CHALMERS
UNIVERSITY OF TECHNOLOGY

Mathematical function for predicting CFPP in diesel fuel blends

Using MLR and neural networks to derive a prediction function for the cold filter plugging point in diesel fuel blends

Master's thesis in Systems, Control and Mechatronics

Stefanus Ivarsson Bergenhem

Department of Electrical Engineering
CHALMERS UNIVERSITY OF TECHNOLOGY
Gothenburg, Sweden 2018

Mathematical function for predicting CFPP in diesel fuel blends

Using MLR and neural networks to derive a prediction function for the cold filter plugging point in diesel fuel blends

Stefanus Ivarsson Bergenhem

CHALMERS
UNIVERSITY OF TECHNOLOGY

Department of Electrical Engineering
Chalmers University of Technology
Gothenburg, Sweden 2018

Mathematical function for predicting CFPP in diesel fuel blends
Using MLR and neural networks to derive a prediction function for the cold filter plugging point in diesel fuel blends
Stefanus Ivarsson Bergenhem

Supervisors:
Robert Lundin, Preem
Thomas Dolff, Preem
Mikael Johansson, Preem
Malin Govik, Preem
Peter Holmqvist, Preem

Examiner: Torsten Wik, Department of Electrical Engineering

Master's Thesis 2018:NN
Department of Electrical Engineering
Chalmers University of Technology
SE-412 96 Gothenburg

Typeset in LaTeX
Printed by [Name of printing company]
Gothenburg, Sweden 2018

Mathematical function for predicting CFPP in diesel fuel blends
Using MLR and neural networks to derive a prediction function for the cold filter plugging point in diesel fuel blends
Stefanus Ivarsson Bergenhem
Department of Electrical Engineering
Chalmers University of Technology

Abstract

The goal of this thesis is to derive a mathematical function for predicting cold filter plugging point (CFPP) in diesel fuel blends. The work is done at Preems refinery in Lysekil, and the data used comes from laboratory test performed by Preem.
The problem is approached from a statistical point of view, using multiple linear regression with and without mixture problem constraints as well as neural network models.
Different sets of prediction variables are tried with varying success. All calculations are done in Matlab, using inbuilt functions. The best result is a model based on cloud point, cetane index, a distillation temperature and the amount of CFPP lowering additive used, with a $R^2 = 0.93$, $RMSE = 1.58$.
A general conclusion drawn is that the additive is the most relevant predictor, and order to derive a better model additional information in the data is needed.

Keywords: CFPP, MLR, Neural network, modelling.

Acknowledgements

I want to thank Viktor Wall Engström. He did his thesis alongside mine, looking at CFPP but from a chemical point of view. We have been working together during the whole project, discussing problems and reaching conclusions together.
A thanks to all the supervisors at Preem. Robert Lundin and Thomas Dolff, our main supervisors and helping with everything and anything. Mikael Johansson for taking an interest in the project and taking the time to explain the control software used. Malin Govik for all the help with explaining the laboratory work.
A thanks to Peter Holmqvist, main supervisor and general go-to guy during the first half of the project, solving any administrative problems and helping with keeping the project on track.
Thanks to Jörgen Sauer for supplying the data used, to Dan Andersson giving a tour and explaining how the refinery works and to Göran Fridolf for showing how the control room works.
Finally, a thanks to Torsten Wik for accepting the role of examiner and being calm and reassuring during the project.

Stefanus Ivarsson Bergenhem, Gothenburg, May 2018

Contents

List of Figures

List of Tables

1
Introduction

The goal of this thesis it to derive a prediction model for cold filter plugging point (CFPP) in diesel fuel relevant enough to be used in a control algorithm. The thesis project was located at Preem in Lysekil. The problem is approached from a statistical point of view, using data gathered from laboratory tests during the last 3 years.
In this chapter some background information to the problem is given. First is a brief description of diesel and CFPP. Then comes a description of the problem from a control system point of view, giving motivation as to why a model is needed. Finishing this chapter are references to work with a similar problem description. The theory for the methods used in this project is given in Chapter 2. Chapter 3 will give a description of the methods used. The results and discussion can be found in Chapter 4 and conclusions will be given in Chapter 5.

1.1 Diesel and CFPP

Diesel is a product produced from crude oil. Crude oil is a highly viscous liquid containing hydrocarbon chains of different lengths and forms. The crude oil can be refined, using different methods such as distillation and cracking. When process is complete all components of roughly the same shape and length have been split up into separate component groups based on boiling point. The groups of interest in this project are kerosene with a boiling point range of about 195-275 $^{\circ}C$ and diesel fuel with a boiling point range of about 275-360 $^{\circ}C$ [7].
Components in these groups are mixed together to a final diesel product. This product has to meet a large number of specifications regarding different properties of the fuel. This covers properties such as density, flash point, kinematic viscosity, cetane number, cloud point or CFPP [15].
For this thesis the focus is on CFPP. Citing the ASTM standard for measuring CFPP, "The CFPP of a fuel is suitable for estimating the lowest temperature at which a fuel will give trouble-free flow in certain fuel systems." [16].
When the temperature of the diesel is lowered some hydrocarbon chains will start to form crystals. The temperature when the crystals starts to appear is called cloud point (CP). Some crystals in the fuel is not a direct problem but as the temperature is lowered more and more crystals will form. At some point there will be enough crystals to clog a filter in a car. When this happens, no fuel will reach the combustion area and the car will not be able to start. The temperature when there are enough crystals to clog a filter is called the cold filter plugging point.

One way to counteract this effect is to add an additive to the mixture. The additive will affect the formation and structure of the crystals which allows the fuel to reach lower temperatures before plugging a filter [17].

1.2 Problem description

The goal of this thesis is to develop mathematical models predicting the CFPP temperature. Such models may then be used in control algorithms. To better motivate the use of a prediction model, consider the blending of a fuel tank from a control system point of view.
Initially, there is an empty tank which will be filled up with diesel. The final product should fulfil the given specifications, which can be seen as constraints. The goal is to get the cheapest possible recipe given the constraints. This result in an initial optimisation problem, were the variables are how much of each component and additive that is used, i.e.

$$\begin{aligned} min \quad & \text{cost} = J(x) \\ s.t. \quad & g(x) \leq \text{specifications} \end{aligned} \tag{1.1}$$

Solving the optimisation problem gives an initial recipe and the process can start. As the tank starts to fill up, samples are taken and analyzed. As no model is perfect and the world is not ideal there will probably be some specifications which are not fulfilled. The remaining filling of the tank can then be seen as a control problem, were the input to the tank is the control signal ($\mathbf{u}$) and the properties to be controlled are the states ($\mathbf{x}$). In state space form the problem could be to determine a controller for the model,

$$\dot{\mathbf{x}} = f(\mathbf{x}, \mathbf{u}). \tag{1.2}$$

The first model of interests is to predict the CFPP of a mixture based on the components and additives used. This could be used both as a part of the initial optimisation and the following control problem. The second model of interests is an observer, predicting the CFPP in a mixture based on other properties in the mixture. This second model is of interest since the standard measurement method for CFPP can take up to 90 minutes to complete [16]. During this time filling of the tank is stopped. An observer based on properties which are quicker to measure could save time.
The data available were set as a boundary to the project. The focus of the project is to attempt to derive a model based on the data given rather than designing new experiments and looking into new tests.

1.3 Related work

Attempts to predict the CFPP in fuels has been done before. Al-Shanableh et al. has developed models for predicting CFPP in bio-diesel from its fatty acid composition using both artificial neural network (ANN) and Fuzzy logic. Both methods showed promising results with R^2 values of 0.96 and 0.98 [13, 14].

Other report the use of ANN to develop a model. Wu et al. used ANN with viscosity, density, refractivity intercept, CFPP of in-going component and weight percentages of each component. They also developed a model for the case when a specific amount of additive is included, but did not use the additive as a parameter. Both models were tested in a refinery with acceptable result [19]. Weimin et al. developed a ANN model for the blending of two components with CFPP and weight percentages of both components as parameters [18].
Semwal et al. report a model based on CFPP and weight percentages which can be used for any number of components [12]. Other authors report models for predicting CFPP based on spectroscopy [1, 10].
There are also reports regarding prediction models for cold properties similar to CFPP [6, 8, 11].
The difference between the reports just mentioned and this thesis is the variables used. The data used in this thesis are the recipe and laboratory analysis result from previous blends done at the refinery during the last 3 years.
Another difference is the additive. In the majority of blends the additive, which is specifically bought is included. The additive has a significant impact on the resulting CFPP any model not including this term will be lacking [2].

2
Theory

The theory presented in Chapter is taken from the books *Introduction to Regression Analysis*, *Experiments with Mixtures*, *Linear regression analysis: theory and computing* and an online course in deep learning. The interested reader can find better and more in-depth explanations there [3, 5, 9, 20].

2.1 Multiple linear regression

Multiple linear regression (MLR) is a method to determine the relationship between a dependent (or response) variable y and a number of independent (or predictor) variables $x_1, x_2, ..., x_k$. The dependent variable y is treated as a random variable. The general MLR model have the following form,

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + ... + \beta_k x_{ki} + \epsilon_i, \tag{2.1}$$

were y_i is the dependent variable, $\{x\}$ are independent variables, $\{\beta\}$ are regression coefficients and $\{\epsilon_i\}$ are random error. The random errors ϵ_i are assumed to have the expected value $E(\epsilon_i) = 0$, variance $Var(\epsilon_i) = \sigma^2$ for $i = 1, 2, ..., n$ and to be *i.i.d.* An MLR model is linear with respect to the regression coefficients, but the predictor variables can take different forms e.g. $x_3 = ln(x_2), x_4 = x_1 * x_2$.
The MLR model can be expressed in matrix form,

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}, \tag{2.2}$$

were

$$\mathbf{X} = \begin{pmatrix} 1 & x_{11} & x_{12} & \cdots & x_{1k} \\ 1 & x_{21} & x_{22} & \cdots & x_{2k} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & x_{n2} & \cdots & x_{nk} \end{pmatrix}, \quad \boldsymbol{\beta} = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_{k-1} \end{pmatrix}, \quad \boldsymbol{\epsilon} = \begin{pmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{pmatrix}. \tag{2.3}$$

The expected value $E(\mathbf{y}) = \mathbf{X}\boldsymbol{\beta}$ and $Var(\mathbf{y}) = Var(\boldsymbol{\epsilon}) = \mathbf{I}\sigma^2$. Giving the following expression for the estimation of y.

$$\hat{y} = \mathbf{X}\boldsymbol{\beta} \tag{2.4}$$

Using matrix format will ease calculations for estimation of $\boldsymbol{\beta}$.
The true regression coefficients $\boldsymbol{\beta}$ will never be known, the goal is however to get a estimate of these coefficients. This is done by the least squares principle:

$$\mathbf{b} = \mathbf{arg\,min}_{\beta}[(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})'(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})], \tag{2.5}$$

were $\mathbf{b}' = (b_0, b_1, ..., b_{k-1})'$ is an estimate of $\boldsymbol{\beta}'$.
The least squares estimation of the regression coefficients is obtained by solving the following equation:

$$\frac{\partial}{\partial \mathbf{b}}[(\mathbf{y} - \mathbf{X}\mathbf{b})'(\mathbf{y} - \mathbf{X}\mathbf{b})] = \frac{\partial}{\partial \mathbf{b}}[\mathbf{y}'\mathbf{y} - 2\mathbf{y}'\mathbf{X}\mathbf{b} + \mathbf{b}'\mathbf{X}'\mathbf{X}\mathbf{b}] = 0. \tag{2.6}$$

Solving this equation gives $\mathbf{X}'\mathbf{X}\mathbf{b} = \mathbf{X}'\mathbf{y}$, and assuming that $(\mathbf{X}'\mathbf{X})$ is non-singular it follows that:

$$\mathbf{b} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}, \tag{2.7}$$

which gives the regression model

$$\hat{y} = x\mathbf{b}. \tag{2.8}$$

It can be shown that the estimator $\mathbf{b}$ is an unbiased estimator of $\boldsymbol{\beta}$:

$$E(\mathbf{b}) = E((\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}) = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'E(\mathbf{y}) = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{X}\boldsymbol{\beta} = \boldsymbol{\beta} \tag{2.9}$$

The variance of $\mathbf{b}$ can be computed as follows:

$$\begin{aligned} Var(\mathbf{b}) &= Var((\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}) \\ &= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'Var(\mathbf{y})((\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}')' \\ &= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{I}\sigma^2((\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}')' \\ &= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\sigma^2 = (\mathbf{X}'\mathbf{X})^{-1}\sigma^2 \end{aligned} \tag{2.10}$$

The variance σ^2 is unknown and is thus replaced by an estimator. An unbiased estimator for σ^2 is given by,

$$s^2 = \hat{\sigma}^2 = \frac{SSE}{n-k}, \tag{2.11}$$

were n is the number of observations, k is the number of regression coefficients and SSE stands for the sum of squares of the residuals (error), which will be discussed in the next section.
If the random error ϵ_i are *i.i.d.* normally distributed then the least squares estimation and the maximum likelihood estimation are the same. If the errors do not have normal distribution then this is no longer necessary true, the use of least squares estimation is however still motivated by the *Gauss-Markov theorem* which say that as long as the errors ϵ_i are uncorrelated and have the same variance then the least squares estimation is the best linear unbiased estimation.

2.2 Model evaluation

2.2.1 Analysis of variance

Analysis of variance (ANOVA) is a way to investigate information about variation in the data used for the model. This can be used to test if the fitted model is statistically significant. It is common to display the information using an ANOVA table, given in Table 2.1.

Table 2.1: ANOVA table

Source of variation	Degrees of freedom	Sum of squares	Mean square
Total	$N-1$	$SST = \sum_{u=1}^{N}(y_u - \bar{y})^2$	
Regression	p	$SSR = \sum_{u=1}^{N}(\hat{y}_u - \bar{y})^2$	$\frac{SSR}{p}$
Residual	$N-(p+1)$	$SSE = \sum_{u=1}^{N}(y_u - \hat{y}_u)^2$	$\frac{SSE}{(N-p-1)}$

SST is the total sum of squares in the data and is computed by summing the squares of the observed y_u about the mean $\bar{y} = (y_1 + ... + y_u + ... + y_N)/N$ were N is the total number of observations. SST has $N-1$ degrees of freedom (d.o.f.).
SSR is the sum of squares due to regression and represent the portion of SST attributed on the fitted model. SSR is therefor the difference between SST and SSE,

$$SSR = SST - SSE, \tag{2.12}$$

with $N-1-(N-p-1) = p$ degrees of freedom, were p is the number of coefficients (not including the intercept).
SSE is the sum of squares of the residuals and is the sum of squares of the difference between the observed y_u and the predicted $\hat{y}_u$. There is $N-p-1$ degrees of freedom associated with SSE.
SSE can be split into two parts, sum of squares for pure error (SS_{PE}) and sum of squares for lack of fit (SS_{LOF}). When an experiment is replicated, and the replication, which have the same conditions gives different values, the sum of squares for variation in these observations is SS_{PE}. SS_{PE} represents the unavoidable error in the data. SS_{PE} have degrees of freedom equal to the number of independent replications.
Removing SS_{PE} from SSE gives the portion of residual due to the model not fitting the data perfectly, i.e.

$$SS_{LOF} = SSE - SS_{PE}. \tag{2.13}$$

The degrees of freedom for SS_{LOF} is the d.f. for SSE minus d.f. for SS_{PE}

2.2.2 F-test & t-test

When a model has been fitted it is important to test if it is statistically significant. This can be done by an overall F-test. The hypothesis to be tested is if at least one of the coefficients β_i, $1 \leq i \leq m$ in the regressed model differs from zero. That would imply that at least one variable is useful in describing the variation of y. This hypothesis is written as follows:

$$H_0 : \beta_1 = \beta_2 = \cdots = \beta_m = 0 \tag{2.14}$$

against

$$H_1 : \text{ At least one } \beta_i \neq 0, \ \ 1 \leq i \leq m. \tag{2.15}$$

The ANOVA approach to doing this test is to look at the ratio between the mean of SSR devided by the mean of SSE.

$$F = \frac{SSR/m}{SSE/(n-m-1)} \tag{2.16}$$

This ratio has an F-distribution, which can be derived from the assumptions made about the error for the general linear model, Equation (2.1). The F-value from Equation (2.16) is compared to table values for the F-distribution, $f_{\alpha,m,n-m-1}$. If $F > f_{\alpha,m,n-m-1}$ then H_0 can be rejected in favor of H_1, i.e. at least one $\beta_i \neq 0$, with a α level significance.
When doing F-test and t-test in a statistical software there is usually a p-value included instead of a table value for $f_{\alpha,m,n-m-1}$ or $t_{n-m-1,\alpha/2}$. The p-value is a measure of the α level certainty of the test. For example, if the F-test given in Equation (2.16) has a corresponding p-value $p = 0.001$, then the F-value is significant at a $0.001 * 100\% = 0.1\%$ level. This means that probability of H_0 being true is 0.1%.
An F-test can also be done for lack of fit, testing the hypothesis H_0: Lack of fit for the model being equal to zero.

$$F_{LOF} = \frac{SS_{LOF}/d.f._{LOF}}{SS_{PE}/d.f._{PE}} \tag{2.17}$$

If F_{LOF} is larger than the corresponding F-distribution table value for the given degrees of freedom and significance, then H_0 is rejected and the conclusion is that the model has lack of fit.
Another common test is the t-test, used to test significance of specific coefficients. Let δ_j denote the jth diagonal element of $(\mathbf{X}'\mathbf{X})^{-1}$ and $s = \sqrt{s^2}$, were s^2 is the unbiased estimator of σ^2 given in Equation (2.11). Then it can be shown that

$$T = \frac{b_j - \beta_j}{s\sqrt{\delta_j}} \tag{2.18}$$

has a t-distribution with $n - m - 1$ degrees of freedom. Based on this it is possible to test

$$H_0 : \beta_j = 0 \tag{2.19}$$

against

$$H_1 : \beta_j \neq 0. \tag{2.20}$$

This is done by calculating

$$T = \frac{b_j}{s\sqrt{\delta_j}}, \tag{2.21}$$

and comparing T with $t_{n-m-1,\alpha/2}$, a table value for t-distribution with $n - m - 1$ degrees freedom and α confidence level. If $|T| > t_{n-m-1,\alpha/2}$ then H_0 can be rejected with α level confidence.
There is a weakness in the t-test. When the test is repeated for a large number of coefficients then the chance for Type 1 error, rejecting H_0 when H_0 is true, increases. To avoid this, it is possible to use a higher level of confidence test, or to just use the F-test for general significance of regression testing.

2.2.3 Coefficient of multiple determination

One way to measure the goodness of fit of a regression model is to use the coefficient of multiple determination R^2, also known as the squared multiple correlation. R^2 is defined as follows

$$R^2 = \frac{SSR}{SST} = 1 - \frac{SSE}{SST}, \tag{2.22}$$

were the second equality comes from Equation (2.12). The R^2 value can be interpreted as the proportion of the total variation in the observations explained by the model.
There is a possible drawback with using the R^2 value. For a set number of N observations, the R^2 value will increase with increasing numbers of parameters. This can be understood by looking at last term in Equation (2.22). SST is fixed for the data set but $SSE = (\mathbf{y} - \mathbf{Xb})'(\mathbf{y} - \mathbf{Xb})$. As shown in Equation (2.5), the $\mathbf{b}$ vector is chosen in order to minimise SEE, When the number of parameters increase the solutions to Equation (2.5) becomes better, SSE will decrease and R will increase.
An adjusted R^2, denoted R^2_A, which takes the addition of new variables into account can be calculated as follows,

$$R^2_A = 1 - \frac{SSE/(N-p)}{SST/(N-1)}. \tag{2.23}$$

In R^2_A the degrees of freedom is considered, which punishes the addition of unnecessary parameters.

2.2.4 Predicted residual error sum of squares (PRESS)

Another common way to evaluate the model is the Predicted residual error sum of squares (PRESS) statistic. The PRESS residual can be seen as the leave-one-out residual. When looking at the *ith* observation, the residual is $e_i = y_i - \hat{y}_i$. But if the *ith* observation is left out when doing the linear regression and then using the new model to predict the y value in the *ith* (left-out) observation, it gives an estimated y value denoted $\hat{y}_{(i)}$. The PRESS residual can then be defined as follows:

$$e_{(i)} = y_i - \hat{y}_{(i)}. \tag{2.24}$$

After doing this for every observation it is possible to calculate the sum of squared PRESS residuals, called PRESS statistics,

$$PRESS = \sum_{i=1}^{N} e^2_{(i)}. \tag{2.25}$$

The PRESS statistics can be used as a measure of how well the model can predict new data as well as a measure of stability for the model. A small value indicates that the model is less sensitive to each sample which is the goal of a good regression model, while a large number could indicate that the model needs to be reworked.
A simpler way to calculate the PRESS statistic, not requiring to keep doing new regressions, is

$$PRESS = \sum_{i=1}^{N} \left(\frac{e_i}{1 - h_{ii}} \right)^2, \tag{2.26}$$

were h_{ii} is the *ith* diagonal element in the hat matrix $H = X(X'X)^{-1}X$.

A R^2 like statistic based on PRESS can be calculated as follows,

$$R^2_{PRESS} = 1 - \frac{PRESS}{SST}, \tag{2.27}$$

which gives a value more easily compared to R^2 and R^2_A.

2.2.5 MAE & RMSE

Two common ways to evaluate models are to look at the mean absolute error (MAE),

$$MAE = \sum_{i=1}^{N} \frac{|e_i|}{N}, \tag{2.28}$$

and the root-mean-squared-error (RMSE),

$$RMSE = \sqrt{\sum_{i=1}^{N} \frac{(\hat{y}_i - y_i)^2}{N}}. \tag{2.29}$$

Both MAE and RMSE fills a similar function, a way to describe how much the estimated value $\hat{y}$ will deviate from the true value y.

2.2.6 Outliers

Given the assumption that the regression model is correctly specified, any individual observation that deviate significantly from the corresponding model value could be considered an outlier.
One way to detect outliers in a vector with random data is to compare each value to the *Median absolute derivation* (MAD),

$$MAD = b * median(|\epsilon_i - median(\boldsymbol{\epsilon})|), \tag{2.30}$$

were $b = 1.4826$ given the assumption of normal distribution [4]. The criterion to remove an outlier is if

$$\epsilon_i < M - 3 * MAD \; or \; \epsilon_i > M + 3 * MAD \tag{2.31}$$

were M is the median of the error vector $\boldsymbol{\epsilon}$.

2.3 Modelling of mixtures

Combining different components to create a product is a mixture problem. When dealing with a mixture problem the model is built around the fractions of each component included in the product. If there are q tanks, then the model will be built around $x_1, x_2, ..., x_q$, were x_i is the fraction of component i. This will lead to a constraint,

$$\sum_{i=1}^{q} x_i = 1, x_i \geq 0, \tag{2.32}$$

representing that the product is a mixture of the components. This new constraint calls for some adjustments to the original linear regression model given in equation (2.1) since the parameters β are no longer unique.

2.3.1 Canonical Polynomials for mixtures

Going back to the original first order polynomial,

$$y = \beta_0 + \sum_{i=1}^{q} \beta_i x_1 + \epsilon, \tag{2.33}$$

and using the identity (2.32) gives the following expression,

$$\hat{y} = \sum_{i=1}^{q} \beta_0 x_i + \sum_{i=1}^{q} \beta_i x_i = \sum_{i=1}^{q} \beta_i^* x_i, \tag{2.34}$$

were $\beta_i^* = \beta_0 + \beta_i$. A similar derivation can be done for second order polynomials. Using the identity given in Equation (2.32), a second order term for component x_i can be written as

$$x_i^2 = x_i(1 - \sum_{j=1, j\neq i}^{q} x_j). \tag{2.35}$$

Using the two constraints given in Equations (2.32) and (2.35), the canonical second order model can be written as

$$\hat{y} = \sum_{i=1}^{q} \beta_i^* x_i + \sum_{i=2}^{q} \sum_{j=1}^{i-1} \beta_{ij} x_i x_j \tag{2.36}$$

2.4 Neural net fitting

This section will explain the basic concept of neural networks. During the project the modelling was done using the Neural network toolbox 9.1 in Matlab, using only shallow networks with 1 hidden layer. The Matlab algorithm is more advanced that the basic concept explained here.

Neural network is a modelling technique which is loosely based on how the human brain works. A network of artificial neurons is set up. A matrix of data is passed through the network and get processed by the nodes in the network and generates an estimation. The estimation is then compared with the observation in each datapoint and the difference is used to update the weights in the network.

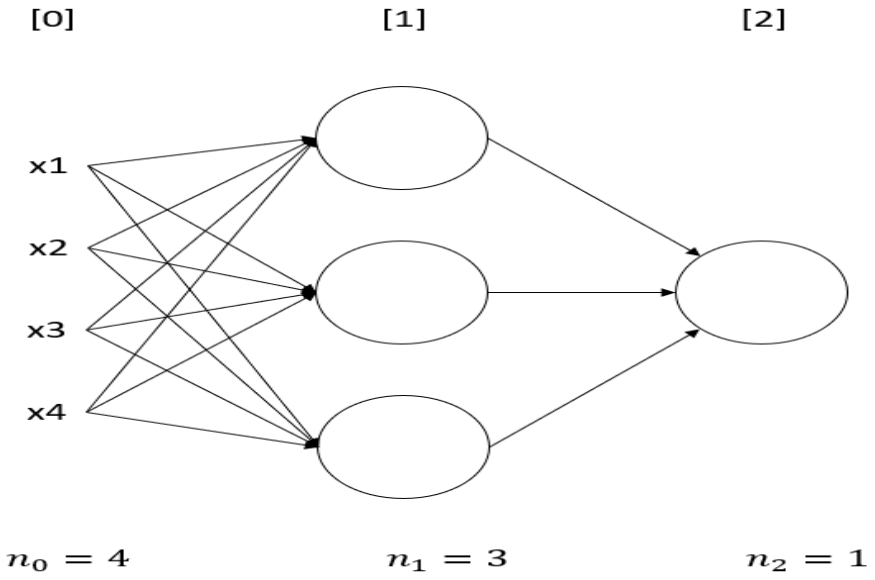

Figure 2.1: Neural network with 1 hidden layer containing 3 nodes.

2.4.1 Forward propagation

Figure 2.1 shows an example of a neural network with 3 nodes in a hidden layer. Each circle represents a node. The nodes are split into vertical columns. Above and below each column there is a number. Each column represents a layer. The number above the column is the number of each layer and the number below is the number of nodes in the layer.

The last layer, [2], is the output layer and the layer in the middle, [1], is the hidden layer. To the left are $x_1, ..., x_4$, the predictors used. The layer with predictors is called the input layer and is the 0:th layer.

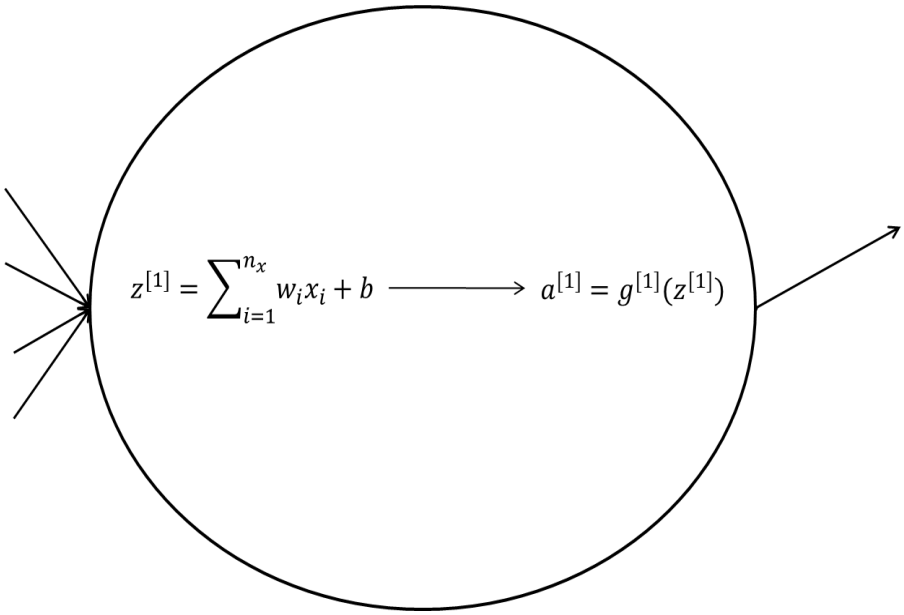

Figure 2.2: The third node in the first layer of the network illustrated in Figure 2.1.

Figure 2.2 shows a node in the hidden layer. The node is split up into two parts, first a linear combination $z^{[1]}$ then a nonlinear activation $a^{[1]}$. The superscript [1] represent that it is a node in the first layer. w_i and b are weights and bias in the linear combination.
$g^{[1]}$ is called activation function. In the hidden layer the activation function used during the project is,

$$g^{[1]}(z^{[1]}) = sigmoid(z^{[1]}) = \frac{1}{1 - e^{-z^{[1]}}}. \tag{2.37}$$

For the output layer the activation is linear, $g^{[2]}(z^{[2]}) = z^{[2]}$. The same calculations are done in each node, but each node has an individual set of weights and bias.
To ease calculations vectorization and matrix form can be used. The whole first layer can be calculated as follows,

$$z^{[1]} = W^{[1]}a^{[0]} + b^{[0]} \tag{2.38}$$

$$a^{[1]} = g^{[1]}(z^{[1]}), \tag{2.39}$$

were

$$z^{[k]} = \begin{pmatrix} z_1^{[k]} \\ \vdots \\ z_{n_k}^{[k]} \end{pmatrix}, \quad a^{[k]} = \begin{pmatrix} a_k^{[k]} \\ \vdots \\ a_{n_k}^{[k]} \end{pmatrix}, \quad W^{[k]} = \begin{pmatrix} w_{1,1}^{[k]} & \cdots & w_{1,n_{k-1}}^{[k]} \\ \vdots & \ddots & \vdots \\ w_{n_k,1}^{[k]} & \cdots & w_{n_k,n_{k-1}}^{[k]} \end{pmatrix} \tag{2.40}$$

and $b^{[k]} = (b_1^{[k]}, ..., b_{nk}^{[k]})^T$. n_k is the number of activations (nodes) in the k:th layer. The inputs are denoted as $x = a^{[0]}$
So far only a forward propagation for a single observation $x^{(i)}$ has been considered. When training a neural network there will be a large number of observations which

all will be taken through the network in each iteration. To ease these calculations additional vectorization is used, e.g.

$$Z^{[k]} = (z^{[k],(1)}, \cdots, z^{[k],(i)}, \cdots, z^{[k],(N)}) \;\; A^{[k]} = (a^{[k],(1)}, \cdots, a^{[k],(i)}, \cdots, a^{[k],(N)}), \tag{2.41}$$

where N is the number of data points. The linear combination and activation for the entire data set can then be calculated simultaneously using matrix calculations,

$$Z^{[k]} = W^{[k]} A^{[k-1]} + B^{[k]} \tag{2.42}$$

$$A^{[k]} = g^{[k]}(Z^{[k]}). \tag{2.43}$$

One thing to notice here is that $B^{[k]}$ is an $n_k \times N$ matrix, but the columns are all identical, that is,

$$B^{[k]} = (b^{[k]}, b^{[k]}, ..., b^{[k]}). \tag{2.44}$$

This expansion is done in order to match the dimensions of all matrices in Equation (2.42).

2.4.2 Cost function

Initially the weights and biases are given small random values. The network is then trained over a number of iterations. The training is done by evaluating a cost function in each iteration and then change the weights and biases according to gradient decent. In this case the mean sum of squared errors is used as cost function, i.e.

$$J(y, a^{[2]}) = \frac{1}{m} \sum_{i=1}^{m} \frac{(a^{[2](i)} - y^{(i)})^2}{2} = \frac{1}{2m} (A^{[2]} - Y)(A^{[2]} - Y)^T. \tag{2.45}$$

where Y is a matrix of measurement data. The goal of the training is to reduce the cost of the model.

2.4.3 Back propagation

Calculating the gradient of each weight and bias starts from the cost function. Taking the derivative of the cost function (2.45) with respect to the final activation gives

$$\frac{\partial J}{\partial A^{[2]}} = \frac{1}{m} (A^{[2]} - Y). \tag{2.46}$$

Next step is to get the partial derivative of the cost function with respect to the linear combination in the output node, $Z^{[2]}$, which turns out to be the same as the partial derivative of the activation since the activation is linear,

$$\frac{\partial J}{\partial Z^{[2]}} = \frac{\partial J}{\partial A^{[2]}} \frac{\partial A^{[2]}}{\partial Z^{[2]}} = \frac{\partial J}{\partial A^{[2]}} * \hat{1} = \frac{1}{m} (A^{[2]} - Y). \tag{2.47}$$

$\hat{1}$ represent a matrix of ones with the same dimensions as $Z^{[2]}$ and the asterisk $(*)$ represent element-wise multiplication. Here the chain rule is used, where the function $A^{[2]} = g(Z^{[2]})$ is the one given in Equation (2.43).

From Equation (2.47) it is easy to calculate the derivative of the weights in the output layer. Using the chain rule,

$$\frac{\partial J}{\partial W^{[2]}} = \frac{\partial J}{\partial Z^{[2]}} \frac{\partial Z^{[2]}}{\partial W^{[2]}} = \frac{1}{m}(A^{[2]} - Y){A^{[1]}}^T, \tag{2.48}$$

where $Z^{[2]}$ is from Equation (2.42). The partial derivative with respect to the bias in the third layer can be calculated in a similar way, first by calculating

$$\frac{\partial J}{\partial B^{[2]}} = \frac{\partial J}{\partial Z^{[2]}} \frac{\partial Z^{[2]}}{\partial B^{[2]}} = \frac{\partial J}{\partial Z^{[2]}} = \frac{1}{m}(A^{[2]} - Y). \tag{2.49}$$

The columns in $\frac{\partial J}{\partial B^{[2]}}$ are then summed in order to get $\frac{\partial J}{\partial b^{[2]}}$,

$$\frac{\partial J}{\partial b^{[2]}} = \frac{1}{m} \sum_{columns} (A^{[2]} - Y). \tag{2.50}$$

With this the output layer is completed, and the weights and biases for the hidden layer can be calculated. The calculations for the remaining layers are the same as for the output layer, with two differences. As activation function *sigmoid* is used instead of the linear combination, and the first step $\frac{\partial J}{\partial A^{[1]}}$ is different.
The complete algorithm for calculating the derivative of weights and biases in a layer l looks as follows:

$$\begin{aligned} \frac{\partial J}{\partial A^{[l]}} &= W^{[l+1]T} \frac{\partial J}{\partial Z^{[l+1]}} \\ \frac{\partial J}{\partial Z^{[l]}} &= \frac{\partial J}{\partial A^{[l]}} * g^{[l]T}(Z^{[l]}) \\ \frac{\partial J}{\partial W^{[l]}} &= \frac{\partial J}{\partial Z^{[l]}} A^{[l]T} \\ \frac{\partial J}{\partial b^{[l]}} &= \sum_{columns} \frac{\partial J}{\partial Z^{[l]}} \end{aligned} \tag{2.51}$$

2.4.4 Updating parameters

The final step of a iteration is to update the weights and biases by gradient decent, i.e.

$$W^{[l]} = W^{[l]} - \alpha \frac{\partial J}{\partial W^{[l]}} \tag{2.52}$$

$$b^{[l]} = b^{[l]} - \alpha \frac{\partial J}{\partial b^{[l]}} \tag{2.53}$$

where α is the learning rate, a design parameter adjusting how quickly the system parameters change. A high α gives a faster learning but might cause instability, while a low value of α gives a more stable and slower learning.

2.4.5 Splitting data into sets

Before a network is trained the data is split into 3 parts, a training set, a validation/development set and a test set. The training set is used to train the system. In

each iteration when the parameter has been updated based on the training set the model is then tested using the validation set. If the model works sufficiently well on the validation set the training is stopped and the model is complete, else a new iteration starts.
When the network training stops, the model is tested on the test set. This is to get an unbiased estimation of how the model preforms, since both the training and development sets are included in the training.

3
Method

3.1 Data

The data used are some of the blends preformed during the last 3 years, 485 blends in total. With each blend comes laboratory results from the components used, the recipe used and laboratory results from the product. Table 3.1 shows all the information available.

Table 3.1: The information available from a single blend (data point/sample).

Tank	Tank 1	···	Tank m	Product
Density (dens.) $[kg/m^3]$	XXX	···	XXX	XXX
Cloud point (CP) $[°C]$	XXX	···	XXX	XXX
Flash point (FLP) $[°C]$	XXX	···	XXX	XXX
Viscosity at 40 $°C$ $[\frac{mm^2}{s}]$ (visk40)	XXX	···	XXX	XXX
Cetane index [no unit] (CI)	XXX	···	XXX	XXX
Cetane number [no unit] (CN)	XXX	···	XXX	XXX
vol% recoverd at 250 $°C$ (R250) [%]	XXX	···	XXX	XXX
vol% recoverd at 350 $°C$ [%]	XXX	···	XXX	XXX
Temperature when 95% recovered (95Rec) $[°C]$	XXX	···	XXX	XXX
CFPP $[°C]$	XXX	···	XXX	XXX
Fraction used (recipe) [%]	XXX	···	XXX	
Additive [ppm]	XXX			

Tank 1,..., Tank m, represents the components used in a specific blend, all components are stored in a tank. XXX represents that a value for this tank component is available. The data comes from 485 tables like this, one for each blend. The following list will give a short clarification of what some of these properties means.

- Cloud point, as mentioned in the background, is the temperature when the first crystals start to appear when cooling the liquid down.
- Flash point is the lowest temperature at which the vapours will ignite given an ignition source.
- Cetane number is measure of the liquids combustion speed and pressure needed to ignite.
- Cetane index is an estimation of cetane number calculated from the density and distillation range.

- Volume-% recovered is how much of the original liquid volume is recovered if the liquid is heated to a specific temperature and the vapor is collected.
- Fraction used is how much of each tank that was used in the recipe, given in %.
- Additive says how much additive, in parts per million, that was used in the recipe.

There are some possible sources of variation and outliers in the data. First of all is the fact that all properties measured by laboratory tests performed after a standard [15]. The standard tests for these are not perfect and allow for some variance. When it comes to the measurement of CPFF there is an occurrence called "false CFPP", which is when the crystals fall out at a temperature 10-15 degrees higher than they "should". When this happens, there is no repetition of the test according to the ASTM standard, so the false CFPP is documented.
The exact amount of additive used in each blend is hard to know. In practice, the mixing of component and additives to create a product is done in large quantities. The components and additive are pumped from large component tanks and mixed together in a pipe. The samples are then taken from this pipe.
This gives two sources of possible variance. First is the assumption of ideal mixing, that the sample contains the exact composition that was given as a set-point to the pumps. This might not always be true. Secondly, the pumps have a set-point, but in practice the flow is approximately normally distributed around this set-point. Looking at data from the additive tank, if the set-point is at 400 PPM, the actual concentration can typically vary from 350 to 450. The exact value of the sample is not analysed. Instead, the set-point is used as the value.
Finally, is the matching of data. The data came from different databases with different notations, so matching a complete data-point was done by hand. While most should be correct there is always the possibility of mistakes.

3.2 Multiple linear regression

All calculations were performed in Matlab. The linear regression, including models with the mixture constraints, were calculated using the Matlab function *fitlm*. ANOVA were calculated using the Matlab function *anova*. The R^2_{PRESS} was not found in the *anova* function and was therefore calculated separately.

3.2.1 Outliers

Handling possible outliers is a sensitive matter. An observation which deviates significantly from the predicted value might be an outlier, but it might also be an indication that the model is not correctly specified. The correct detection of an outlier should be a combination of statistics and understanding of the data. Given the sources of possible outliers mentioned in the previous section, especially the

"false CFPP" and the possible mismatch of data, there is a potential for outliers. However, to be careful not to exclude any "false outliers", any exclusions of outliers will be denoted as possible outliers, *and ANOVA for each model when no potential outliers have been removed from the data is given in Appendix.*
Possible outliers were detected and removed using the following method.

1. Model is regressed.
2. Check for and remove potential outliers using Equation (2.31)
3. Repeat until no more potential outliers are found

3.3 Neural network

Neural networks with one hidden layer was tried as prediction function. The modeling was done using the neural network toolbox 9.1 in Matlab. For each set of prediction variables different sizes of network were evaluated, from 1 to 15 nodes in the hidden layer.
Each network setup was trained 10 times with different initial values of weights and biases. The training of a network might get stuck in a local minimum and give a misleading result. Repeating the training 10 times from different initial conditions should give a more representative result. Sigmoid function was used as activation function and Levenberg-Marquardt back-propagation was used as training algorithm. The data is randomly split 70/15/15 into training, validation and test set before each training of a network.
The evaluation of each network was done based on R^2, RMSE and MAE as defined in Chapter 2.

4

Results and discussion

4.1 Correlation matrix

The first step is to investigate if there are any correlations between CFPP and each of the other properties. This is done by investigate a correlation matrix for the data, looking for linear correlations.

$$\begin{pmatrix} 1.0 & 0.27 & 0.6 & 0.18 & 1.7 \cdot 10^{-5} & -0.11 & -0.3 & -0.053 & 0.41 & 0.046 & 0.054 \\ 0.27 & 1.0 & 0.4 & 0.14 & 0.46 & 0.17 & -0.12 & -0.045 & 0.52 & -0.17 & 0.53 \\ 0.6 & 0.4 & 1.0 & 0.18 & 0.46 & 0.33 & -0.22 & -0.03 & 0.22 & 0.12 & 0.12 \\ 0.18 & 0.14 & 0.18 & 1.0 & 0.16 & 0.047 & -0.086 & -0.021 & 0.14 & 0.098 & 6.4 \cdot 10^{-3} \\ 1.7 \cdot 10^{-5} & 0.46 & 0.46 & 0.16 & 1.0 & 0.56 & -0.17 & -0.07 & 0.25 & 0.06 & 0.25 \\ -0.11 & 0.17 & 0.33 & 0.047 & 0.56 & 1.0 & -0.095 & 0.066 & -0.016 & 0.03 & 0.094 \\ -0.3 & -0.12 & -0.22 & -0.086 & -0.17 & -0.095 & 1.0 & 0.024 & -0.12 & -0.019 & -0.082 \\ -0.053 & -0.045 & -0.03 & -0.021 & -0.07 & 0.066 & 0.024 & 1.0 & -0.095 & 0.049 & -0.051 \\ 0.41 & 0.52 & 0.22 & 0.14 & 0.25 & -0.016 & -0.12 & -0.095 & 1.0 & -0.33 & 0.35 \\ 0.046 & -0.17 & 0.12 & 0.098 & 0.06 & 0.03 & -0.019 & 0.049 & -0.33 & 1.0 & -0.76 \\ 0.054 & \underline{0.53} & 0.12 & 6.4 \cdot 10^{-3} & \underline{0.25} & 0.094 & -0.082 & -0.051 & \underline{0.35} & \underline{-0.76} & 1.0 \end{pmatrix} \tag{4.1}$$

The properties of this correlation matrix can be found in Table 4.1

Table 4.1: Properties (variables) of the correlation matrix (4.1)

$x_1 = dens$	$x_2 = CLP$	$x_3 = FLP$	$x_4 = Visk40$	$x_5 = CI$	$x_6 = CN$
$x_7 = R250$	$x_8 = R350$	$x_9 = 95Rec$	$x_{10} = Additive$	$x_{11} = CFPP$	

The properties that are somewhat correlated to CFPP are CP, CI, 95Rec and additive, underlined in the correlation matrix (4.1). This does only show linear correlations, possible nonlinear correlations might still exist.

4.2 MLR based on properties in the diesel product

Looking at the correlation matrix shows that CP, CI, 95 Rec and additive are most (linearly) correlated to CFPP. A first order linear regression model was therefor set up with these as predictors.

4.2.1 First order linear model based on properties in the diesel product

Equation (4.2) shows a first order polynomial regression model based on the predictors given in Table 4.2.

$$\hat{y} = 61.93 + 1.16x_1 + 0.71x_2 - 0.29x_3 - 0.0235x_4 \tag{4.2}$$

Table 4.2: Independent variables based on properties in the diesel product.

$x_1 = CP$	$x_2 = CI$	$x_3 = 95Rec$	$x_4 = additive$

This model suggests that CFPP will increase with increasing CP and CI while decreasing with increasing 95Rec and concentration additive used. The ANOVA table for the model is given in Table 4.3.

Table 4.3: ANOVA table for the first order linear model (4.2) with the predictors given in Table 4.2.

Variation	d.f.	Sum of squares	Mean square	F-score	p-value
Total	484	$SST = 18493$			
Regression	4	$SSR = 14257$	3564.3	403.87	4.56e-152
Residual	480	$SSE = 4236.1$	8.8252		
LOF	479	$SS_{LOF} = 4236.1$	8.8252	inf	0
PE	1	$SS_{PE} = 0$	0		
$R^2 = 0.771$		$R^2_A = 0. = 0.769$		$R^2_{PRESS} = 0.762$	
$RMSE = 2.97$		$MAE = 2.31$			

The F-test for regression and corresponding p-value shows that the regression is significant with high probability. The R^2 value indicate that 77% of the variance in the data is covered by the model. R^2_A and R^2_{PRESS} are close to R^2 indicating that the R^2 value can be trusted. The F-test for lack of fit and corresponding p-value indicates that the model can be improved. While this conclusion might be true, the F-test in this case is questionable given that there is only one replication, and knowing the measurement methods the pure error should be nonzero.
The algorithm given for detecting and removing outliers revealed 15 potential outliers in this case. Removing the potential outliers gave the following model:

$$\hat{y} = 86.019 + 1.0082x_1 + 0.94537x_2 - 0.38838x_3 - 0.027104x_4. \tag{4.3}$$

Overall the coefficients are of the same size, indicating that no major change in trends has happened. The ANOVA for this model is given in Table 4.4.

Table 4.4: ANOVA table for first order linear model (4.3) with the predictors given in Table 4.2. 15 outliers removed from the data.

Variation	d.f.	Sum of squares	Mean square	F-score	p-value
Total	469	$SST = 16754$			
Regression	4	$SSR = 13890$	$3472,6$	563.78	8.48e-177
Residual	465	$SSE = 2864.1$	6.16		
LOF	464	$SS_{LOF} = 2864.1$	6.17	inf	0
PE	1	$SS_{PE} = 0$	0		
$R^2 = 0.829$		$R^2_A = 0.828$		$R^2_{PRESS} = 0.825$	
$RMSE = 2.48$		$MAE = 2.02$			

The ANOVA for the model when the outliers have been removed shows an improvement. The F-score and R^2 values has increased and RMSE/MAE has also decreased.

4.2.2 Second order polynomial model based on properties in the diesel product

A second order polynomial was investigated, using the predictors given in Table 4.2, i.e.

$$\hat{y} = \beta_0 + \sum_{i=1}^{4} \beta_i x_i + \sum_{i=1}^{4} \sum_{j=i}^{4} \beta_{ij} x_i x_j \tag{4.4}$$

It turns out that the matrix $(\mathbf{X}'\mathbf{X})^{-1}$, which is a part of the linear regression (Equation (2.7)), was close to singular, so the results cannot be completely trusted. However, looking at the t-statistics for the coefficients (found in Appendix Table A.3) the t-score for coefficients for terms including additive has a higher certainty compared to others.

4.2.3 Second order model with respect to additive and interactions with additive based on properties in the diesel product

Inspired by the t-statistics for the second order polynomial model, given in Appendix Table A.3, a model with second order term for additive and interactions between additive and other parameters were set up. This model looks as follows,

$$\hat{y} = \beta_0 + \sum_{i=1}^{4} \beta_i x_i + \sum_{i=1}^{4} \beta_{i4} x_i x_4, \tag{4.5}$$

with the predictors are given in Table 4.2.
The resulting ANOVA after 22 potential outliers have been removed is given in Table 4.5. The ANOVA table for the model with no outliers removed from the data can be found in Appendix, Table A.4.

Table 4.5: ANOVA table for the model (4.5) with predictors given in Table 4.2. 22 outliers have been removed from the data.

Variation	d.f.	Sum of squares	Mean square	F-score	p-value
Total	462	$SST = 17755$			
Regression	8	$SSR = 16539$	2067.4	771.81	7.94e-259
Residual	454	$SSE = 1216.1$	2.6796		
LOF	453	$SS_{LOF} = 1216.1$	2.6796	inf	0
PE	1	$SS_{PE} = 0$	0		
$R^2 = 0.932$		$R^2_A = 0.93$		$R^2_{PRESS} = 0.928$	
$RMSE = 1.64$		$MAE = 1.26$			

The F-test and corresponding p-value indicates that the regression is significant. The R^2 indicates that the model covers 93% of the variance in the data and R^2_A and R^2_{PRESS} are close to R^2 indicating that the R^2 value can be trusted.
Looking at the t-statistics for the coefficients (Appendix, Table A.5) the values of β_0, β_2 and β_3 are not certain to be different from zero. Removing these terms gives the following reduced model,

$$\hat{y} = \beta_1 x_1 + \beta_4 x_4 + \sum_{i=1}^{4} \beta_{i4} x_i x_4. \tag{4.6}$$

The ANOVA for the reduced model (4.6) can be found in Table 4.6. 26 potential outliers has been removed from the data, ANOVA for the model with no outliers removed can be found in appendix Table A.6.

Table 4.6: ANOVA table for the reduced second order model (4.6) with predictors given in Table 4.2. 26 potential outliers removed from the data.

Variation	d.f.	Sum of squares	Mean square	F-score	p-value
Total	459	$SST = 17387$			
Regression	6	$SSR = 16249$	2708.2	1078.6	1.29e-264
Residual	453	$SSE = 1137.4$	2.5108		
LOF	414	$SS_{LOF} = 1112.3$	2.6867	4.1704	4.16e-7
PE	39	$SS_{PE} = 25.125$	0.644		
$R^2 = 0.933$		$R^2_A = 0.933$		$R^2_{PRESS} = 0.931$	
$RMSE = 1.58$		$MAE = 1.23$			

The F-score and corresponding p-value indicate that the regression is significant. The R^2, R^2_A and R^2_{PRESS} values indicates that the model covers 93% of the total variation in the data. The F-score for lack of fit and corresponding p-value indicates that the model can be improved. The reason for a reasonable PE value appearing in this model is that for some blends there is no additive added. This makes all the predictors except CP equal to zero in the model (Equation (4.6)). For a number of these the CP is the same, which gives replications to calculate PE.

T-statistics for the coefficients indicates that all coefficients are different from zero with high probability (Appendix, Table A.7).
Comparing the ANOVA of the reduced model (Table 4.6) with the complete model (Table 4.5) shows that both have similar R^2 and RMSE values, indicating a similar performance. A difference to be found is the p-value related to F-score for regression. The value for the reduced model is about 10^4 times smaller, indicating a more significant model. Given this and the fact that all the coefficients in the reduced model are significant, the reduced model seems better.
Some time was spent on improving this model, but with no success. This model could be used as an observer, but cannot easily be used to predict the CFPP in a product based on components. This because the predictors are properties in the already finished product, not based on components and their properties.

4.3 Mixture models

The diesel blending is a mixture of different components. Regarding the problem as a mixture brings the mixture constraint, Equation (2.32), and corresponding models. An assumption in this case is that a given component tank always contains the same kind of oil, i.e. no variation over time. The predictors in this case are the fractions of each component. The additive is seen as a component in this case. The assumption can be checked by looking at the properties in each tank over time, as they stay more or less the same this assumption has been taken.
There is one component tank which has only been used in two blends. These two blends have therefore been removed from the data used in the mixture models and the remaining 483 blends are used as the data-set.

4.3.1 First order canonical mixture model

The first order canonical mixture model looks as follows,

$$\hat{y} = \sum_{i=1}^{9} \beta_i x_i. \tag{4.7}$$

The predictors for this model are given in Table 4.7.

Table 4.7: Predictors used in mixture model. x_i represents the fraction of tank i used in the blend.

x_1 = tank 1	x_2 = tank 2	x_3 = tank 3	x_4 = tank 4	x_5 = tank 5
x_6 = tank 6	x_7 = tank 7	x_8 = tank 8	$x_9 = additive$	

The ANOVA for this model is given in Table 4.8. 8 possible outliers have been removed, ANOVA table for model based on the complete set can be found in Appendix, Table A.8

Table 4.8: ANOVA table for first order mixture (4.7) with predictors given in Table 4.7. 8 potential outliers removed from the data.

Variation	d.f.	Sum of squares	Mean square	F-score	p-value
Total	475	$SST = 16636$			
Regression	9	$SSR = 12246$	1360.6	144.43	9.2e-129
Residual	466	$SSE = 4390$	9.42		
LOF	431	$SS_{LOF} = 4333.9$	10.06	6.256	4.86e-9
PE	35	$SS_{PE} = 56.167$	1.605		
$R^2 = 0.736$		$R^2_A = 0.731$		$R^2_{PRESS} = 0.724$	
$RMSE = 3.07$		$MAE = 2.44$			

The F-score value and corresponding p-value for regression indicate that the regressed model is significant. The R^2, R^2_A and R^2_{PRESS} values indicate that the model covers 73% of the variance in the data. The F-score and corresponding p-value for lack of fit indicate that the model can be improved.
After looking at the t-statistics (Appendix, Table A.9) it is reasonable to suspect that the coefficients β_1, β_2 and β_6 might not differ from zero. However, given the mixture constraint (2.32) just straight removing these terms from the model is not recommended. The workaround will be shown in the next section.

4.3.2 First order mixture model based on reduced mixture components

As some coefficients in the model given in Equation (4.7) might be equal to zero the model can be simplified by removing these. This is done by first assuming,

$$\beta_1 = \beta_2 = \beta_6 = \bar{\beta}, \;\; \bar{x} = x_1 + x_2 + x_6 \tag{4.8}$$

Using the mixture constraint $\bar{x}$ can be replaced,

$$\bar{x} = 1 - x_3 - x_4 - x_5 - x_7 - x_8 - x_9. \tag{4.9}$$

This yields the following first order reduced mixture model,

$$\hat{y} = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \beta_5 x_5 + \beta_6 x_6. \tag{4.10}$$

with predictors given in Table 4.9. To keep things clear, the mixture models when the 3 tanks have been removed will be called reduced mixture model.

Table 4.9: Predictors used in reduced mixture models.

x_1 = tank 3	x_2 = tank 4	x_3 = tank 5
x_4 = tank 7	x_5 = tank 8	$x_6 = additive$

Anova for the first order reduced mixture model (4.10) is given in Table 4.10. 8 possible outliers have been removed. ANOVA for regression with no outliers removed can be found in Appendix, Table A.10.

Table 4.10: ANOVA table for the first order reduced mixture model (4.10) with the predictors given in Table 4.7. 8 potential outliers removed from the data.

Variation	d.f.	Sum of squares	Mean square	F-score	p-value
Total	474	$SST = 16636$			
Regression	6	$SSR = 12239$	2039.8	217.12	8.89e-132
Residual	468	$SSE = 4396.8$	$9,3949$		
LOF	433	$SS_{LOF} = 4340.6$	10.03	6.256	5.07e-9
PE	35	$SS_{PE} = 56.167$	1.605		
$R^2 = 0.736$		$R^2_A = 0.732$		$R^2_{PRESS} = 0.727$	
$RMSE = 3.07$		$MAE = 2.44$			

The F-score and corresponding p-value for regression indicates that the model is significant. The R^2, R^2_A and R^2_{PRESS} values indicates that this model covers about 73% of the variance in the data. The F-score and corresponding p-value for lack of fit indicates that the model can be improved.
Comparing this reduced mixture model (Equation (4.10)) with the canonical model (Equation (4.7)) shows that both have similar R^2, RMSE and lack of fit. A difference can be found in the p-values corresponding to the F-score for regression in both models. The reduced mixture model has a smaller value, indicating a more significant regression. This is the result of both models preforming equally well, but the reduced mixture model is simpler having 3 terms less.

4.3.3 Second order polynomial reduced mixture model

A second order polynomial model for the reduced mixture component is set up,

$$\hat{y} = \beta_0 + \sum_{i=1}^{6} \beta_i x_i + \sum_{i=1}^{6} \sum_{j=i}^{6} \beta_{ij} x_i x_j. \tag{4.11}$$

The independent variables are given in Table 4.9. ANOVA for this model is given in Table 4.11 with 9 potential outliers removed. The ANOVA table for the model with no outliers removed from the data is given in Appendix, Table A.12

Table 4.11: ANOVA table for the second order mixture model (4.11) with predictors given in Table 4.9. 9 potential outliers removed from the data.

Variation	d.f.	Sum of squares	Mean square	F-score	p-value
Total	473	$SST = 16299$			
Regression	27	$SSR = 13454$	498.3	78.13	1.53e-150
Residual	446	$SSE = 2844.7$	6.38		
LOF	411	$SS_{LOF} = 2788.6$	6.79	4.23	1.27e-6
PE	35	$SS_{PE} = 56.167$	1.605		
$R^2 = 0.825$		$R^2_A = 0.815$		$R^2_{PRESS} = 0.800$	
$RMSE = 2.53$		$MAE = 1.95$			

The F-score and corresponding p value indicate that the model is significant. The R^2, R^2_A and R^2_{PRESS} values indicate that this model covers about 81% of the variance in the data. Looking at the t-statistics for the coefficients, given in appendix Table A.13, most of the coefficients seems to be not significantly different from zero. That indicates that the model should probably not be completely trusted and there might be unnecessary terms in the model.
After some investigation it became clear that the only added term that significantly improved the results was the second order additive term. The final model for the reduced mixture predictors then looks as follows,

$$\hat{y} = \beta_0 + \sum_{i=1}^{6} \beta_i x_i + \beta_7 x_6^2, \tag{4.12}$$

with the independent variables given in Table 4.9. ANOVA for this model is given in Table 4.12 with 14 potential outliers removed. ANOVA for the model when no outliers are removed from the data is given in Appendix, Table A.14.

Table 4.12: ANOVA table for the model (4.12) with predictors given in Table 4.9. 14 potential outliers removed from the data.

Variation	d.f.	Sum of squares	Mean square	F-score	p-value
Total	468	$SST = 15949$			
Regression	7	$SSR = 12852$	1836	273.3	1.26e-159
Residual	461	$SSE = 3097.3$	6.72		
LOF	426	$SS_{LOF} = 3041.1$	7.14	4.45	6.28e-7
PE	35	$SS_{PE} = 56.167$	1.605		
$R^2 = 0.806$		$R^2_A = 0.803$		$R^2_{PRESS} = 0.798$	
$RMSE = 2.59$		$MAE = 2.01$			

F-score and corresponding p-value for regression indicate that the model is significant. The R^2, R^2_A and R^2_{PRESS} indicate that the model covers 80% of the variation in the data. F-score and corresponding p-score value indicate that the model can be improved.
Comparing this model (Equation (4.12)) with the complete second order model (Equation (4.11)). The complete model has a bit higher R^2 and a bit lower RMSE. However, looking at the t-statistics for this model (Appendix, Table A.15) all coefficients, with the possible exception of the intercept, are relevant with high probability. This fact outweighs a minor loss in RMSE/R^2.

4.4 Simplified mixture components

In total 8 different component tanks (given in Table 4.7) have been used in the blends during the last 3 years. Some of these tanks have somewhat varying content during this time. There are also some tanks with similar content and could therefore be considered as a group. This motivates an attempt to group the component tanks by their CFPP.

A histogram showing the CFPP measured in the component tanks before each blend on the x-axis and the number of time each CFPP has been measured on the y-axis is given in Figure 4.1. The large number of measurements does not mean that there are many component tanks, just that the existing tanks are frequently measured.

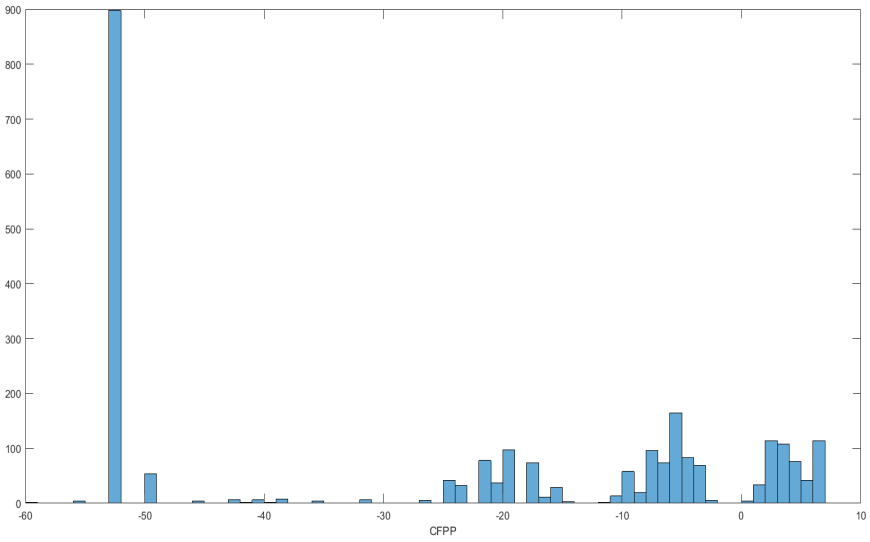

Figure 4.1: Histogram showing the CFPP measured in all component tanks before each blend for the entire data set. The x-axis shows the CFPP measured and the -axis shows how many times each CFPP have been measured.

Figure 4.1 shows 4 clear groupings,

- G1: $CFPP \in (-53, -48)$
- G2: $CFPP \in (-27, -14)$
- G3: $CFPP \in (-12, -2)$
- G4: $CFPP \in (0, 10)$

A part of the data with $CFPP \in (-47, -30)$ is not included in the given groups and are left out of the modelling in the attempt to derive a functional, if a bit more narrow, model. The left-out part corresponds to 40 blends out of 485, resulting in 445 data points to be used in total.
A first order mixture model based on component groupings just described was derived,

$$\hat{y} = \sum_{i=1}^{5} b_i x_i, \tag{4.13}$$

with the independent variables given in Table 4.13. Models based on these predictors will be called simplified components mixture models.

Table 4.13: Predictors for the simplified mixture component models

$x_1 = G1$	$x_2 = G3$	$x_3 = G3$	$x_4 = G4$	$x_5 = additive$

A model evaluation is given in Table 4.14. 3 possible outliers have been removed. ANOVA for the model with no outliers removed can be found in Appendix, Table A.17.

Table 4.14: ANOVA table for the first order simplified components mixture model (4.13) with predictors given in Table 4.13. 3 possible outliers removed from the data.

Variation	d.f.	Sum of squares	Mean square	F-score	p-value
Total	442	$SST = 15215$			
Regression	5	$SSR = 10711$	2142.2	207.87	4.366e-113
Residual	437	$SSE = 4503.6$	10.306		
LOF	404	$SS_{LOF} = 4449.5$	11.014	6.71	5.07e-9
PE	33	$SS_{PE} = 54.167$	1.614		
$R^2 = 0.704$		$R^2_A = 0.701$		$R^2_{PRESS} = 0.696$	
$RMSE = 3.21$		$MAE = 2.58$			

The F-score value and corresponding p-value given in Table 4.14 indicates that the regression is significant with a high degree of probability. The R^2, R^2_A and R^2_{PRESS} indicates that the model covers about 70% of the variation in the data.
Comparing this model to the reduced component linear mixture model given in Equation (4.10), R^2 and RMSE are a slightly better in the reduced components compared to simplified components.
The strength of the simplified components grouping is that it does not require the assumption that a specific tank always contain the same kind of oil, but is instead based on a property of the oil. This makes the model more flexible to use. The chosen grouping by CFPP value in components is simple, and an improved way of grouping could probably be found.

4.4.1 Simple components, second degree additive model

In an attempt to improve the simple components model (Equation (4.13)) a second order additive term is added to the model. According to the mixture constraint given in Equation (2.35) the new model can be written as follows,

$$\hat{y} = \sum_{i=1}^{5} b_i x_i + \sum_{i=1}^{4} b_{i5} x_i x_5. \tag{4.14}$$

A model evaluation is given in Table 4.15. 6 possible outliers have been removed. ANOVA for the model when no outliers have been removed is given in Appendix, Table A.18

Table 4.15: ANOVA table for the simplified components mixture model with second order additive term (4.14) with predictors given in Table 4.13. 6 possible outliers removed from the data.

Variation	d.f.	Sum of squares	Mean square	F-score	p-value
Total	439	$SST = 14757$			
Regression	9	$SSR = 11727$	1303	184.91	8.814e-142
Residual	430	$SSE = 3030.2$	7.05		
LOF	397	$SS_{LOF} = 2976$	7.5	4.567	9.255e-7
PE	33	$SS_{PE} = 54.167$	1.614		
$R^2 = 0.795$		$R^2_A = 0.791$		$R^2_{PRESS} = 0.784$	
$RMSE = 2.64$		$MAE = 2.08$			

The F-score value and corresponding p-value given in Table 4.15 indicate that the regression is significant with a high degree of probability. The R^2, R^2_A and R^2_{PRESS} indicates that the model covers about 79% of the variation in the data.
Looking at the coefficients for variables interacting with and including additive, see Table A.19 in the appendix, they are all of similar size. A simplification can be made, i.e.

$$b_5 \approx -b_{15} \approx -b_{25} \approx -b_{35} \approx -b_{45} = \bar{b}. \tag{4.15}$$

Using the approximation given in Equation (4.15) and the mixture constraint given in Equation (2.32) the model can be simplified to

$$\hat{y} = \sum_{i=1}^{4} b_i x_i + \bar{b} x_5^2 \tag{4.16}$$

4.5 Neural network

4.5.1 Network based properties in the product

The predictors used for these networks, given in Table 4.16 are the same as used in MLR model based on properties in the product, Equation (4.2). The data-set now consists of 485 samples.

Table 4.16: Independent variables used for training neural networks

$x_1 = CP$	$x_2 = CI$	$x_3 = 95Rec$	$x_4 = additive$

The networks were trained as described in method. The network with highest R^2 values of the 10 replications were then saved. A graph displaying the R^2 values of these networks against the number of nodes in the hidden layer can be found in Figure 4.2.

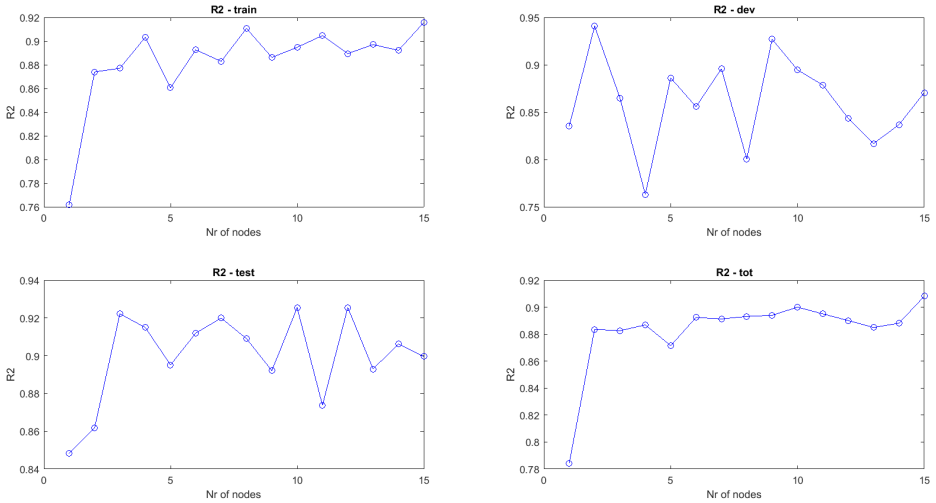

Figure 4.2: The figure shows the R^2 values for different network setup. Top left shows values for training set, top right shows values for validation (development) set, bottom left shows values for independent test set and bottom right shows values for the complete data set. Properties in the diesel product are used as predictors, Table 4.16.

Looking at Figure 4.2 the network with 10 hidden units is considered the best. This is because the R^2 values in all data-sets are both high and reasonable close to each other, with the independent test set having the highest R^2 value. Table 4.17 shows the RMSE, MAE and R^2 values for the network with 10 hidden nodes.

Table 4.17: RMSE, MAE and R^2 values for network with 10 hidden nodes. Predictors used can be found in Table 4.16.

	RMSE	MAE	R^2
Training set	1.952	1.371	0.895
Validation set	2.199	1.573	0.895
Test set	1.671	1.257	0.925
Complete set	1.952	1.384	0.900

The R^2 values for the test set and complete set are similar, with the independent test set having a higher value. This indicates that R^2 value for the complete set can be trusted, the model covers 90% of the variation in the data.
Comparing this model to the second order additive MLR model based on the same predictors (Equation (4.6)) and the corresponding ANOVA (Table 4.6), the RMSE, MAE and R^2 values are slightly better for the MLR model. The MLR model is also simpler, having only 6 coefficients.
In the MLR model 24 possible outliers have been removed from the data, which is not the case for the network model. A more suitable comparison might therefore be with the ANOVA for the MLR model when no outliers have been removed (Appendix Table A.6). In this case the network model preforms slightly better.

4.5.2 ANN models based on component fractions

For these networks the predictors are the fractions from each component tank used in the blends (Table 4.18), the same as used for the first order canonical mixture model (Equation (4.7)). The data-set contains 483 data points.

Table 4.18: Predictors used in ANN network for component fractions.

$x_1 = tank1$	$x_2 = tank2$	$x_3 = tank3$
$x_4 = tank4$	$x_5 = tank5$	$x_6 = tank6$
$x_7 = tank7$	$x_8 = tank8$	$x_9 = additive$

The networks were trained as described in Chapter 3. The network with highest R^2 values of the 10 replications were then saved. A graph displaying the R^2 values of these networks against the number of nodes in the hidden layer can be found in Figure 4.3.

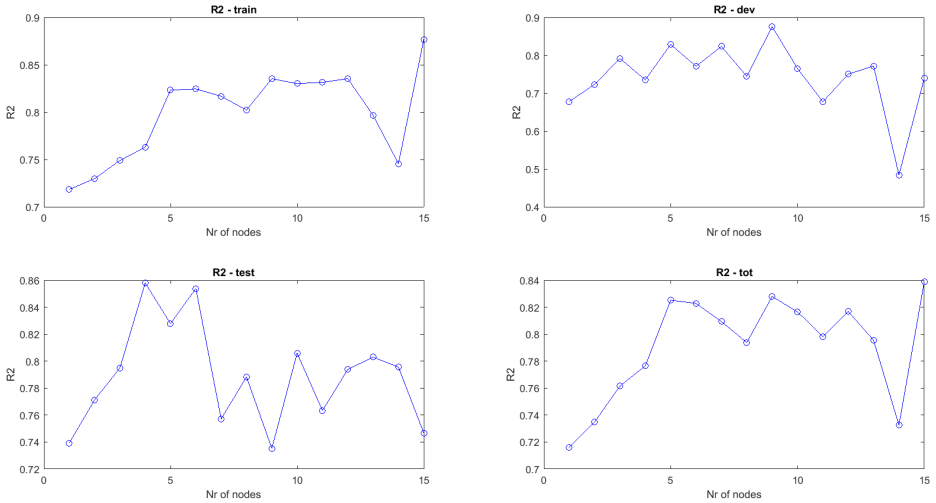

Figure 4.3: The figure shows the R^2 values for different network setup. Top left shows values for training set, top right shows values for validation (development) set, bottom left shows values for independent test set and bottom right shows values for the complete data set. Fractions of each tank are used as predictors, Table 4.18.

Looking at Figure 4.3, the network with 5 hidden units is considered the best. This is due to all R^2 values being comparably high and the number of nodes is low. Simpler is better, if given the choice. Table 4.19 shows the RMSE, MAE and R^2 values for the network with 5 hidden nodes.

Table 4.19: RMSE, MAE and R^2 values for network with 5 hidden nodes. Predictors used can be found in Table 4.18.

	RMSE	MAE	R^2
Training set	2.4947	1.9034	0.8233
Validation set	2.6260	2.1290	0.8286
Test set	2.7651	2.0048	0.8279
Complete set	2.5565	1.9522	0.8252

The R^2 values for the independent test set and the R^2 values for the complete set are similar, indicating that the R^2 value for the complete set can be trusted. Comparing this network with the result for the second order additive reduced mixture model given in Equation (4.12), and corresponding ANOVA given in Table 4.12. The network shows a slightly better performance in terms of RMSE, MAE and R^2. The MLR model is however much simpler, only containing 8 coefficients in total while the network model has 61.

In the case of the MLR model 14 possible outliers for the model has been removed. A more fair comparison might be to look at the ANOVA for the case when no outliers has been removed, appendix Table A.14. In this case the difference in performance is larger, with the network preforming better.

4.5.3 Neural networks based on simplified mixture components

These network models are based on the simplified mixture components used in Equation (4.13). The predictors are given in Table 4.20. The complete data set consists of 445 samples.

Table 4.20: Predictors used for neural network models.

$x_1 = G1$	$x_2 = G2$	$x_3 = G3$
$x_4 = G4$	$x_5 = additive$	

The networks were trained as described in method. The network with highest R^2 values of the 10 replications were then saved. A graph displaying the R^2 values of these networks against the number of nodes in the hidden layer can be found in Figure 4.4.

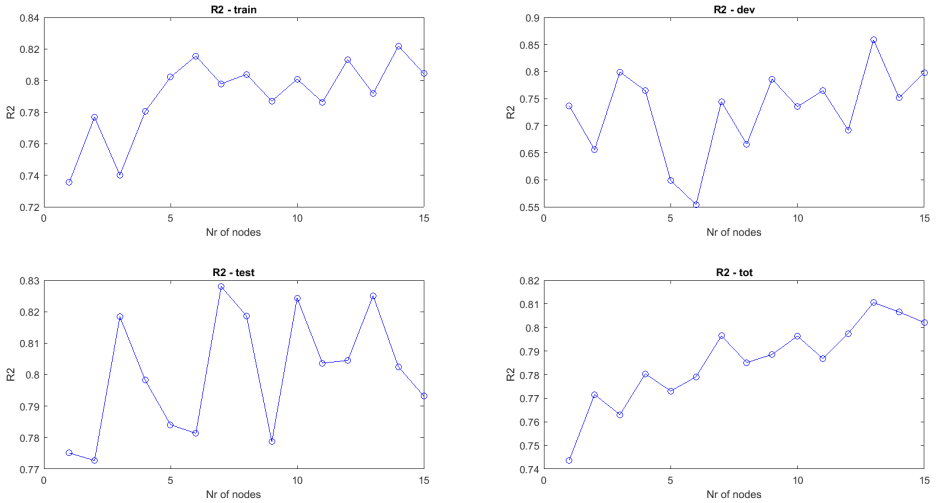

Figure 4.4: The figure shows the R^2 values for different network setup. Top left shows values for training set, top right shows values for validation (development) set, bottom left shows values for independent test set and bottom right shows values for the complete data set. Fractions of each tank used as predictors, Table 4.20.

Looking at Figure 4.2 the network with 13 hidden units is considered the best. This is because the R^2 values in all data-sets are both high and of reasonably close to each other. Table 4.21 shows the RMSE, MAE and R^2 values for the network with 13 hidden nodes.

Table 4.21: RMSE, MAE and R^2 values for network with 13 hidden nodes. Predictors used can be found in Table 4.20.

	RMSE	MAE	R^2
Training set	2.5932	1.9801	0.7917
Validation set	2.2687	1.7223	0.8586
Test set	2.7378	2.0342	0.8251
Complete set	2.5697	1.9494	0.8105

Looking at the R^2 values for the different sets given in Table 4.21, we see that the R^2 value for the independent test set is close to the R^2 value of the complete set, indicating that the R^2 value for the complete set can be trusted.
Comparing this network model with the simplified components mixture model with second order additive term based on same predictors (Equation (4.14)) and the corresponding ANOVA (Table 4.15), the RMSE, MAE and R^2 values are slightly better in the network model. The difference in parameters are rather large. With 13 hidden nodes the network model has 65 weights and 14 biases, compared to the MLR model which can be reduced to 5 coefficients. The ANOVA given in Table 4.15 is based on the case when 6 possible outliers have been removed. Another relevant comparison is therefore with the ANOVA for the MLR model when no outliers has been removed, given in Appendix Table A.18. The difference between the network

model and mixture model is slightly larger in this case with the network showing the best performance.
Comparing this network based on simplified components with the network based on normal mixture components given in the previous section, the network with normal mixture component preforms slightly better and has fewer nodes in the hidden layer. However, given the generality of the simplified components compared to the assumption that components stay the same in the case of normal mixture model the difference is not as large as one could expect.

4.6 MLR Model based only on additive

Given all the results so far it is reasonable to suspect that additive is by far the most influential predictor. To investigate further a model only based on additive is set up,

$$\hat{y} = \beta_0 + \beta_1 x + \beta_2 x^2, \tag{4.17}$$

were $x = additive$. ANOVA for this model is given in Table 4.22 for the case when no outliers have been removed and in Table 4.23 when 17 possible outliers have been removed.

Table 4.22: ANOVA table for model based only on additive given in Equation (4.17). No outliers removed from the data.

Variation	d.f.	Sum of squares	Mean square	F-score	p-value
Total	484	$SST = 18493$			
Regression	2	$SSR = 11417$	5708.5	388.84	2.83e-101
Residual	482	$SSE = 7076.2$	14.61		
LOF	19	$SS_{LOF} = 786.1$	41.38	3.056	1.94e-5
PE	463	$SS_{PE} = 6290.1$	13.59		
$R^2 = 0.617$		$R^2_A = 0.616$		$R^2_{PRESS} = 0.612$	
$RMSE = 3.83$		$MAE = 2.72$			

Table 4.23: ANOVA table for model based only on additive, given in Equation (4.17). 17 possible outliers removed from the data.

Variation	d.f.	Sum of squares	Mean square	F-score	p-value
Total	467	$SST = 15432$			
Regression	2	$SSR = 11084$	5542	592.7	1.24e-128
Residual	465	$SSE = 4348$	9.35		
LOF	19	$SS_{LOF} = 772.63$	9.35	5.08	5.21e-11
PE	446	$SS_{PE} = 3575.2$	8.02		
$R^2 = 0.718$		$R^2_A = 0.717$		$R^2_{PRESS} = 0.714$	
$RMSE = 3.06$		$MAE = 2.37$			

Looking at the ANOVA for these two it is reasonable to assume that additive is by far the most influential predictor. Comparing the R^2 value of this model for R^2 values based on mixture models, both reduced and simplified components, it seems that the additional predictors in the mixture models does not have a huge impact. This might be an explanation to the fact that the models based on reduced components and models based on simplified components do not differ much in performance.
Comparing this model to that based on properties in the product, there is a significant impact when the other predictors are added. This is likely, in large, because of the CP term included in those models. By definition the cloud point in diesel is the temperature when crystals start to appear, while CFPP is the temperature when the crystals clog a filter, so CFPP must always happen after CP.

5

Conclusions and Discussion

First of all, it is important to point out that any statistical model is based on the data used to derive it. Any conclusions drawn based on the data used in this project might not be true for another set of data. However, since the data used in this case comes from analysis of actual laboratory test on regular blends done at Preem, it is fair to say that the conclusions regarding the data should be applicable for this given process.

A general conclusion is that the additive is by far the most influential parameter, almost surprisingly so given that it is only included in the blends at PPM levels.

The goal of creating an observer model based on other properties in the diesel product was the most successful one. The model given in Equation (4.6) has $R^2 = 0.93$, $RMSE = 1.58$ and $MAE = 1.23$ in the case when 26 potential outliers are removed and $R^2 = 0.88$, $RMSE = 2.18$ and $MAE = 1.54$ when none have been removed. Given that measuring CFPP can take over an hour, a quicker estimate based on this model could be useful. The reason for this model being more appropriate as an observer is that the predictors are other properties (not CFPP) in the *product*, not based on properties in the individual components (tanks).

The attempts to derive a model which could be used for optimisation or in a control algorithm were less successful. The best model for this purpose were the second order additive reduced mixture model given in Equation (4.12) with an $R^2 = 0.81$, $RMSE = 2.59$ and $MAE = 2.01$ in the case when 14 potential outliers are removed and $R^2 = 0.74$, $RMSE = 3.07$ and $MAE = 2.44$ with nothing removed. Neither the case with potential outliers removed or the one with none removed shows a promising model and is probably not good enough to use in a control algorithm.

The algorithm used to remove potential outliers is likely too aggressive. As most 26 potential outliers were detected and removed, in the model mentioned above. In a set of 485 blends this represents a bit over 5%. While a few could be expected, over 5% is a bit excessive. Another way of looking at it is that for the region spanned by the remaining 95 % there is a well preforming model which can be utilised. There is also the potential for someone with a deeper chemical understanding to study the difference between what was removed and what was kept.

The grouping of components based on CFPP was somewhat successful. The performance of the models based on the simplified components is comparable to the performance of the reduced mixture models. The simplified component setup is more flexible and less potential outliers were found for these models. What should not be forgotten is that 40 blends were removed when regressing models based on the simplified components, and the reason why comparably few outliers were found for these models might be because "difficult" blends were removed. Looking further

into a more advanced system of dividing the components into groups could be of interest.
The neural networks models performed better in general compared to the MLR models, given that no outliers were removed for these models. It is possible that the performance could be improved further by additional nodes in the hidden layer. The question is if it is possible/desirable to implement such a model in a control algorithm. Depending on the software used this might be a problem because of the more complicated terms in the neural network model compared to the MLR models.
During the project other sets of predictors have been investigated in the attempt to find a better model, but with equal or less success to the results given in this report. Given this set of data it seems unlikely that the "perfect model" can be found. Some additional predictor(s), containing new information, are probably needed.

Bibliography

[1] Julio Cesar L Alves, Claudete B Henriques, and Ronei J Poppi. "Determination of diesel quality parameters using support vector regression and near infrared spectroscopy for an in-line blending optimizer system". In: *Fuel* 97 (2012), pp. 710–717.

[2] Kazimierz Baczewski, Piotr Szczawiński, and Milena Kamińska. "Experimental testing of influence of commercial depressants on diesel fuels low temperature properties". In: *Journal of KONES* 22.1 (2015), pp. 7–14.

[3] John A. Cornell. *Experiments with Mixtures*. Third Edition. Wiley series in probability and statistics. Wiley-Interscience, 2002. ISBN: 0-471-39367-3.

[4] Christophe Leys et al. "Detecting outliers: Do not use standard deviation around the mean, use absolute deviation around the median". In: *Journal of Experimental Social Psychology* 49.4 (2013), pp. 764–766.

[5] H.A. Cho M.A. Goldberg. *Introduction to Regression Analysis*. WIT Press, 2004. ISBN: 1-85312-624-1.

[6] Arpit S Maheshwari and Jitendra G Chellani. "Correlations for Pour Point and Cloud Point of middle and heavy distillates using density and distillation temperatures". In: *Fuel* 98 (2012), pp. 55–60.

[7] Robert E. Maples and Knovel (e-book collection). *Petroleum refinery process economics*. English. 2nd. Tulsa, Okla: PennWell Corp, 2000;2003;

[8] Fátima IC Mirante and João AP Coutinho. "Cloud point prediction of fuels and fuel blends". In: *Fluid Phase Equilibria* 180.1-2 (2001), pp. 247–255.

[9] Andrew Ng. *Deep Neural Networks (DNNs) by Andrew Ng [full course]*. Online course about Deep Neural Networks given by Andrew NG, adjunct professor at Stanford University. URL: `https://www.youtube.com/watch?v=7PiK4wtfvbA&list=PLBAGcD3siRDguyYYzhVwZ3tLvOyyG5k6K`. (accessed: 02.02.2018).

[10] N Pasadakis, S Sourligas, and Ch Foteinopoulos. "Prediction of the distillation profile and cold properties of diesel fuels using mid-IR spectroscopy and neural networks". In: *Fuel* 85.7-8 (2006), pp. 1131–1137.

[11] Sakorn Saiban and Trevor C Brown. "Kinetic model for cloud-point blending of diesel fuels". In: *Fuel* 76.14-15 (1997), pp. 1417–1423.

[12] Prem B Semwal and Ram G Varshney. "Predictions of pour, cloud and cold filter plugging point for future diesel fuels with application to diesel blending models". In: *Fuel* 74.3 (1995), pp. 437–444.

[13] Filiz Al-Shanableh, Evcil Ali, and Mahmut Ahsen Savaş. “Fuzzy logic model for prediction of cold filter plugging point of biodiesel from various feedstock”. In: *Procedia Computer Science* 120 (2017), pp. 245–252.

[14] Filiz Al-Shanableh, Ali Evcil, and Mahmut Ahsen Savaş. “Prediction of cold flow properties of biodiesel fuel using artificial neural network”. In: *Procedia Computer Science* 102 (2016), pp. 273–280.

[15] *Standard Specification for Diesel Fuel Oils*. URL: `https://www.astm.org/Standards/D975.htm`. (accessed: 27.05.2018).

[16] *Standard Test Method for Cold Filter Plugging Point of Diesel and Heating Fuels*. URL: `https://www.astm.org/Standards/D6371.htm`. (accessed: 27.05.2018).

[17] *VISCOPLEX®cold flow improvers (CFIs)*. URL: `http://oil-additives.evonik.com/product/oil-additives/Downloads/cfi-brochure-rebrand-en-v10-interactive.pdf`. (accessed: 28.05.2018).

[18] Li Weimin, Li Wenkai, and Chi-Wai Hui. “Integrating neural network models for refinery planning”. In: *Computer Aided Chemical Engineering*. Vol. 15. Elsevier, 2003, pp. 1304–1309.

[19] Chuanjie Wu et al. “Artificial neural network model to predict cold filter plugging point of blended diesel fuels”. In: *Fuel processing technology* 87.7 (2006), pp. 585–590.

[20] Xin Yan, Xiaogang Su, and Ebook Central (e-book collection). *Linear regression analysis: theory and computing*. English. Hackensack, NJ;Singapore; World Scientific, 2009;2014; ISBN: 9789812834102;9812834109;

A

Appendix 1

A.1 Models based on properties in the diesel product

A.1.1 First order linear model based on properties in the diesel product

Table A.1: Coefficient statistics for first order linear model based on properties in the diesel product. Model given in Equation (4.2) and independent variables given in 4.2. No outliers removed

	Estimate	SE	tStat	pValue
(Intercept)	61.9346	18.9175	3.2739	0.0011
x1	1.1630	0.0768	15.1495	0.0000
x2	0.7061	0.1250	5.6470	0.0000
x3	-0.2856	0.0504	-5.6712	0.0000
x4	-0.0235	0.0007	-31.6853	0.0000

Table A.2: Coefficient statistics for first order linear model based on properties in the diesel product. Model given in Equation (4.2) and independent variables given in 4.2. 15 potential outliers removed.

	Estimate	SE	tStat	pValue
(Intercept)	86.0190	16.5720	5.1906	0.0000
x1	1.0082	0.0699	14.4161	0.0000
x2	0.9454	0.1089	8.6835	0.0000
x3	-0.3884	0.0446	-8.7118	0.0000
x4	-0.0271	0.0007	-39.7902	0.0000

A.1.2 Second order model based on properties in the diesel fuel product

Table A.3: Coefficient statistics for second order polynomial model based on properties in the diesel product. Model given in Equation (4.4) and independent variables given in 4.2. No outliers removed

	Estimate	SE	tStat	pValue
(Intercept)	-1720.3540	967.1779	-1.7787	0.0760
x1	-1.3248	4.7786	-0.2772	0.7817
x2	31.1687	8.8893	3.5063	0.0005
x3	4.9291	5.1289	0.9610	0.3371
x4	0.3556	0.0551	6.4562	0.0000
x1:x1	-0.0078	0.0112	-0.6975	0.4858
x1:x2	0.1100	0.0330	3.3357	0.0009
x1:x3	-0.0103	0.0134	-0.7662	0.4440
x1:x4	0.0014	0.0002	5.8730	0.0000
x2:x2	0.0174	0.0449	0.3870	0.6989
x2:x3	-0.0909	0.0253	-3.5921	0.0004
x2:x4	0.0016	0.0004	3.8492	0.0001
x3:x3	-0.0002	0.0071	-0.0226	0.9820
x3:x4	-0.0014	0.0001	-9.2076	0.0000
x4:x4	4.23e-05	0.0000	25.9757	0.0000

A.1.3 Second order additive terms model based on properties in the diesel fuel product

Table A.4: ANOVA table Second order additive terms model based on properties in the diesel fuel product. Model given in Equation (4.5) and predictors given in Table 4.2. No outliers removed.

Variation	d.f.	Sum of squares	Mean square	F-score	p-value
Total	484	$SST = 18493$			
Regression	8	$SSR = 16241$	2030.2	429.17	3.50e-212
Residual	476	$SSE = 2251.7$	4.73		
LOF	475	$SS_{LOF} = 2251.7$	4.74	inf	0
PE	1	$SS_{PE} = 0$	0		
$R^2 = 0.878$		$R^2_A = 0.876$		$R^2_{PRESS} = 0.872$	
$RMSE = 2.17$		$MAE = 1.54$			

Table A.5: Coefficient statistics for second order additive model based on properties in the diesel product. Model given in Equation (4.5) and independent variables given in 4.2. 22 possible outliers removed from the data.

	Estimate	SE	tStat	pValue
(Intercept)	-3.7412	18.1926	-0.2056	0.8372
x1	0.9249	0.0650	14.2231	0.0000
x2	0.2624	0.1274	2.0594	0.0400
x3	-0.0305	0.0527	-0.5788	0.5630
x4	0.2920	0.0554	5.2754	0.0000
x1:x4	0.0016	0.0002	7.5458	0.0000
x2:x4	0.0013	0.0004	3.4270	0.0007
x3:x4	-0.0011	0.0001	-7.6840	0.0000
x4:x4	4.188-05	0.0000	25.1883	0.0000

A.1.4 Reduced second order additive terms model based on properties in the diesel fuel product

Table A.6: ANOVA table for reduced second order additive model given in Equation (4.6) based on predictors given in Table 4.2. No outliers removed from the data

Variation	d.f.	Sum of squares	Mean square	F-score	p-value
Total	485	$SST = 19348$			
Regression	6	$SSR = 17066$	2844.3.2	596.93	1.09e-218
Residual	479	$SSE = 2282.4$	4.9093		
LOF	414	$SS_{LOF} = 2155.2$	4.909	1.544	0.046
PE	40	$SS_{PE} = 127.21$	3.1802		
$R^2 = 0.877$		$R^2_A = 0.875$		$R^2_{PRESS} = 0.872$	
$RMSE = 2.18$		$MAE = 1.54$			

Table A.7: Coefficients and t-statistics for reduced second order additive model given in Equation (4.6) based on predictors given in Table 4.2 with 26 possible outliers removed.

	Estimate	SE	tStat	pValue
x1	1.0537	0.0301	35.0624	0.0000
x4	0.2922	0.0335	8.7329	0.0000
x1:x4	0.0010	0.0002	6.1610	0.0000
x2:x4	0.0023	0.0002	10.1872	0.0000
x3:x4	-0.0013	0.0001	-15.6263	0.0000
x4:x4	4.0989e-05	0.0000	28.5509	0.0000

A.2 Mixture models

A.2.1 First order mixture model

Table A.8: ANOVA table for first order mixture model given in Equation (4.7) based on predictors given in Table 4.7. No outliers removed from the data

Variation	d.f.	Sum of squares	Mean square	F-score	p-value
Total	482	$SST = 18055$			
Regression	9	$SSR = 12291$	1365.7	112.32	1.41e-111
Residual	474	$SSE = 5763.5$	12.159		
LOF	438	$SS_{LOF} = 5706.8$	13.03	8.278	3.99e-11
PE	36	$SS_{PE} = 56.67$	1.57		
$R^2 = 0.681$		$R^2_A = 0.674$		$R^2_{PRESS} = 0.66$	
$RMSE = 3.49$		$MAE = 2.61$			

Table A.9: Coefficients and t-statistics for first order mixture model given in Equation (4.7) with predictors given in Table 4.7. 9 possible outliers removed.

	Estimate	SE	tStat	pValue
x1	-3.6836	2.6388	-1.3959	0.1634
x2	-4.8698	1.9903	-2.4468	0.0148
x3	-6.4797	1.0051	-6.4471	0.0000
x4	-8.4947	0.9913	-8.5690	0.0000
x5	3.9855	1.0509	3.7923	0.0002
x6	-2.7039	1.5492	-1.7453	0.0816
x7	-12.0108	2.3312	-5.1521	0.0000
x8	-27.4492	2.6334	-10.4234	0.0000
x9	-26023.3445	751.0808	-34.6479	0.0000

A.2.2 Reduced first order mixture model

Table A.10: ANOVA table for the reduced first order mixture model given in Equation (4.10) based on predictors given in Table 4.7. No outliers removed from the data.

Variation	d.f.	Sum of squares	Mean square	F-score	p-value
Total	482	$SST = 18055$			
Regression	6	$SSR = 12265$	2044.1	168.02	3.74e-114
Residual	476	$SSE = 5790.1$	12.16		
LOF	440	$SS_{LOF} = 5733.4$	13.03	8.278	3.99e-11
PE	36	$SS_{PE} = 56.67$	1.57		
$R^2 = 0.679$		$R^2_A = 0.675$		$R^2_{PRESS} = 0.665$	
$RMSE = 3.49$		$MAE = 2.60$			

Table A.11: Coefficients and t-statistics for the reduced first order mixture model given in Equation (4.10) with predictors given in Table 4.7. 8 possible outliers removed.

	Estimate	SE	tStat	pValue
(Intercept)	-3.6174	1.1060	-3.2707	0.0012
x1	-2.8480	1.4411	-1.9763	0.0487
x2	-4.9142	1.4314	-3.4331	0.0006
x3	7.4369	1.3214	5.6282	0.0000
x4	-8.0393	2.7999	-2.8713	0.0043
x5	-23.5795	3.0790	-7.6581	0.0000
x6	-26042.6346	749.7306	-34.7360	0.0000

A.2.3 Second order polynomial reduced mixture model

Table A.12: ANOVA table for the second order polynomial reduced mixture model given in Equation (4.11) based on predictors given in Table 4.9. No outliers removed from the data.

Variation	d.f.	Sum of squares	Mean square	F-score	p-value
Total	482	$SST = 18055$			
Regression	27	$SSR = 14166$	524.7	61.4	2.4e-133
Residual	455	$SSE = 3888.4$	8.5		
LOF	419	$SS_{LOF} = 3831.7$	9.1	5.8	9.4e-9
PE	36	$SS_{PE} = 56.67$	1.57		
$R^2 = 0.785$		$R^2_A = 0.772$		$R^2_{PRESS} = 0.745$	
$RMSE = 2.92$		$MAE = 2.16$			

Table A.13: Coefficients and t-statistics for the second order polynomial reduced mixture model given in Equation (4.11) based on predictors given in Table 4.9. 9 possible outliers have been removed from the data.

	Estimate	SE	tStat	pValue
(Intercept)	-10.5648	3.5510	-2.9752	0.0031
x1	12.0207	19.1026	0.6293	0.5295
x2	27.6351	11.6525	2.3716	0.0181
x3	-0.2774	9.4778	-0.0293	0.9767
x4	20.2683	18.0569	1.1225	0.2623
x5	30.8612	19.5315	1.5801	0.1148
x6	-46417.1363	5256.8061	-8.8299	0.0000
x1:x1	-9.6216	18.3710	-0.5237	0.6007
x1:x2	-46.2092	25.4033	-1.8190	0.0696
x1:x3	18.7066	24.4176	0.7661	0.4440
x1:x4	-61.3237	31.7796	-1.9297	0.0543
x1:x5	-26.3497	37.9431	-0.6945	0.4878
x1:x6	6848.8649	6878.1764	0.9957	0.3199
x2:x2	-24.2570	9.7601	-2.4853	0.0133
x2:x3	-21.2303	14.5252	-1.4616	0.1446
x2:x4	-21.4156	23.0795	-0.9279	0.3540
x2:x5	-80.7694	24.2254	-3.3341	0.0009
x2:x6	-2338.6010	7635.8656	-0.3063	0.7595
x3:x3	30.7913	8.5996	3.5805	0.0004
x3:x4	-38.6708	21.8795	-1.7674	0.0778
x3:x5	-18.7579	26.1585	-0.7171	0.4737
x3:x6	5897.6053	6769.1274	0.8713	0.3841
x4:x4	-31.8407	29.9456	-1.0633	0.2882
x4:x5	-45.9485	39.3019	-1.1691	0.2430
x4:x6	25012.8600	13075.3150	1.9130	0.0564
x5:x5	-45.8858	34.3070	-1.3375	0.1817
x5:x6	-30448.8207	13296.9357	-2.2899	0.0225
x6:x6	32396870.3958	2669515.1223	12.1359	0.0000

A.2.4 Second order additive reduced mixture model

Table A.14: ANOVA table for the second order additive reduced mixture model given in Equation (4.12) based on predictors given in Table 4.9. No outliers removed from the data.

Variation	d.f.	Sum of squares	Mean square	F-score	p-value
Total	482	$SST = 18055$			
Regression	7	$SSR = 13307$	1901.1	190.22	2.05e-133
Residual	475	$SSE = 4747.3$	9.99		
LOF	439	$SS_{LOF} = 4690.6$	10.7	6.79	8.78e-10
PE	36	56.67	1.57		
$R^2 = 0.737$		$R^2_A = 0.733$		$R^2_{PRESS} = 0.724$	
$RMSE = 3.16$		$MAE = 2.27$			

Table A.15: Coefficients and t-statistics for the second order additive reduced mixture model given in Equation (4.12) based on predictors given in Table 4.9. 14 possible outliers have been removed from the data.

	Estimate	SE	tStat	pValue
(Intercept)	-2.0668	0.9565	-2.1609	0.0312
x1	-3.7289	1.2278	-3.0369	0.0025
x2	-5.4860	1.2345	-4.4440	0.0000
x3	8.2475	1.1383	7.2453	0.0000
x4	-9.0567	2.3940	-3.7831	0.0002
x5	-20.6613	2.6581	-7.7730	0.0000
x6	-46693.6334	1769.1894	-26.3927	0.0000
x7	33725851.3910	2590273.6315	13.0202	0.0000

A.3 Simplified components

A.3.1 First order simplified components mixture model

Table A.16: Coefficient statistics for simple component linear mixture model given in Equation 4.13. 3 possible outliers removed from the data.

	Estimate	SE	tStat	pValue
x1	-19.7830	1.8992	-10.4165	0.0000
x2	-6.8607	1.2616	-5.4380	0.0000
x3	-8.2822	0.8276	-10.0070	0.0000
x4	4.1923	1.0893	3.8486	0.0001
x5	-24958.0675	792.2473	-31.5029	0.0000

Table A.17: ANOVA table for simple component linear mixture model given in Equation 4.13. based on predictors given in Table 4.13. No outliers removed from the data.

Variation	d.f.	Sum of squares	Mean square	F-score	p-value
Total	445	$SST = 15510$			
Regression	5	$SSR = 10646$	2129.2	192.63	2.22e-108
Residual	440	$SSE = 4863.5$	11.05		
LOF	407	$SS_{LOF} = 4809.3$	11.820	7.2	1.87e-9
PE	33	$SS_{PE} = 54.17$	1.64		
$R^2 = 0.687$		$R^2_A = 0.684$		$R^2_{PRESS} = 0.678$	
$RMSE = 3.32$		$MAE = 2.65$			

A.3.2 Second order additives simplified mixture model

Table A.18: ANOVA table for second order additive simplified components mixture model given in Equation (4.14). Predictors given in Table 4.13. No outliers removed from the data.

Variation	d.f.	Sum of squares	Mean square	F-score	p-value
Total	445	$SST = 15510$			
Regression	5	$SSR = 11933$	1325.9	161.62	7.05e-133
Residual	436	$SSE = 3576.8$	8.2		
LOF	403	$SS_{LOF} = 3522.6$	8.74	5.33	1.22e-7
PE	33	54.17	1.64		
$R^2 = 0.770$		$R^2_A = 0.765$		$R^2_{PRESS} = 0.756$	
$RMSE = 2.86$		$MAE = 2.18$			

Table A.19: Coefficient statistics for second order additive simplified components mixture model given in Equation (4.14). Predictors given in Table 4.13. 6 possible outliers removed from the data.

	Estimate	SE	tStat	pValue
x1	-14.9126	2.3949	-6.2268	0.0000
x2	-8.4837	2.0534	-4.1316	0.0000
x3	-6.8318	1.0182	-6.7097	0.0000
x4	5.8877	1.3604	4.3278	0.0000
x5	36118128.0895	2770924.4882	13.0347	0.0000
x1:x5	-36185944.2049	2772523.0974	-13.0516	0.0000
x2:x5	-36158500.0206	2773049.3000	-13.0393	0.0000
x3:x5	-36171803.8897	2772721.3172	-13.0456	0.0000
x4:x5	-36150330.9086	2772652.6287	-13.0382	0.0000

Table A.20: Coefficients and t statistics for model based only on additive given in Equation (4.17). 17 possible outliers removed from the data.

	Estimate	SE	tStat	pValue
(Intercept)	-5.0234	0.2971	-16.9073	0.0000
x1	-0.0423	0.0020	-20.8742	0.0000
x2	2.87e-05	0.0000	9.6739	0.0000

Table A.21: Coefficients and t statistics for model based only on additive given in Equation (4.17). No outliers removed from the data.

	Estimate	SE	tStat	pValue
(Intercept)	-5.1910	0.3696	-14.0455	0.0000
x1	-0.0419	0.0025	-16.6982	0.0000
x2	2.83e-05	0.0000	7.6890	0.0000

A.4 Relevant matlab code

A.4.1 MLR

```
%% General MLR fitting
% X is a (N,m) matrix, N= #samples and m= #predictors
% Y is a (N,1) matrix containg CFPP values
Model1m=fitlm(X,Y)

%% Mixture models
Model1m=fitlm(X,Y,'intercept',false)
%% ANOVA
anova(Model1m,'summary')

%% PRESS

H=X*inv(X'*X)*X'; %Hat matrix
E=Model1m.Residuals.Raw;
PRESS=0;
for i=1:length(E)
    PRESS=PRESS+(E(i)/(1-H(i,i)))^2;
end
R2_PRESS=1-PRESS/Model1m.SST


%% outlier detection/removeal
asd=1;
while asd==1
    keep=ones(size(Y));
```

```
    Model1m=fitlm(X,Y);
    outliers=isoutlier(Model1m.Residuals.raw)
    keep(outliers)=0;
    X_temp=X(find(keep==1),:);
    Y=Y(find(keep==1));
    X=X_temp;
    if length(X)==length(keep)
        asd=0;
    end
end

function [ outliers ] = isoutlier( E )
% Detects outliers based on Median absolute derivation
% There is a actual matlab function for this, with the same name.
%But I have a older version of matlab were it is not included.

    A=median(E);
    b=1.4826;
    MAE=b*median(abs(E-A));
    outliers=find(E<A-3*MAE|E>A+3*MAE);
end
```

A.4.2 Neural networks

```
% X is a (m,N) matrix, N= #samples and m= #predictors
% Y is a (1,N) matrix containg CFPP values

Node=15;% number of nodes
rep=10; % number of repetitions
Networks= ShallowANN( X,Y,Node,rep );



function [ Tot_Networks ] = ShallowANN( X,Y,N,rep )
    %  X - input data.
    %  Y - target data.
    %   N - Max nr layers
    %   rep - repeated tries on each network
    Tot_Networks=cell(1,N);

for ii=1:N
    Networks=cell(1,rep);
    for jj=1:rep


%   X - input data.
%   Y - target data.

X
```

```
x = X;
t = Y;
%
% Choose a Training Function
% For a list of all training functions type: help nntrain
% 'trainlm' is usually fastest.
% 'trainbr' takes longer but may be better for challenging problems.
% 'trainscg' uses less memory. Suitable in low memory situations.

trainFcn = 'trainlm';  % Levenberg-Marquardt backpropagation.
%
% Create a Fitting Network
hiddenLayerSize = ii;
net = fitnet(hiddenLayerSize,trainFcn);
%
% Setup Division of Data for Training, Validation, Testing
net.divideParam.trainRatio = 70/100;
net.divideParam.valRatio = 15/100;
net.divideParam.testRatio = 15/100;
% Choose a Performance Function
% For a list of all performance functions type: help nnperformance
net.performFcn = 'mse';  % Mean Squared Error
% Train the Network
[net,tr] = train(net,x,t);
%
% Test the Network
y = net(x);
e = gsubtract(t,y);
performance = perform(net,t,y);

%===RMSE/MAE/R2=====

% ===Train===
e_train=e((0==isnan(tr.trainMask{1})));
t_train=t((0==isnan(tr.trainMask{1})));
RMSE(1)=sqrt(sumsqr(e_train)/length(e_train));
MAE(1)=sum(abs(e_train))/length(e_train);
SS_tot_train=sumsqr(t_train-mean(t_train));
R2(1)=1-sumsqr(e_train)/SS_tot_train;


%=== Dev ===
e_dev=e((0==isnan(tr.valMask{1})));
t_dev=t((0==isnan(tr.valMask{1})));
% RMSE/MAE
```

```
RMSE(2)=sqrt(sumsqr(e_dev)/length(e_dev));
MAE(2)=sum(abs(e_dev))/length(e_dev);
SS_tot_dev=sumsqr(t_dev-mean(t_dev));
R2(2)=1-sumsqr(e_dev)/SS_tot_dev;

% ==== Test ====
e_test=e((0==isnan(tr.testMask{1})));
t_test=t((0==isnan(tr.testMask{1})));
% RMSE/MAE
RMSE(3)=sqrt(sumsqr(e_test)/length(e_test));
MAE(3)=sum(abs(e_test))/length(e_test);
SS_tot_test=sumsqr(t_test-mean(t_test));
R2(3)=1-sumsqr(e_test)/SS_tot_test;

%==== Total====
RMSE=zeros(4,1);
MAE=zeros(4,1);
R2=zeros(4,1);
% RMSE
RMSE(4)=sqrt(sumsqr(e)/length(e));
MAE(4)=sum(abs(e))/length(e);
SS_tot=sumsqr(t-mean(t));
R2(4)=1-sumsqr(e)/SS_tot;


%
Table_1=table(RMSE,MAE,R2,'VariableNames',{'RMSE','MAE','R2'},'RowNames',

Networks{jj}=struct();
Networks{jj}.net=net;
Networks{jj}.tr=tr;
Networks{jj}.table=Table_1;
    end
    Tot_Networks{ii}=Networks;
end



end
```