cdot)$ is given by the Pythagorean formula. The euclidean distance between $\mathbb{x}_a$ and $\mathbb{x}_b$ is defied as

$$d_E(\mathbb{x}_a, \mathbb{x}_b) = \sqrt{\sum_{j=1}^{n}(x_{a,j} - x_{b,j})^2}. \tag{3}$$

Standardized Euclidean distance metric. The euclidean distance can be further developed by using standardization. To obtain the standardized euclidean distance $d_{SE}(\cdot, \cdot)$, every observation $\mathbb{x}_i = (x_{i,1}, \ldots, x_{i,n})$ in the observed set $\mathscr{X}$ is standardized. The standardized euclidean distance is the defined as

$$d_{SE}(\mathbb{x}_a, \mathbb{x}_b) = \sqrt{\sum_{j=1}^{n}\left(\frac{x_{a,j} - \widehat{\mu}_j}{\widehat{\sigma}_j} - \frac{x_{b,j} - \widehat{\mu}_j}{\widehat{\sigma}_j}\right)^2} \tag{4}$$

where $\widehat{\mu}_j$ and $\widehat{\sigma}_j$ are sample mean and sample standard deviation of the j-th coordinate calculated on the observations in set $\mathscr{X}$.

Mahalanobis distance metric. The Mahalanobis distance is defined as

$$d_{Mah}(\mathbb{x}_a, \mathbb{x}_b) = \sqrt{(\mathbb{x}_a - \mathbb{x}_b)S^{-1}(\mathbb{x}_a - \mathbb{x}_b)^T} \tag{5}$$

where S^{-1} is the inverse of the point estimate of the covariance matrix based on the observed set. The Mahalanobis distance metric is a multidimensional generalization of the idea of measuring how far apart two observations e.g. $\mathbb{x}_a$ and $\mathbb{x}_b$ are in terms of standard deviations.

2.1.3 Categorical variables

A categorical variable is a variable that can take one of a finite number of values, which can be referred to as groups or levels. The levels of a categorical variable do not need to have any particular order amongst the levels.

Dummy coding. There are a number of ways used in analysis of categorical variables. The one being used in this thesis is so called *Dummy coding*. A dummy variable is binary, it takes the value 0 or 1 to indicate the absence or presence of a specific categorical level effect. When using dummy coding to substitute categorical variables there is one dummy variable for each level of the categorical variable. If x is categorical with l levels then x is described in terms of dummy variables as

$$x = (D_{x,1}, D_{x,2}, \cdots, D_{x,l-1})$$

where $D_{x,\cdot}$ takes the value 1 if that specific levels effect is present and 0 otherwise. The number of dummy variables needed is l-1. By using this approach to categorical variables they are treated as numerical and can be incorporated in e.g. regression modeling, but this comes with the cost of dimensionality.

Penalized Distances for categorical covariates. Introducing distance for categorical variables is in general hard and not unique procedure. In this thesis, we use the simple modification of existing metrics by incorporating penalty. Consider two observation vectors consisting of continuous and categorical variables. For each of the categorical variable we compare the values in two vectors. If the values are the same the penalty is set to zero and if they are different a penalty is set to a covariate specific constant. Finally the distance between the two observations is calculated as the sum of the penalties and the distance (calculated as in 2.1.2.1-2.1.2.4) between the continuous part of the vector. The choice of covariate specific penalty is an element which affects this type of distance significantly, however proper prior studies allow determining a value specific to given applications.

For further studies about the distance for categorical variables, we refer to Boriah et al. (2008).

2.1.4 Neighborhood

Consider set $\mathscr{X}$ and a new point $\mathbb{x}_0$. The neighborhood $N_{\mathscr{X}}(\mathbb{x}_0, k, d)$ is a set of points from any point in $\mathscr{X}$ that includes the k nearest neighbours of point $\mathbb{x}_0$ using distances measured with help of an appropriate distance function $d_{(\cdot)}(\cdot, \cdot)$.

2.1.5 Inverse Distance Weighting

Recall that $\mathscr{X} = \{\mathbb{x}_i,\ i = 1, \ldots, m\}$ is a set of n-dimensional observations and $\mathbb{x}_0$ is a new observation. Assuming that each observation $\mathbb{x}_i$ has a local influence that diminish with increased distance the Inverse Distance Weighting function can be introduced. The IDW function assign weights in relation to the distance $d_{(\cdot)}(\mathbb{x}_0, \mathbb{x}_i)$ between $\mathbb{x}_0$ and all $\mathbb{x}_i$ in $\mathscr{X}$. Following the reasoning in Shepard (1968) the IDW function is defined as

$$w(\mathbb{x}_0, \mathbb{x}_i) = \begin{cases} \frac{d(\mathbb{x}_0, \mathbb{x}_i)^{-u}}{\sum_{i=1}^{m} d(\mathbb{x}_0, \mathbb{x}_i)} & \text{if } d(\mathbb{x}_0, \mathbb{x}_i) \neq 0 \text{ for all } \mathbb{x}_i \\ 1 & \text{if } d(\mathbb{x}_0, \mathbb{x}_i) = 0 \text{ for some } \mathbb{x}_i \end{cases} \tag{6}$$

and the parameter $u > 0$. The effect of distance on weight is dependent on parameter u, a larger u magnify the effect and vice versa. If $u = 0$ the weights are equally distributed among all observations in $\mathscr{X}$.Observe further that

$$\sum_{i=1}^{m} w(\mathbb{x}_0, \mathbb{x}_i) = 1. \tag{7}$$

2.2 Regression

Regression analysis is a set of techniques used to model and analyze several variables with focus on the relationship between a dependent response variable y and independent explanatory variables $x_1, ..., x_n$. There are a number of different regression techniques and in this section a few of them are described. More precisely, Multiple Linear Regression (MLR), k-nearest neighbors regression (KNN) and Local Regression. Among these the three the first, MLR, is parametric which assumed that sample data comes from a population that follows a probability distribution based on a fixed set of parameters (Geisser and Johnson, 2006). The KNN is non parametric which basically implies that assumptions about the origin of the data are made.The Local Regression can be seen as a semi parametric and is a hybrid method of the MLR and the KNN regression.

Throughout Section 2.2 we assume that we have m observations from a set $\mathscr{V}$ consisting of $n + 1$-dimensional vectors $(\mathbb{x}_i, y_i)$ where $\mathbb{x}_i \in \mathbb{R}^n$, $i = 1, ..., m$ and $y_i \in \mathbb{R}$ are vectors of explanatory variables and corresponding response variable, where $\mathbb{x}_i = (x_{i1}, x_{i2}, \cdots, x_{in})$ is a vector of parameters of the i-th observation with n explanatory variables. $\mathscr{X}$ is the set of observed explanatory variables and $\mathscr{Y}$ the corresponding set of observed response variables. Further is a new observation denoted as $\mathbb{x}_0 \in \mathscr{X}_0$ and predictions of the response denoted $\hat{y}_0$.

2.2.1 Multiple Linear Regression

Assume that y_i is observations of Y_i. A multiple linear regression model is described as

$$Y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_n x_{in} + \varepsilon_i, \quad i = 1, \ldots, n \tag{8}$$

where Y_i is the prediction of the observation y_i and ε_i are independent identically distributed r.v's with $E(\varepsilon_i) = 0$ and $V(\varepsilon_i) = \sigma^2$. Assuming we have observed m pairs $(\mathbb{x}_i, y_i)$, using matrix notation the design matrix is defined as

$$\mathbf{X} = \begin{bmatrix} \mathbb{x}_1^{\mathrm{T}} \\ \mathbb{x}_2^{\mathrm{T}} \\ \vdots \\ \mathbb{x}_m^{\mathrm{T}} \end{bmatrix} = \begin{bmatrix} x_{11} & \cdots & x_{1n} \\ x_{21} & \cdots & x_{2n} \\ \vdots & \ddots & \vdots \\ x_{m1} & \cdots & x_{mn} \end{bmatrix}.$$

By letting $\mathbf{y} = (y_1, ..., y_m)^T$,$\boldsymbol{\varepsilon} = (\varepsilon_1, ..., \varepsilon_m)^T$ and $\boldsymbol{\beta} = (\beta_1, ..., \beta_m)^T$ then the multiple linear regression model can be written as

$$\mathbf{y} = \boldsymbol{X\beta} + \boldsymbol{\varepsilon} \tag{9}$$

where

$$\mathbf{y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_m \end{bmatrix} \mathbf{X} = \begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1n} \\ x_{21} & x_{22} & \cdots & x_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ x_{m1} & x_{m2} & \cdots & x_{mn} \end{bmatrix} \boldsymbol{\beta} = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_m \end{bmatrix} \boldsymbol{\varepsilon} = \begin{bmatrix} \varepsilon_o \\ \varepsilon_1 \\ \vdots \\ \varepsilon_n \end{bmatrix}$$

If matrix $\mathbf{X}^T\mathbf{X}$ is invertible then $\hat{\boldsymbol{\beta}}$ describes the least square estimator of the $\boldsymbol{\beta}$

$$\hat{\boldsymbol{\beta}} = \left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T\mathbf{y}.$$

The errors of a fitted model can then be calculated as

$$\hat{\mathbf{y}} = \mathbf{X}\hat{\boldsymbol{\beta}} \implies \mathbf{e} = \mathbf{y} - \hat{\mathbf{y}} = \mathbf{y} - X\hat{\boldsymbol{\beta}}. \tag{10}$$

From Equation 10 a prediction of a new observation $\mathbb{x}_0$ is given by

$$\hat{y}_0 = \mathbb{x}_0\,\hat{\boldsymbol{\beta}} \tag{11}$$

2.2.2 KNN Regression

The KNN is a non-parametric regression method dependent on the distance $d(\cdot, \cdot)$ between observations in $\mathscr{X}$. This method involves a tuning parameter k which is the number of observations $(\mathbb{x}_i, y_i)$ that are included in the nearest neighborhood $N_{\mathscr{X}}(\mathbb{x}_0, k, d)$.

The KNN regression procedure use averaging to predict $\hat{y}_0$. This averaging is done within the nearest neighborhood $N_{\mathscr{X}}(\mathbb{x}_0, k, d)$. A prediction for a new observation at point $\mathbb{x}_0$ using KKN is estimated as

$$\hat{y}_0 = \sum_{i=1}^{k} \frac{1}{k} y_i \tag{12}$$

where all k observed points y_i are from the neighborhood $N_{\mathscr{X}}(\mathbb{x}_0, k, d)$.

Weighted KNN Interpolation (IDW-KNN). To modify the KNN regression, weights can be introduced in the prediction process. Instead of predicting using the mean value of the k nearest neighbors as in Equation 12 one can introduce IDW. By applying the weights from Equation 6 prediction of $\hat{y}_0$ for new observation $\mathbb{x}_0$ using weighted KNN regression is defined as

$$\hat{y}_0 = \sum_{i=1}^{k} w(\mathbb{x}_0, \mathbb{x}_i) \cdot y_i \tag{13}$$

where all responses y_i corresponds to the responses within the neighborhood $N_{\mathscr{X}}(\mathbb{x}_0, k, d)$.

2.2.3 Local Regression

Local regression is a non-parametric regression method for fitting non-linear functions and computing the fit at a target point (Wasserman, 2006). This is done by using a hybrid of the KNN and the multiple linear regression. The procedure is described by

1. Determine the neighborhood $N_{\mathscr{X}}(\mathbb{x}_0, k, d)$. See Section 2.1.4.
2. Fit multiple linear regression model to all observations in $N_{\mathscr{X}}(\mathbb{x}_0, k, d)$ and predict $\hat{y}_0$ using Equation 11.

2.2.4 Model Parameters

Described methods are affected by the choice of parameters. Depending on the application optimal parameter structures will vary. In Table 1 are the optional parameters for each model described.

Table 1: The table describes model, model predictor and model parameters

Model	Predictors	Parameters
MLR	$\mathscr{X}$	
KNN	$\mathscr{X}$	k, $d_{(\cdot)}(\cdot,\cdot)$
IDW-KNN	$\mathscr{X}$	k, u, $d_{(\cdot)}(\cdot,\cdot)$
Local Regression	$\mathscr{X}$	k, $d_{(\cdot)}(\cdot,\cdot)$

2.3 Model Performance Measures

Given a set $\mathscr{X}_0$ of some given number of observations m_0 observed $n+1$-dimensional vectors $(\mathbb{x}_i, y_i)$ and m corresponding predictions $\hat{y}_i$ the following performance measures are defined as follows in section 2.3.1-2.3.2.Consider the following abosolute and relative errors

$$e_i = \hat{y}_{0i} - y_{0i}, \quad i = 1, \ldots, n$$
$$e_i^{rel} = \left(\frac{\hat{y}_{0i} - y_{0i}}{y_{0i}} \right) \quad i = 1, \ldots, n.$$

2.3.1 Mean Square Error and Root Mean square Error

The Mean Square Error (MSE) is defined as

$$MSE = \frac{1}{m_0}\sum_{i=1}^{m_0}(e_i)^2$$

where $\hat{y}_i$ is the prediction of y_i. The Root Mean Square Error is defined as

$$RMSE = \sqrt{MSE}.$$

Sometimes relative performance measure are requested so the MSE_{Rel} and RMSE_{Rel} introduced as

$$\text{MSE}_{\text{rel}} = \frac{1}{m_0}\sum_{i=1}^{m_0}(e_i^{rel})^2$$

and

$$RMSE_{Rel} = \sqrt{MSE_{Rel}}.$$

2.3.2 Quantile Performance

In practice not only the mean performance might be of interest when comparing the model success. It is often of interest to investigate the behaviour of extreme errors in the prediction process. Now by finding the empirical quantiles of both samples we can get an insight in the error distributional properties. For example, the empirical quantile of level 95% correspond to a treshold of the error for the 95% of the individuals. Therefore, we know that 95% of the predicted vehicles had an error smaller then this specific value.

2.4 Validation

To estimate the test error rate a number of techniques can be applied by using the available training data. These are methods that estimate the test error by excluding a set of observations from the training data when applying chosen statistical learning method to the training data during the fitting process. Then by using the withheld set of observations as test observations Cross-Validation can be performed. This method of using a set of observations can also be used as a method for fine tuning of model parameters in a statistical learning model. In that regard, $\frac{1}{K}$ of the observations are excluded as test set while the remaining $\frac{K-1}{K}$ observations are used as training set for the fine tuning or optimization process. As with most techniques discussed in this thesis there are numerous ways to apply different tools and that goes for cross validation too. I have decided to rely on two particular techniques which are k-fold cross validation and Leave One Out cross validation.

2.4.1 K-Fold Cross-Validation

The validation technique k-fold cross-validation is an approach that incorporates randomness by dividing the observations in K random groups or folds of almost equal size. One of the folds is treated as test data and the remaining $K-1$ groups are used as training data for fitting of the model. The mean square error, MSE_k, is then computed on the observations in the test fold excluded. This procedure is then repeated K times so

that each group is used as a test set once and training set $K-1$ times. This procedure results in K different training procedures of the model and K estimates of the test group model error,$MSE_1, MSE_2, ..., MSE_K$ where $k = 1, 2, \cdots, K$. The K-fold Cross-Validation is estimate is given by (James et al., 2013)

$$CV_{(K)} = \frac{1}{K} \sum_{k=1}^{K} MSE_k$$

Leave-One-Out Cross-Validation. The Leave One Out Cross Validation (LOOCV) is a special case of the k-fold cross validation where the number of folds are equal to the number of observations. When applying LOOCV the random factor is included in the k-fold cross validation is eliminated.

This approach has almost as little bias as possible whereas almost the whole data set is used training the model. It also tends not to overestimate the test error rate. There is no randomness in choosing the training observations and the test observation, because all combinations are evaluated the same result is obtained if its run multiple times. The down side with this method is that it can be time consuming, if each individual model takes time to train and/or if is large.

2.4.2 Cross-Validation Bias-Variance trade-off

An important advantage of the k-fold cross validation over the LOOCV is that it often gives more accurate estimates of the test error than the LOOCV does due to a bias-variance trade-off. When it comes to bias and bias reduction the LOOCV is to be preferred over k-fold cross validation as it uses way more of the observations for the training set than the k-fold cross validation. But, bias is not the only problem when working with estimations, the variance must be considered as well. Because when the LOOCV is performed, n models are trained on very similar training sets the variance of the LOOCV is higher than that of the k-fol crossvalidation. The test error estimate resulting from k-fold cross validation tend to generate a lower variance than that of the LOOCV method. In regards of choosing an optimal K, James et al. (2013) state that empirical studies have shown that $k = 5$ or $k = 10$ gives test error rates without both bias and variance being higher than necessary.

3 Data

The data used in this thesis is derived both from Scania sources and simulated in VECTO. The data collected comes with limitations, which are described below together with how these limitations are handled. Further are the observed data reduced and transformed in order to crate a solid and representative foundation for regression modeling as possible.

3.1 Scania Data Sources

Due to confidentiality reasons, information on PIDAT and other Scania data sources can not be described in further detail in this thesis.

3.2 VECTO

The simulated data set from VECTO contains numerous output parameters describing the performance of the vehicle simulated. The one used in this thesis is the amount of CO_2 that a vehicle emits in relation to it's weight and the distance travelled, $[gCO_2/tkm]$.

3.3 Sub sets

The data used in this thesis is based on subsets rather than all data available. The reasoning behind this is foremost that gathering data with as high relevance as possible affect the accuracy of a prediction model.

Further selections are made due to constraints embedded within VECTO. Because everything surrounding the coming law is yet to be established regarding certification of components and vehicle parameters, some of these are based on Scania's best guesses.

To limit the studied data subset of vehicles they are filtered out in regards to there specifications. The set consists of HDV's, both rigid trucks and tractors. The wheel configurations are limited to 4x2 and all types of 6x2 vehicles. The vehicles studied are exclusively vehicles sold in the European Union, Norway and Switzerland, in 2015. Engines are another constraint that limit the data set. Only vehicles with the Euro 6 type of engines are considered.

3.3.1 Constraints in VECTO

When it comes to VECTO there are also few constraints restricting the choice of data. The type of gearboxes that are supported in VECTO is manual and automated manual, currently automatic gearboxes are not supported. For this reason the automatic gearboxes are excluded from analysis in this thesis. Also when it comes to the wheels on vehicles simulated, there are still constraints in regards of which dimensions that can be handled.

3.4 Data construction

The first step to create a data set to build desired predictive model is to collect a set of vehicle specifications representative for the purpose. This data set is constructed as illustrated in Table 2. In the table each row corresponds to a vehicle and each column corresponds to a variable. The first column contains the unique chassis number for each vehicle so that they can be recognized and backtracked as needed. The last column is left

empty to be filled with gCO_2/tkm after simulated in VECTO. This data set constitute the foundation of the prediction model.

Table 2: The table shows a fraction of how observed vehicles are described prior to simulation in VECTO.

chassis nr.	Country	Rear axle	Engine	Gearbox	...	gCO_2/tkm
2097563	Switzerland	R780	DC13 125	GRS905R	...	-
2105568	Sweden	R780	DC13 115	GRS905	...	-
3107771	Sweden	R660	DC09 113	GRS895	...	-
⋮	⋮	⋮	⋮	⋮	⋮	⋮
7107579	Great Britain	R660	DC09 108	GR875	...	-

As there is no way to take a single vehicle specification and directly use it as in-data in VECTO, a standardized way to do this must be implemented as a first step. As described in Section 1.2, the in-data required by VECTO consists of a set of files describing a vehicle with all its components and specifics that are required to full fill the simulation process. There are basically four file-structures needed; The Job file containing the Vehicle file, the Engine file and the Gearbox file. All these need to be manipulated in some way to run VECTO. Since Scania's present vehicle specifications do not meet VECTO requirements some variables has to be manipulated and/or redefined. As previously stated, a first step to meet the aim of this thesis is to construct a way to translate Scania's vehicle specifications to VECTO in-data, and fill the holes where required data is missing or not sufficient. Table 3 illustrates the differences between the VECTO input structure and the available vehicle specification. Note that due to previously stated reasons some parameters are set fix for all observed vehicles and marked "-" in Table 3. Presently there are a number of variables that have to be constructed to fit the VECTO in-data structure, foremost RRC for all wheel axles on the HDV and $CDxA$.

Regarding the RRC, these values are collected from a separate document provided by the tire suppliers. As there are no direct connection between Scania's vehicle specification and the documentation from the tire suppliers another way to distinguish which exact tire fitted on each vehicle have to be constructed.

The $CDxA$ is not included in the vehicle specification presently. The legislators in close correspondence with the HDV manufactures are trying to come up with a way to solve the question of how the certification of $CDxA$ is going to appear. Due to the situation this is resolved by using Scania's best guess. That means that the $CDxA$ value used in the modeling are individually simulated for each vehicle and added to the set as a numerical continuous variable.

When the law is instated there will be classifications regarding the air drag of a vehicle, but these are not yet remotely finalized and the methods for measuring this is not yet established either. The air drag accounts for a major part of the power demand running HDV's, and so does the rolling resistance.

Regarding the FC map and the loss maps needed the legislative process is not complete. Hence are Scania's best guess sufficient the best option. Further, in the gearbox loss map received from Scania more points are added trough interpolation since VECTO demand higher resolution. Some of the parameters used simulating the vehicles in VECTO

deviate from what is actually required as input in VECTO. These simplifications are made due to the absence of clear guidelines from the legislators. A few simplifications are also made because acquiring the correct values is not a priority and due to time constraints.

Table 3: Displays the Table showing VECTO input parameters and Scania's corresponding variables found in vehicle specifications and other various documentation.

VECTO	**Scania**	**Notes**
FC map	Engine	Scania best guess
Displacement	Engine	
Transmission type	Opticruise	Determine if MT or AMT. All AT are excluded
Gear ratios	Gearbox	
Gear loss maps	Gearbox	Scania best guess
Rear axle ratio	Axle type	
Rear axle loss map	Rear axle file	Scania best guess
Wheels	Wheel dimension	
$\text{RRC}_{front\,axle}$	$\text{RRC}_{front\,axle}$	Calculated from separate documentation
$\text{RRC}_{driving\,axle}$	$\text{RRC}_{driving\,axle}$	Calculated from separate documentation
$\text{RRC}_{extra\,axle}$	$\text{RRC}_{extra\,axle}$	Calculated from separate documentation (if given vehicle have three axles)
CDxA	CDxA	Scania best guess
Curb Weight	Weight	
Axle configuration type	Configuration	
Rear axle ratio	Rear axle ratio	
Auxiliaries	-	Fixed in this thesis
Vehicle category	chassis adaptation	
Rolling-circumference	Rolling-circumference	

3.5 Simulation in VECTO

The constructed data set is run in VECTO to create a dependent response. VECTO returns numerous output variables and among those gCO_2/tkm from the reference load simulation is the one used as response variable representing VECTO. This output is most likely the one that is of significance regarding the coming law.

3.6 Predictor Selection

After collecting relevant output we start building a predictive model. What first needs to be addressed is the number of explanatory variables that are used in the prediction process. The initial variable choice is done based on the specific in-data VECTO requires, as this was what we used to generate the response variable. These variables are described in Table 3. Prior to the prediction modeling we decided to transform two explanatory variables possible to resemble those in VECTO. The explanatory variables transformed are

- RRC. The axle wise RRC values calculated is transformed in to a full vehicle RRC in combination with wheel configuration weight load share between axles and the FZ value. From Equation 1 in Section 1.2.4

$$RRC_{vehicle} = \sum_{i=1}^{A} wls \cdot RRC_i \cdot (w_{loading} + w_{vehicle} + w_{massextra}) \cdot wls \cdot (16.64 \cdot FZ)^{-0.1}.$$

  where

  - A is the number of axles on each vehicle
  - wls is not specified in the vehicle specification from Scania and is thus equaly distributed between the number of axles on each vehicle
  - $w_{loading}$ is set to the vehicle reference load
  - $w_{vehicle}$ is set to the vehicle weight
  - $w_{massextra}$ is set to zero, since this is not included in the legislation

- The cruising-rev can be described as the engine speed required at a given vehicle speed in relation to the ratio on the highest gear in the gearbox, the axle ratio and the wheel rolling-circumference. The transformed explanatory variable C_{rpm} is calculated as

$$C_{rpm} = (Axle_{ratio} \cdot GBX_{ratio})/(RollCirc).$$

  where

  - $Axle_{ratio}$ is the rear axle ratio as given in the vehicle specification
  - GBX_{ratio} is the ratio on the highest gear in the gearbox as given in the gearbox specification
  - $RollCirc$ is the rolling-circumference as given in the vehicle specification

These variables contain information from several separate variables, both categorical and continuous in two continuous variables.

3.6.1 Variable set

The set of explanatory variables and the corresponding reference variable is the described in Table 4. These remain the same through out all model testing. If so, it is clearly stated in the model description.

Table 4: This table show the set of all explanatory variables that are used in the modeling process.

Variable Name	**Type**	**Number of Levels**	**Description**
$RRC_{vehicle}$	Continuous	-	Calculated as described in Equation 1
CDxA	Continuous	-	Calculated as the vehicle individual CD value times the front sectional area og the vehicle
Cruising-rev (C_{rpm})	Continuous	-	Calculated as combination of rolling circumference on the driving axle, axle gear ratio and ratio on the highest gear
Weight	Continuous	-	The weight of the vehicle as given in the vehicle specification
Engine Displacement	Continuous	-	The displacement of the engine as described in the engine specification
Engine	Categorical	13	Type of engine
Gearbox	Categorical	12	Type of gearbox
Rear Axle	Categorical	7	Type of rear axle
Wheel configuration	Categorical	5	The number of wheel axles including which ones are driving and steering
chassis Adaptation	Categorical	2	States if the HDV is a truck or a tractor
Respons variables	**Type**	**Number of Levels**	**Description**
y_i [gCO_2/tkm]	Continuous response	-	Simulated in VECTO

4 Method

To meet the aim of the thesis and build an accurate predictive model a few different types of regression are applied to observed vehicles and analyzed. Initially a heuristic exploratory model training phase is conducted to learn as much as possible about the relation between covariates and the response. The best performing of these models are further analyzed and implemented as a final model which is tested to get the final result. Further, all modeling referring to multiple linear regression in this thesis is done without any interaction terms or higher order terms.

4.1 General Model Structure

All models are based on a set $\mathscr{V}$ of $m = 28972$ observed vehicles with corresponding VECTO output. Formally we define $\mathscr{V} = \{(\mathbb{x}_i, y_i),\ i = 1, \ldots, m\}$ the corresponding set of the explanatory variables is $\mathscr{X} = \{\mathbb{x}_i,\ i = 1, \ldots, m\}$ and $\mathscr{Y} = \{y_i,\ i = 1, \ldots, m\}$ represent the vehicle explanatory variables and response variable, respectively. Further a set of new vehicles $\mathscr{V}_0 = \{(\mathbb{x}_j, y_j),\ j = 1, \ldots, p\}$, where $p = 7243$, is used as a test set for the final model. Prediction for a new vehicle with explanatory variables $\mathbb{x}_0$ is denoted by $\hat{y}_0$. The overall modeling process is illustrated in Figure 1 and this chapter concern the model prediction stage.

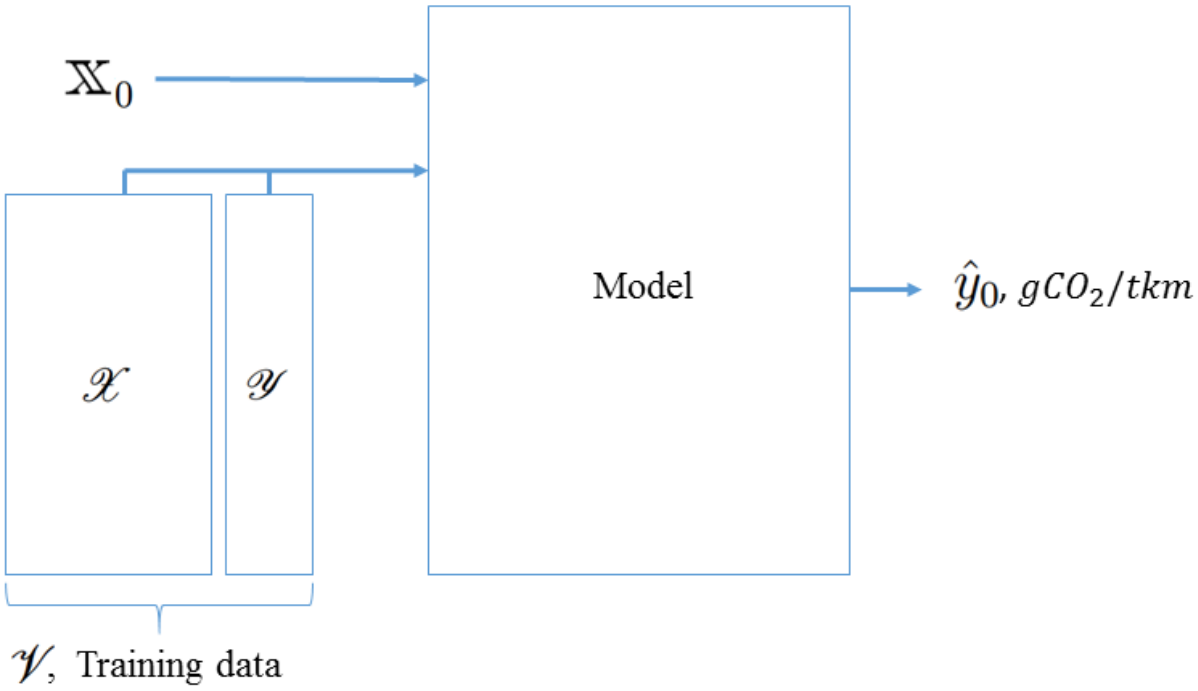

Figure 1: The picture illustrates the general process of a new vehicle$\mathbb{x}_0$ from $\mathscr{X}_0$ being predicted. $\mathscr{V}$ represents the training data of previously observed vehicles. The Model refers to chosen prediction modeling technique applied.

4.2 Exploratory Models

To gain information and knowledge about the relation between the explanatory variables and and the response variable, the exploratory process is heuristic in nature and results are analyzed continuously. Four types of exploratory models are implemented and tested using various theories and parameter variations.

For each type of the exploratory model we try to incorporate the information from continuous and discrete covariates. The discrete covariates are either incorporated by fitting a model for each of the categories combination separately or by using the penalised distances. It is worth to observe that the first type of model may become unfeasible in the prediction process if the new observation does not belong to any categories combination present in the training data.

All these variations are not described, instead a general description of the specific modeling is presented. Further, when testing the exploratory models the set of of training observations is divided in two parts. One test set $\mathscr{V}_{test}$ and one training set $\mathscr{V}_{train}$ where $\frac{1}{5}$ of the observations in $\mathscr{V}$ are placed in $\mathscr{V}_{test}$ and the remaining $\frac{4}{5}$ in $\mathscr{V}_{train}$. The exploratory models are trained on $\mathscr{V}_{train}$ and subsequently evaluated on $\mathscr{V}_{test}$. The performance measures that are being analyzed during the exploratory phase are as described in Section 2.3. The MSE, RMSE are analyzed but foremost is the largest relative prediction error in the empirical 95% quantile what determines how good a model is.

4.2.1 Model 1, Multiple Linear Regression.

The first exploratory model is applied to the whole training set $\mathscr{V}$. The modeling is done by fitting a multiple linear regression with all observed vehicles in $\mathscr{X}_{train}$ as explanatory variables and $\mathscr{Y}_{train}$ as predictive variables. The result is evaluated based on the the prediction errors from the fitted regression model applied on the vehicles in $\mathscr{X}_{test}$ as in Equation 10 in Section 2.2.1.

4.2.2 Model 2, Multiple Linear Regression on sub groups

The set of vehicles are now divided in subsets, this division is done in three ways.

- First, the dividing factor is wheel configuration.
- Second, the dividing factor is a combination of Wheel configuration and chassis adaptation. This way of grouping constitute the two classes of reference loads that VECTO use in simulation.
- The third and last division in to subsets is done by placing all vehicles with the exact same set of categorical variables in the same subset. All three ways to partition the data apply the theory of Multiple Linear Regression as in Model 1.

the MLR fitted to the new vehicle $\mathbb{x}_0$ is then based vehicles from corresponding partition.

Model 2.1 Partitions based on Wheel configuration. divides the set of observed vehicles depending on wheel configuration. All vehicles with wheel configuration 4x2 is placed in one partition, and all vehicles with variations of wheel configuration 6x2 is placed in the second partition. Subsequently a MLR model is fitted to each of these partitions.

Model 2.2 Partitions based on Vehicles Reference load. Following Model 2.1 and documentation on VECTO (CLIMA, 2014), another way of dividing vehicles is implemented. Model 2.2 divide all vehicles with wheel configuration 4x2 and chassis adaptation "Basic" in to one partition, and all other vehicles are placed in a second partition. This constitutes two subsets which in VECTO corresponds to the two different reference loads that are used in the Long Haulage mission profile.

Model 2.3 Categorical covariate partitions. divide the the observed vehicles $\mathcal{V}_{train}$ in subsets where every observation in the same group have the exact same configuration for the categorical covariates described in Table 4. The MLR fitted to the new vehicle $\mathbb{x}_0$ is based on vehicles from corresponding partition.

4.2.3 Model 3, Local Regression

In these models is the concept of local regression is applied as described in Section 2.2.3. What sets the two models apart is the way of distinguishing the nearest neighborhood, $N_{\mathcal{X}_{train}}(\mathbb{x}_0, k, d)$. Model 3.1 divide the observed vehicles in partitions based on the categorical variables, where as model 3.2 use all observed vehicles but introduce penalized distances. The models are all fitted to observed vehicles $\mathcal{V}_{train}$ and evaluated using the new observations in $\mathcal{X}_{test}$. There are two parameters explored throughout model 3 they are found in Table 5.

Table 5: The list of the parameters and their tested values for Models 3.1 and 3.2.

Parameter	Tested values	Model
k	5, 10, 15	Model 3.1
	25, 50, 75	Model 3.2
Distance Metric	$d_{MH}(\cdot,\cdot)$, $d_E(\cdot,\cdot)$ $d_{SE}(\cdot,\cdot)$	Model 3.1-3.2

Model 3.1 Local Regression restricted to continuous variables on Categorical covariate partitions. This model consider the continuous explanatory variables exclusively after partition of vehicles with the same categorical structure is done. Both $N_{\mathcal{X}_{train}}(\mathbb{x}_0, k, d)$ and the fitting of the MLR model is hence performed exclusively on observed vehicles with the same categorical structure. Model 3.1 can be summarized by:

- Local Regression.
- Neighborhood $N_{\mathcal{X}_{train}}(\mathbb{x}_0, k, d)$ based on vehicles from partition with the same categorical structure.
- A MLR is fitted using all continuous covariates exclusively.

Model 3.2 Local regression for all covariates with penalized distances on categorical variables. In this model all covariates are considered. Penalized distances are introduced to the categorical variables in order to construct neighborhood $N_{\mathscr{X}_{train}}(\mathbb{x}_0, k, d)$ consisting of vehicles with similar categorical structure regarding: the wheel configuration, the chassis adaptation and the engine. Thereafter is a MLR fitted to the neighborhood $N_{\mathscr{X}_{train}}(\mathbb{x}_0, k, d)$ using all covariates. Model 3.2 can be shortly summarized by:

- Local Regression.
- Neighborhood $N_{\mathscr{X}_{train}}(\mathbb{x}_0, k, d)$ is based on all vehicles. Distances calculated using penalized distances.
- It places incorporates penalized distances. The penalty's are set to 10^9 on categorical values; Engine, Wheel configuration and Chassis adaptation. The remainder of categorical covariates get penalty's set to 1.
- An MLR is fitted using all covariates.

4.2.4 Model 4, KNN Regression

The modeling technique applied in these models are based on KNN regression and IDW-KNN found in Section 2.2.2 and 2.2.2. In model 4.1 and 4.3 the observed vehicles are divided in partitions based on the categorical variables, whereas model 4.2 and 4.4 introduce penalized weights. Further are model 4.1-4.2 using KNN regression and model 4.3-4.4 use IDW-KNN. Throughout model 4 there are three common parameters that are altered, namely the size of k, the distance metric and in Model 4.3-4.4 the parameter u described in Section 2.1.5. Table 6 soh which parameters were tested for what explanatory models.

Table 6: The list of the parameters and their tested values for Models 4.1 - 4.4.

Parameter		Model
k	5, 7, 10, 15	Model 4.1-4.4
Distance Metric	$d_{MH}(\cdot,\cdot)$, $d_E(\cdot,\cdot)$ $d_{SE}(\cdot,\cdot)$, $d_{Mah}(\cdot,\cdot)$	Model 4.1-4.4
u	1, 1.5, 2, 2.5	Model 4.3-4.4

Model 4.1 KNN Regression on categorical covariate partitions. This model consider partition of vehicles with the same categorical structure and base the neighborhood $N_{\mathscr{X}_{train}}(\mathbb{x}_0, k, d)$ on these partitions. Thereafter the KNN regression method is applied. Model 4.1 can be summarized by:

- KNN regression
- Neighborhood $N_{\mathscr{X}_{train}}(\mathbb{x}_0, k, d)$ based on vehicles from partition with the same categorical structure.

Model 4.2 KNN Regression for all covariates with penalized distances on categorical variables. This model use KNN regression on all covariates and all observed vehicles without partition. Penalized distances are introduced to the categorical variables in order to construct neighborhood $N_{\mathscr{X}_{train}}(\mathbb{x}_0, k, d)$ consisting of vehicles with similar categorical structure regarding; the wheel configuration, the chassis adaptation and the engine. Model 4.2 can be summarized by:

- KNN Regression
- Neighborhood $N_{\mathscr{X}_{train}}(\mathbb{x}_0, k, d)$ based on all vehicles. Distances calculated using penalized distances.
- It incorporates penalized distances. The penalty's are set to 10^9 on categorical values; Engine, Wheel configuration and Chassi adaptation. The remainder of categorical covariates get penelies set to 1.

Model4.3, IDW-KNN on categorical covariate partitions. This model consider the continuous explanatory variables exclusively after partition of vehicles with the same categorical structure is done, as done in model 3.1 and 4.1. Model 4.3 can be summarized by:

- IDW-KNN
- Neighborhood $N_{\mathscr{X}_{train}}(\mathbb{x}_0, k, d)$ based on vehicles from partition with the same categorical structure.

Model4.4 IDW-KNN for all covariates with penalized distances on categorical variables. Model 4.4 use IDW-KNN on all covariates and all observed vehicles without partition. Penalized distances are introduced to the categorical variables in order to construct neighborhood $N_{\mathscr{X}_{train}}(\mathbb{x}_0, k, d)$ consisting of vehicles with similar categorical structure regarding; the wheel configuration, the chassis adaptation and the engine. Model 4.4 can be summarized by:

- IDW-KNN
- Neighborhood $N_{\mathscr{X}_{train}}(\mathbb{x}_0, k, d)$ is based on all vehicles. Distances calculated using penalized distances.
- It incorporates penalized distances. The penalty's are set to 10^9 on categorical values; Engine, Wheel configuration and Chassis adaptation. The remainder of categorical covariates get penalty's set to 1.

4.2.5 Model performance measures

The performance of each model is determined using MSE, RMSE, and also the largest prediction error of the emperical 95%, 99% and 99.9% quantiles as described in Section 2.3. All performance measures are foremost based on relative values.

4.3 Selected models

The exploratory studies lead up to some model for final analysis, it is chosen based on previous heuristic studies. The model is the most general one tested, as expected.

4.4 Final model

The final model is then established from the exploratory phase. The combination of parameters that performs best regarding the largest relative error in the 95% quantile is chosen as final. The final model is tested on $\mathcal{V}_0$ using $\mathcal{V}$ as training data. The errors are recorded and the prediction power of the final model is recorded.

5 Results

5.1 Exploratory models

This section present the results from the exploratory modeling.

We present for each of the models the introduced performance measures. For the models using the categorical covariate partitions we also present the number of failed predictions. This number corresponds to the number of new observed vehicles run through a model that does not belong to any categories combination present in the training data and hence cant be predicted, in relation the number of vehicles predicted.

5.1.1 Model 1, Multiple Linear regression

Model 1 is a global model where all covariates are used to fit a MLR. The following results show that there is a clustering among the observed relative errors. The results of exploratory Model 1 is presented in Table 7 and 8. It has a $RMSE_{rel}$ of 5.00% and the empirical 95% quantile of the relative error is 11.34%, which is not a very good prediction. The errors are also illustrated as a histogram seen in Figure 2 and 3 . These results show that there are two clusters of observations as seen in the scatterplots in Figure 2. We found that these clusters are due to the reference load between vehicles differ in VECTo depending on the vehicle specification.

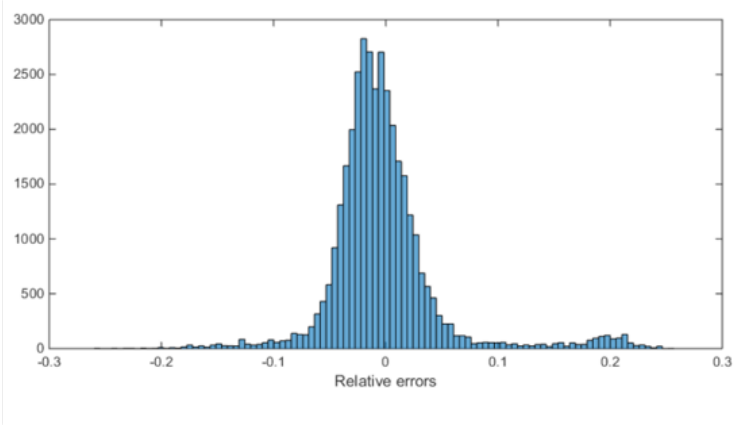

Figure 2: Displays histograms over relative prediction errors from model 1.

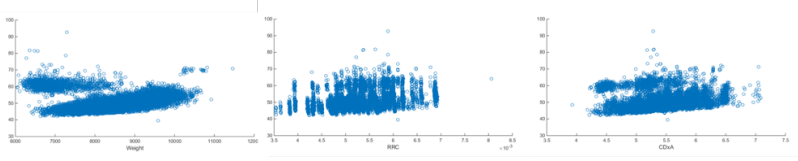

Figure 3: Scatterplott of gCO_2/tkm vs single covariates from Model 1.

Table 7: Performance measures MSE, MSE_{rel}, RMSE and $RMSE_{rel}$ for Model 1.

	Rel. MSE [%]	MSE[$(gCO_2/tkm)^2$]	Rel. RMSE [%]	RMSE [gCO_2/tkm]
Model 1	25.04	7.81	5.00	2.80

Table 8: The empirical 95%, 99% and 99.9% quantiles of relative errors for Model 1.

	95%	99%	99.9%
Model 1 [%]	11.34	20.78	23.62

5.1.2 Model 2, Multiple Linear Regression on sub groups

The results for Model 1 indicates the presence of a clustering among observed vehicles. In Models 2.1 and 2.2 subsets are created in order to find the cause of the cluster. Model 2.3 divide all training data in partitions based on the categorical combination for every vehicle. The performanse measures resulting from Model 2 are found in Table 9 and 10.

Model 2.1 Partition based on Wheel configuration. Dividing the set of observed vehicles in two depending on wheel configuration give the result presented in Table 9. The $RMSE_{rel}$ is 4.380% and the empirical 95% quantile of the relative error is 7.30%.

What is most important from the results in model 2.1 is that we see that the clustering is still present for all 4x2 vehicle partition but it is not present for all 6x2 vehicle partition, this is seen in figure 5. The histograms of all observations in each sub-group is plotted in Figure 4 and a third sub-plot with all errors from the full model plotted in the same histogram.

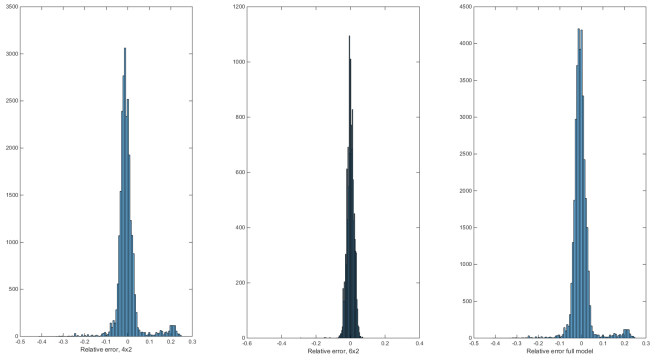

Figure 4: Histograms of the relative prediction error from Model 2.1. 4x2 partition (left histogram), 6x2 partition (center histogram) full model histogram (right histogram).

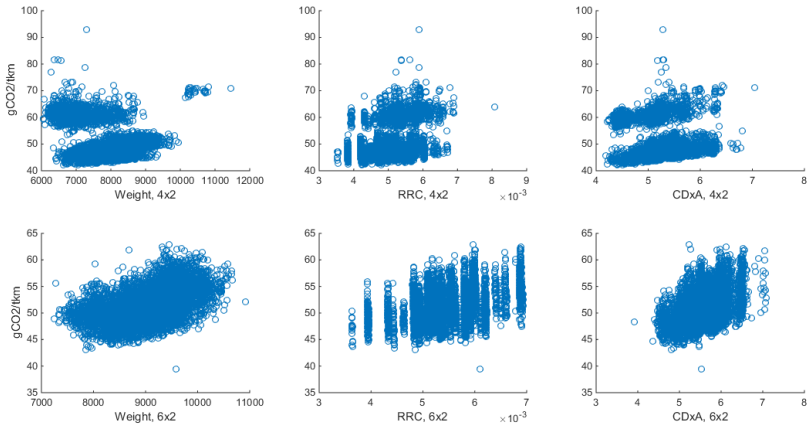

Figure 5: Scatterplott of gCO_2/tkm vs single covariates. Top plots are vehicles with wheel configuration 4x2. The bottom plots are vehicles with wheel configuration 6x2. Left column shows the relation with weight, the center plots shows the relation with RRC and the right plots show the relation with CDxA.

Model 2.2 Partition based on Vehicles Reference load. Studying the results for Model 2.1 and the literature on VECTO (CLIMA, 2014) we found that the reference load put on a vehicle when simulated in VECTO is dependent on two categorical variables, namely the wheel configuration and the chassis adaptation. By dividing the training data in these partitions Model 2.2 was created an it results in a MSE_{rel} of 1.76% and the best largest error in the empirical 95% quantile is 3.71%/1.82 gCO_2/tkm, as seen in 9 and 10. Figure 7 show scatter plot of the individual covariates plotted against the response variable gCO_2/tkm between the two sub-groups and it is clear that the two clusters are now gone. This is also seen in Figure 6, which show three histograms of relative prediction errors from running Model 2.2. One for each of the two partitions and one for the full model.

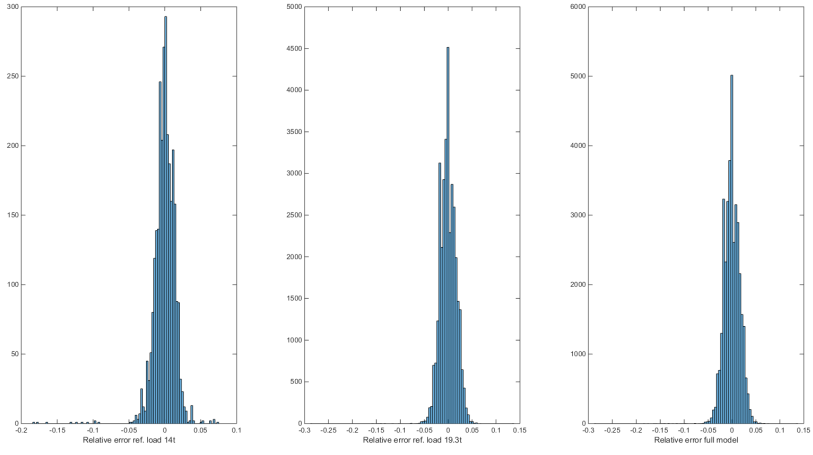

Figure 6: Displays histograms over relative prediction errors from model 2.2. The training data is divided in two partitions based on the reference load set by VECTO when simulated. The 14t reference load partition histogram is displayed to the left, the 19t partition in the middle and the full model histogram to the right.

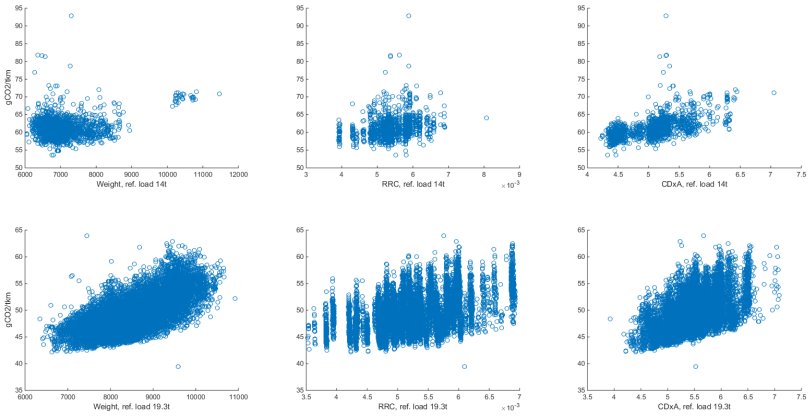

Figure 7: Scatterplott of gCO_2/tkm vs single covariates from Model 2.2. The top plots are for vehicles with reference load 14t and the bottom ones are vehicles with reference load 19.3t. Left column shows the relation with weight, the center plots shows the relation with RRC and the right plots show the relation with CDxA.

Model 2.3 Categorical covariate partitions. In model 2.3 all variables are divided in partitions of vehicles with the same categorical combinations, before fitting an MLR model. Running this model gave a $RMSE_{rel}$ of the relative prediction errors of 162.51% and the largest relative error in the empirical 95% quantile of 18.05%, as seen in 9 and 10. In Figure 8 a histogram is displayed which indicates that large prediction errors was made, but most predictions was better, which is why the relative error empirical 95% quantile show a relative error of 18.05% but the $RMSE_{rel}$ is high, 162.5%. Analysis of the output resulted in identification of a few extreme outliers which affects the linear regression models. Observe that the model results in failed predictions when combinations in the test sets were not observed in the training sets. This is an additional disadvantage of such modelling strategy.

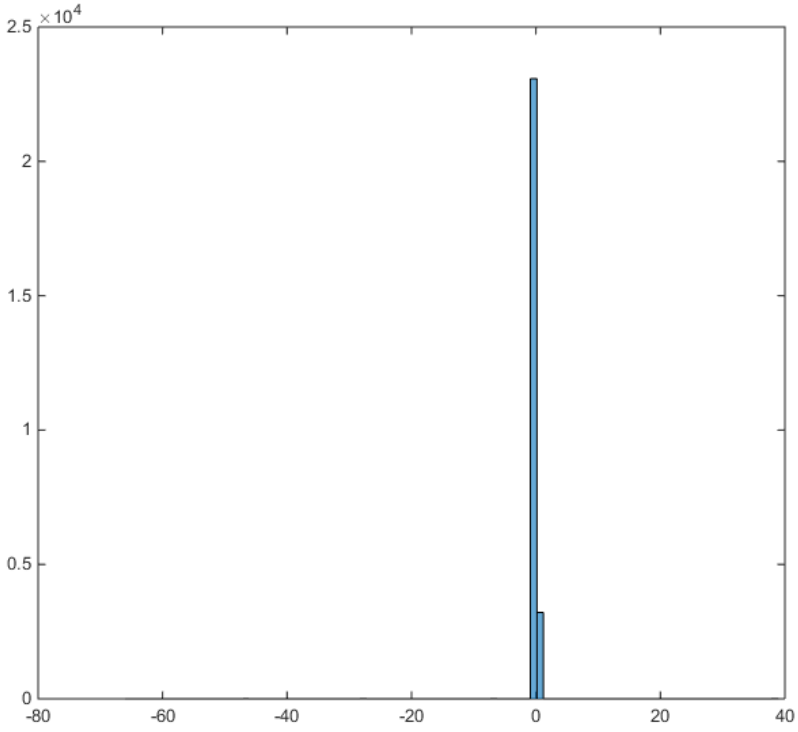

Figure 8: Histograms over relative prediction errors from model 2.3

Table 9: Performance measures MSE, MSE_{rel}, RMSE and $RMSE_{rel}$ for Model 2.1-2.3.

	Rel. MSE [%]	MSE[$(gCO_2/tkm)^2$]	Rel. RMSE [%]	RMSE [gCO_2/tkm]
Model 2.1 (configuration)	19.18	5.90	4.38	2.43
Model 2.2 (reference load)	3.10	0.76	1.76	0.87
Model 2.3 (categorical combinations)	7406.05	264.08	86.06	16.25

Table 10: The empirical 95%, 99% and 99.9% quantiles of relative errors for Model 2.1-2.3 together with percentage of failed predictions in relation to how many predictions were made.

	95%	99%	99.9%	Nr. of failed predictions
Model 2.1 (configuration) [%]	7.30	21.32	24.67	-
Model 2.2 (reference load)[%]	3.71	4.81	6.63	-
Model 2.3 (Categorical combinations)[%]	18.05	29.65	3215.84	0.17%

5.1.3 Model 3, Local regression

In this section the results the two exploratory variations of the Local regression is presented. Model 3.1 incorporates the partition of the traning data based on the exclusive category combination, and in Model 3.2 the all covariates are considered but penalized distances are introduced instead.

Model 3.1 Local Regression restricted to continuous variables on Categorical covariate partitions. The best parameter combination of Model 3.1 from the exploratory phase was when $k = 15$. The 95% quantile of relative error for model 3.1 is 84.02% which can be seen in Table 11. The $RMSE_{rel}$ is found in Table 12 and is $4.35 \cdot 10^5\%$. Figure 9 show a Histogram over the relative model prediction errors and illustrates the magnitude of the relative prediction errors. Similarly to model 2.3 this model fails to predict some observations, which happens in 0.18% of the cases.

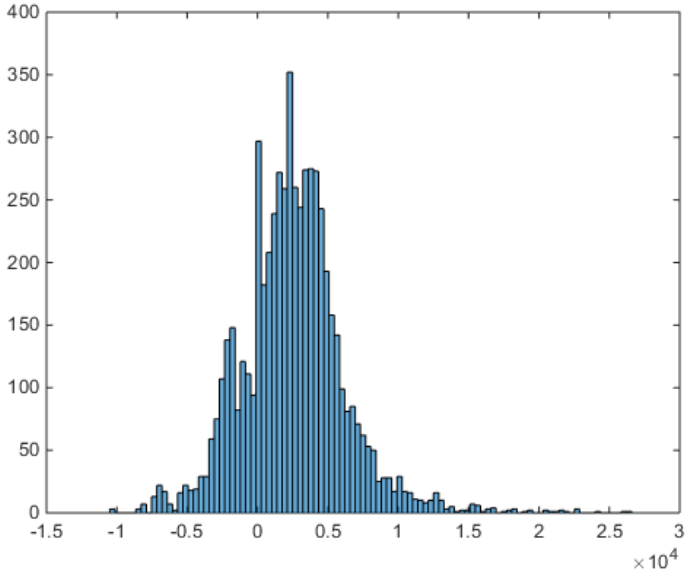

Figure 9: Histogram of relativer prediction errors from model 3.1.

Model 3.2 Local regression for all covariates with penalized distances on categorical variables. The best parameter combination was obtained using $d_{SE}(\cdot, \cdot)$ and $k = 50$. The 95% quantile of relative errors for model 3.2 is 2.78%, as seen in Table 11. Further is the $RMSE_{rel}$ of Model 3.2 is 1.14%. We can observe that the same problem with outliers as in Model 2.3 occurs for this model. In general we can see that Local regression is inferior to the global regression This can be explained by the fact that it is more sensitive to extreme values than the global models. The strategy of using the penalized distances rather than fitting separate models to each combination of categorical variables are significantly better. This can be observed by in specting the results in Tables 11 and 12.

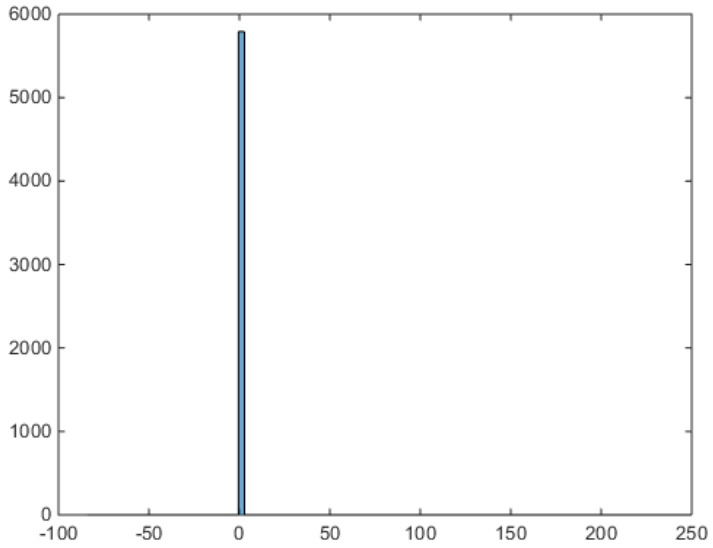

Figure 10: Displays histograms over relative prediction errors from model 3.2.

Table 11: The empirical 95%, 99% and 99.9% quantiles of relative errors for Model 3.1-3.2 together with percentage of failed predictions in relation to how many predictions were made.

	95%	99%	99.9%	Nr. of failed predictions
Model 3.1 Rel. error[%]	84.02	131.17	226.15	0.18%
Model 3.2 Rel. error[%]	2.78	4.57	119.89	-

Table 12: Performance measures MSE, MSE_{rel}, RMSE and RMSE_{rel} for Model 3.1-3.2.

	MSE	MSE_{rel}	RMSE	$RMSE_{rel}$
Model 3.1	$4.30 \cdot 10^{10}$	$1.90 \cdot 10^{11}$	$2.07 \cdot 10^{5}$	$4.35 \cdot 10^{5}$
Model 3.2	0.33	1.31	0.57	1.10

5.1.4 Model 4, KNN Regression

Observing that fitting linear models are not a successful strategy, we moved on to using the KNN regression and the IDW modification. Model 4.1 and 4.3 are using the categorical covariate partitions whereas model 4.2 and 4.4 use all covariates with penalized distances on categorical variables instead.

Model 4.1 KNN Regression on categorical covariate partitions. The best parameter combination of Model 4.1 is obtained by using $k = 7$ and the standardized euclidean distance metric. The 95% quantile of relative errors for Model 4.1 is 3.37% and the RMSE_{rel} is 5.71%. As in all other model using the categorical combination partitions failed predictions occur which for Model 4.1 was 0.24%. The numerical results obtained are presented in Table 13 and 14, and the histogram of the relative prediction errors is presented in Figure 11. Comparing these results from the results in previous sections we see a significant improvement even though we use the categorical covariate partition strategy.

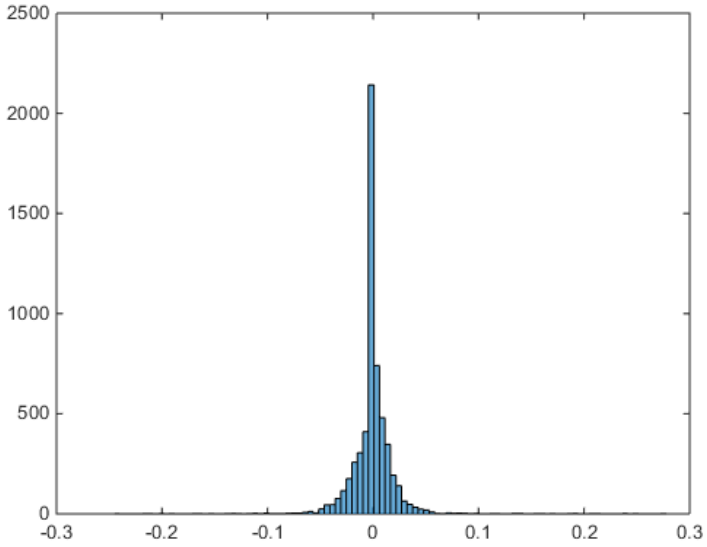

Figure 11: Displays histograms over relative prediction errors from model 4.1.

Model 4.2 KNN for all covariates with penalized distances on categorical variables. This model applies the KNN regression for all covariates and introduce the penalized distances for the categorical covariates. This model did not improve the performance from previous Model 4.1 but for model 4.2 there are no failed predictions. The best parameter combination was obtained by using the standardized euclidean distance metric and $k = 10$. This combination gave a largest relative prediction error in the empirical 95% quantile of 3.73% and a RMSE_{rel} of 2.81%.

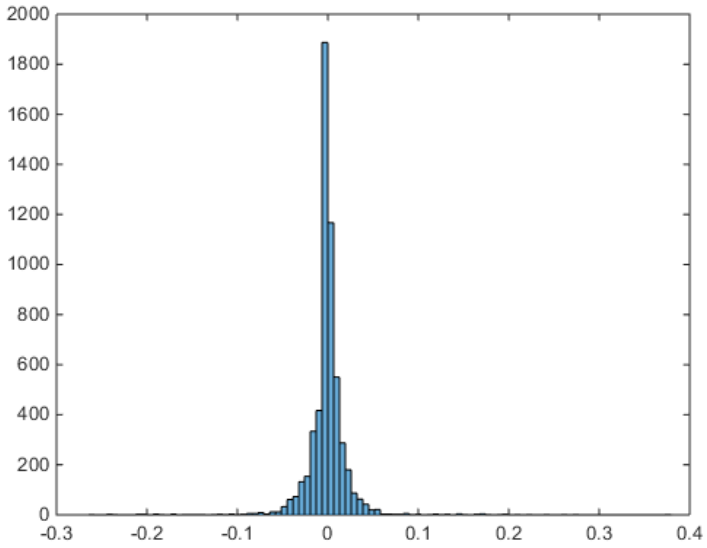

Figure 12: Displays histograms over relative prediction errors from model 4.2

Model 4.3 IDW-KNN on categorical covariate partitions. The best combination of the tested parameter combination of model 4.3 is obtained using $k = 7$, $u = 1.5$ and standardized euclidean distances. The 95% quantile of relative errors is 1.71% and the RMSE_{rel} is 0.81% (see, Table 14 and **??**). Not all categorical combinations are represented which leads to that the number of failed prediction being 0.26%. which also is presented in Table 13 and 14. In Figure 13 a histogram of the relative prediction errors for Model 4.3 is presented. Comparing this model with Model 4.1 we see the incorporating IDW procedure improves the predictive power of the model significantly.

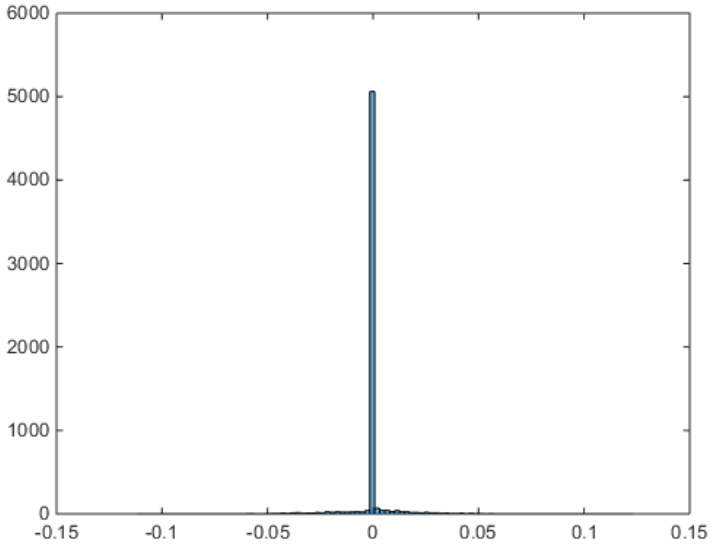

Figure 13: Displays histograms over relative prediction errors from model 4.3

Model 4.4 IDW-KNN for all covariates with penalized distances on categorical variables. The best tested parameter combination of model 4.4 was obtained using the standardized eucidean distance metric, $k = 10$ and $u = 2$. The 95% quantile of relative errors is 1.62%, and the RMSE_{rel} is 0.82%, as seen in Tables 13 and 14. Once again incorporating the IDW results in significant improvement (compared to Model 4.2).

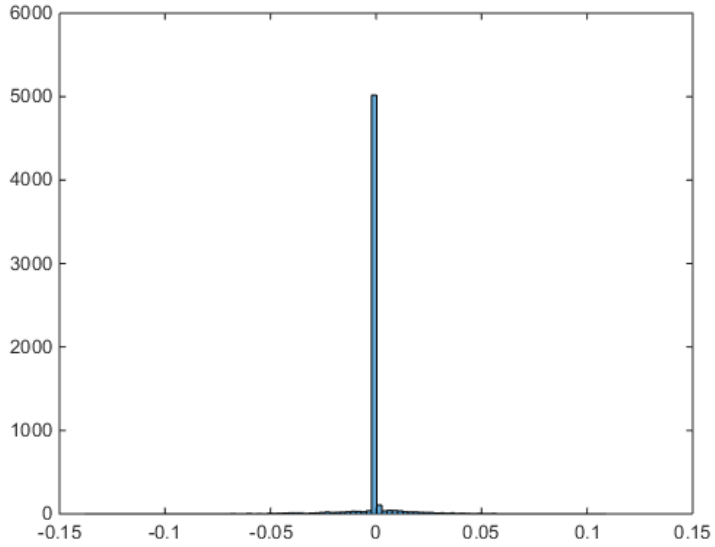

Figure 14: Displays histograms over relative prediction errors from model 4.4.

Table 13: The empirical 95%, 99% and 99.9% quantiles of relative errors for Model 4.1-4.4 together with percentage of failed predictions in relation to how many predictions were made.

	95%	99%	99.9%	Nr. of failed predictions
Model 4.1	3.37	5.51	10.26	0.24%
Model 4.2	3.73	7.54	21.61	-
Model 4.3	1.71	4.10	7.18	0.26%
Model 4.4	1.62	4.05	7.05	-

Table 14: Performance measures MSE, MSE_{rel}, RMSE and RMSE_{rel} for Model 4.1-4.4.

	MSE	MSE_{rel}	RMSE	$RMSE_{rel}$
Model 4.1	10.34	32.57	3.22	5.71
Model 4.2	2.29	7.88	1.51	2.81
Model 4.3	0.17	0.653	0.41	0.81
Model 4.4	0.17	0.67	0.41	0.82

5.2 Final Model

The Model selected as our final model is the Model 4.4. This choice is based on the results from the exploratory analysis. Observe that the model is using the penalized distances to avoid failed predictions. It was chosen since it have the best predictions in the exploratory phase among the models that predicted all vehicles, regardless of specification.

5.3 Final Model performance on test data set

The final model is now tested on $\mathscr{V}_0$, which was excluded form the exploratory analysis and parameter optimization procedures.The results of the prediction on $\mathscr{V}_0$ for the final model trained on the whole set $\mathscr{V}$ are presented in Tables 15 and 16. Further Figure 15 shows the histogram of relative errors. With such model we obtained the 95% quantile relative error to be 0.85% and the $RMSE_{rel}$ to be 0.58%, and even the 99.9% quantile relative error is 5.75%. Figure 15 show the histogram of the relative errors from testing the final model.

Table 15: The empirical 95%, 99% and 99.9% quantiles of relative errors for the final model

Training data	Test data	95%	99%	99.9%
$\mathscr{V}$	$\mathscr{V}_0$	0.85	3.08	5.75

Table 16: Performance measures MSE, MSE_{rel}, RMSE and RMSE_{rel} for final model.

Training data	Test data	MSE $[(\text{gCO}_2/\text{tkm})^2]$	MSE_{rel} [%]	RMSE $[\text{gCO}_2/\text{tkm}]$	$RMSE_{rel}$ [%]
$\mathscr{V}$	$\mathscr{V}_0$	$8.85 \cdot 10^{-2}$	0.34	0.298	0.58

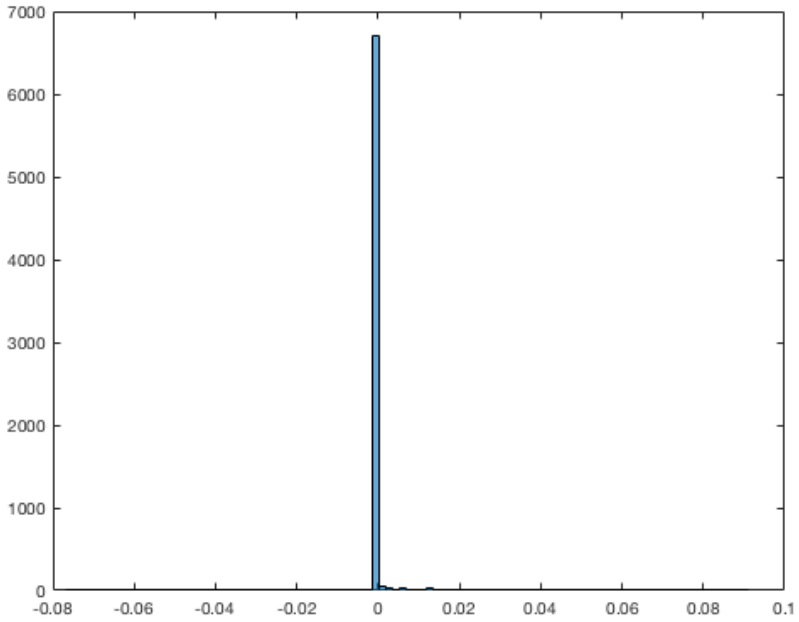

Figure 15: Histogram over the relative errors from testing the final model on test data $\mathscr{V}_0$ using training data $\mathscr{V}$.

6 Discussion

This chapter covers a discussion of the methods used and of the results obtained in Chapter 5.

The Exploratory modeling phase is done heuristically and by testing these initial models an understanding of the relation between covariates and the response variable gCO_2/tkm is established. In the first two models an important discovery regarding a clustering among the observed vehicles is done. A distinct difference could be found, constituting two groups dependent on the reference load that VECTO place on the vehicles during simulation.

In addition to that finding we know from previous studies done at Scania that the gCO_2/tkm is highly dependent on the fuel consumption map, hence dependent on the Engine. This gives three categorical variables proving to be of importance for the modeling accuracy.

The process of establishing a representative neighborhood is enhanced when emphasis is put on categorical variables expressed as categorical variables using penalized distances. This leads to a majority of all predicted vehicles are predicted based on observed vehicles with the same categorical variable structure, which improve the prediction accuracy.

The IDW-KNN model is chosen as final. This method generates the most accurate predictions. In comparison to the KNN regression methos this is superior. Compared to the Local regression modeling the IDW-KNN method also proves to be far more stable. In a case where the vehicle predicted turn out to be different from the majority of the vehicles in the nearest neighborhood, the fitted multiple linear regression cause extreme errors in some cases. This can be seen in Figure 10 where the model in general provide good prediction accuracy, but few extreme outlier are present.

As the observed set of vehicles is as big as it is, interpolation proves to be a suitable method of prediction. For all Vehicles where the distance to the closest neighbor is zero the prediction is set to the observed value of that neighbor. This could be bad if all information is not caught in chosen covariates studied, but as it turns out this is not the case. Figure 15 indicates that most of the prediction errors are zero which they prove to be when studying all errors individually.

It is worth to observe that the final model is tested on training data $\mathscr{V}$ that contains less observed vehicles than the model will in practice be tested on $\mathscr{V} \cup \mathscr{V}_0$. This cahoice was made to include more variation in the models which otherwise would have been missed.

Hence we can expect that the prediction accuracy of the actual model that will be implemented will have better performance. This could be better illustrated by testing the whole model using all observation and LOOCV, but this isnt done since the model should not be tested and trained on the same data to avoid bias.

Summarizing, the construction of a prediction model seem to be a success, depending on the accuracy requirements set by Scania. The IDW-KNN model predicts gCO_2/tkm as simulated in VECTO for any given new HDV in less than a quarter of a second with a prediction error being less than 0.85% for 95% of all vehicles tested.

7 Conclusion

The aim of this thesis is to create a predictive model that can estimate grams of CO_2 per ton and kilometre (gCO_2/tkm) as simulated in VECTO as run in the Long Haulage mission profile. The fact that there are no established performance requirements makes the question of whether the final model reaches the aim of the thesis or not is rather subjective. In spite of that, the resulting prediction accuracy proves to predict new vehicles with a 0.85% relative error for 95% of the tested vehicles and a maximum of 5.75% for the 99.9% of the tested vehicles. The final accuracy of the final model will be even higher in practice since it will be trained on all observed vehicles. So, is the IDW-KNN model accurate enough or not? That is up to Scania to decide. With that being stated, it should be noted that the accuracy of the model is highly dependent on the number of observed vehicles used in the model.

Future recommendations. First and foremost to apply this model in practice the legislative requirements must be established, followed by re-simulation of sold/observed vehicles used as foundation of the model. The fact that the effect of the variables that was fixed in the initial VECTO simulations is unknown must be emphasised, and should be studied further to determine their effect.

To improve the model even further it is one could try to introduce higher order terms in the multiple linear regression (both global and local) setting as this has not been done. We have seen that the parametric model have been very sensitive to the outliers and further investigation of the data set is required to understand the nature of the extreme observations.

Methods that do not assume any linearity should be implemented, e.g. more advanced non-parametric modeling techniques. These were not implemented in this thesis due to software restrictions.

References

Shyam Boriah, Varun Chandola, and Vipin Kumar. Similarity measures for categorical data: A comparative evaluation. *red*, 30(2):3, 2008.

DG CLIMA. Development and validation of a methodology for monitoring and certification of greenhouse gas emissions from heavy duty vehicles through vehicle simulation. 2014.

Seymour Geisser and Wesley O Johnson. *Modes of parametric statistical inference*, volume 529. John Wiley & Sons, 2006.

Gareth James, Daniela Witten, Trevor Hastie, and Robert Tibshirani. *An introduction to statistical learning*, volume 112. Springer, 2013.

Olov Petren. *VECTO - Parameter sensitivity*, 2014.

Donald Shepard. A two-dimensional interpolation function for irregularly-spaced data. In *Proceedings of the 1968 23rd ACM national conference*, pages 517–524. ACM, 1968.

Larry Wasserman. *All of nonparametric statistics*. Springer Science & Business Media, 2006.