Engine Out NO$_X$ Observer

Fusion of Models and Sensor Data for Estimation of NO$_X$ Flow

Systems, Control and Mechatronics

JOHAN ERIKSSON
ALBIN HESSLERYD

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

Engine Out NO$_\mathrm{x}$ Observer

Fusion of Models and Sensor Data for Estimation of NO$_x$ Flow

JOHAN ERIKSSON
ALBIN HESSLERYD

CHALMERS
UNIVERSITY OF TECHNOLOGY

Fusion of Models and Sensor Data for Estimation of NO$_x$ Flow

JOHAN ERIKSSON
ALBIN HESSLERYD

Cover: A Volvo truck diesel engine and after treatment system.

Typeset in LaTeX
Gothenburg, Sweden 2018

Engine Out NO$_x$ Observer
Fusion of Models and Sensor Data for Estimation of NO$_x$ Flow
JOHAN ERIKSSON
ALBIN HESSLERYD
Department of Electrical Engineering
Chalmers University of Technology

Abstract

Transportation of goods is a very important task in today's world where trucks have a central role. The level of emissions allowed for vehicles are regulated by law. Over time these levels are decreased which call for a continuous technology development regarding the efficiency for the after treatment system. One of the species that is strictly regulated is NO$_x$ which today is measured with a sensor in the after treatment system. This NO$_x$ sensor have several limitations and the purpose of this thesis is to design an observer that improves the estimation of the NO$_x$ flow in the after treatment system rather than only use the NO$_x$ sensor reading.

Five different observers have been developed, a fraction estimator, a delayed Kalman filter, an unknown input observer, an steady state unknown input observer, and an extended state vector observer. It was shown that all five observers have quicker transient behaviour than the sensor. The delayed Kalman filter had the absolute highest computational effort. The unknown input observer had the best overall performance on all tested engines in terms of MSE.

The steady state unknown input observer was also implemented and tested in TargetLink at Volvo. This decision was based on that when taking both accuracy and calculation time in to consideration was this the best observer. Therefore the steady state unknown input observer is the observer which the authors suggest Volvo should use to improve the engine out NO$_x$ estimation.

Acknowledgements

Thanks to our examinator Tomas McKelvey for guiding us through the thesis, making sure we were heading the right track and thanks to Johan Dahl, our supervisor at Volvo who has given us very useful tips and helped us with our questions. Also thanks to EATS control team at Volvo for letting us have the opportunity to do our thesis at their office and giving us support during the process.

Contents

Contents

List of Figures

List of Tables

List of Tables

Nomenclature

List of Variables

AFR Air fuel ratio

C_{corr} Correction factor based on cylinder wall temperature

C_{cyl} Cylinder temperature correction

$C_{\omega\tau}$ Engine speed and torque correction

d Sensor delay

D Difference between any model and Horiba

D_{SP} Difference between semiphysical model and Horiba

D_{RB} Difference between radial basis model and Horiba

$\widehat{D}$ Estimated difference between any model and Horiba

$\widehat{D}_{SP}$ Estimated difference between semiphysical model and Horiba

$\widehat{D}_{RB}$ Estimated difference between radial basis model and Horiba

E Accuracy function for any model

E_S Accuracy function for the sensor

E_{SP} Accuracy function for semiphysical model

E_{RB} Accuracy function for radial basis model

H_N Horiba

M Any model

M_{SP} Semiphysical model

M_{RB} Radial basis model

P_R Rail pressure

$\tilde{P}_R$ Steady state value on the rail pressure

S NO$_x$ sensor

T_B Boost temperature

$\tilde{T}_B$ Steady state value of the boost temperature

α_{EGR} EGR-valve position

γ Slack variable for fraction estimator

ζ Variable describing which model that is most trustworthy

$\widehat{\zeta}$ Variable describing which developed model that is most trustworthy

μ Mean value

ξ Tuning variable for fraction estimator in unknown input observer

σ Covariance values

τ_{Eng} Engine torque calculated by the engine model

ψ Tuning variable for fraction estimator in unknown input observer

ω_{Eng} Engine speed

$\tilde{\omega}_{Eng}$ Steady state value of the engine speed

Acronyms

CO	Carbon monoxide
DKF	Delayed Kalman filter
EATS	Engine after treatment system
ECU	Electrical control unit
EO	Engine out
EGR	Exhaust gas recirculation
eSCR	Extended selective catalyst reduction engine without EGR cooler
ESV	Extended state vector observer
FE	Fraction estimator
KF	Linear Kalman filter
MSE	Mean squared error
NH_3	Ammonia
NRMSE	Normalized root mean square
NO_x	Nitrogen oxide
rpm	Revolutions per minute
SCR	Selective catalytic reduction
SO	System out
SSUIO	Steady state solution for unknown input observer
TC	Turbo compound, heat recovery device
UIO	Unknown input observer

1
Introduction

Transportation of goods have since the beginning of the civilization been an important task where the transportation vehicle have been of large importance. In modern times, trucks have been extensively used as transportation vehicles and most of them are powered by diesel engines [1]. Diesel engines produces, among other things, both carbon monoxide (CO), different kinds of nitrogen oxides (NO_x), and soot [2]. These emissions are harmful for both humans [3] and the environment [4], and are therefore constrained by legislations for example in EU [5]. These regulations are continuously getting more and more strict which increases the requirements on the truck manufacturers [6]. The EU regulation of these emissions are named Euro I-VI (Euro 1-6) and are shown in Figure 1.1 where it can be seen that the regulations are getting tighter for each generation of legislation [7, 8, 9, 10, 11].
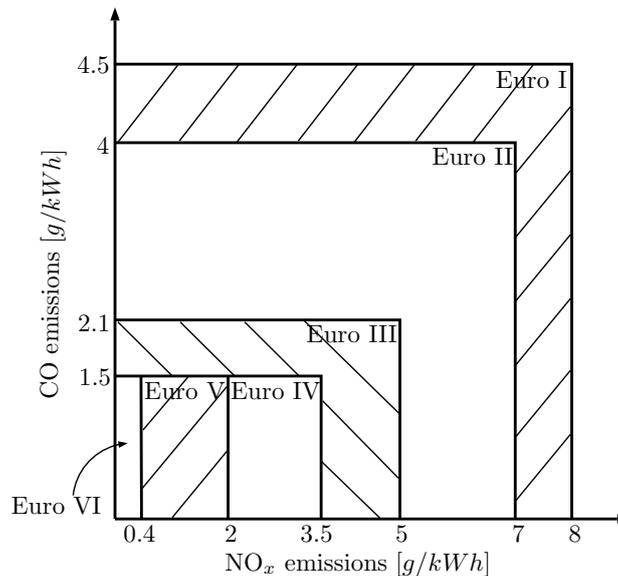
Figure 1.1: The allowed emissions of CO and NO_x from Euro I to VI.

To remove most of the NO_x gases an engine after treatment system (EATS) is used. EATS uses a selective catalyst reduction (SCR) to reduce the tail pipe NO_x emissions. The SCR makes a chemical reaction such that most of the NO_x gases are removed but it requires that portions of urea are added to these gases [12]. This process is described further in Section 1.1.2.

In the SCR, the urea dosing is crucial since underdosing can result in worse performance and increased emitted NO_x [13]. If, on the other hand, the amount of

urea is overdosed there is a risk that ammonia (NH_3) slips through the SCR [14], which is also harmful for both humans and the environment [15]. It should therefore be kept low.

To be able to dose the correct amount of urea to the NO_x gases, the system must have knowledge about the current NO_x flow. The common practice to estimate NO_x flow in the truck industry is by having a NO_x sensor that measures the amount of NO_x flow before the EATS. However, sensors of this kind have several issues where sensor delay is one of the most severe. By designing an observer that takes the sensor delay into account, and combines this with other real time estimates of the NO_x flow, a potentially better estimation of the NO_x flow before the EATS can be made. This would result in an improved control of the urea injection and hopefully lead to an improved control of the tail pipe NO_x, which is beneficial for both the truck manufacturer and the general public.

1.1 Background

In this section the background of the thesis will be presented. This includes the definitions of engine out NO_x and system out NO_x. It will also describe the exhaust system, the existing sensor in the trucks and the existing NO_x flow models at Volvo Trucks (henceforth called Volvo).

1.1.1 Definition of Engine Out NO_x and System Out NO_x

There are two types of NO_x flows that will be discussed in this thesis, engine out (EO) NO_x, and system out (SO) NO_x. The difference between these are that the EO NO_x is the NO_x that is coming directly out from the engine and will be processed in the EATS. The SO NO_x on the other hand is the NO_x coming out from the vehicle and is released into the environment.

1.1.2 Reduction of NO_x in the Exhaust System Using SCR

Urea is stored in a tank that is located on the trucks. The urea is mixed with the exhaust gases from the engine and through hydrolysis [16] the urea is decomposed into NH_3. After this, the gases are transported through the SCR. Here NO_x and NH_3 are transformed to N_2, H_2O, NO_x, and NH_3. There is still NO_x present in the exhaust gases but the amount of it have been decreased by approximately 95-99.9%. The method of using the SCR to reduce NO_x in the exhaust system in the trucks is shown in Figure 1.2. Notice that this figure only shows the output generated from the NO_x and not other gases.
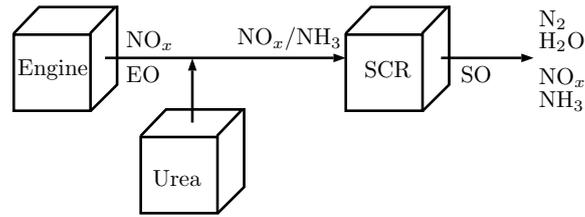
Figure 1.2: A system overview of how the amount of EO NO_x in the exhaust gases are reduced using urea in a selective catalytic reduction (SCR) process. This figure only shows what happens with the NO_x in the system and not any other gases.

1.1.3 NO_x Sensor in the Exhaust System

Today the method used for estimating the EO NO_x flow is to use readings from the NO_x sensor which measures the EO NO_x in *ppm*. It has several limitations, where sensor delay, offset in sensor data, and degeneration are some of them. One of the most severe issues is the sensor delay, where the sensor is delayed d samples. This problem is illustrated in Figure 1.3.
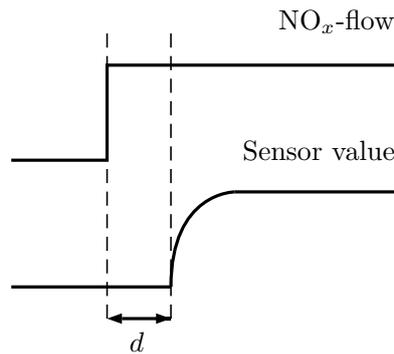
Figure 1.3: Visualization of the sensor delay problem where d is the delay.

Since the sensor reading is used for controlling the amount of urea, the system will always add the amount of urea that was needed d samples ago. This is okay if the system is operating in steady-state, i.e. the engine is running on constant speed and load but the performance will be drastically reduced if the system is changing quickly. Today the implementation of using the NO_x sensor is the standard solution in the industry and is considered to be "good-enough", but in the future, with more strict regulations of SO NO_x, this approach may be insufficient. It is therefore necessary to improve the NO_x flow estimation to get a better control of the SO NO_x.

1.1.4 Existing Models of the Engine Out NO_x

Today two models exist that are used at Volvo to estimate the EO NO_x for simulation purposes. One model is empirically constructed by radial basis functions, and the other one is a semiphysical model [17]. Semiphysical means that the model is partly built on physical equations and partly by empirical values.

Both models give the EO flow of NO_x in g/s and they can potentially give better estimation accuracy in some situations compared to the NO_x sensor. Especially in transient phases the models are outperforming the sensor due to the sensor delay. The models are shown in Figure 1.4. The inputs to the models are very similar and some examples of inputs to the models are; engine speed, torque, and air flow.
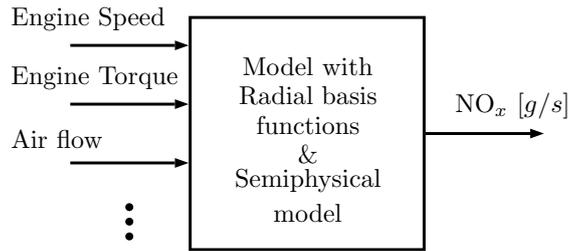
Figure 1.4: Existing models for the EO NO_x.

The difference between the two models are that the one with radial basis functions is implemented both in MatLab© and the electrical control unit (ECU) while the semiphysical model is only implemented in MatLab©. The models use similar inputs but must be treated separately regardless.

The radial basis model uses several stationary points of "normal" NO_x values for given situations to estimate the EO NO_x. It then interpolates between these values to make an estimation of the EO NO_x. The semiphysical model contains partly mathematical expressions and partly fix values. These fix values correspond to "normal values" and should be good in most of the driving scenarios.

1.1.5 Diesel Engine System Description

In Figure 1.5 an overview of a six cylinder diesel engine with exhaust gas recirculation (EGR) is shown. EGR is a process which recirculate some of the EO gas to the intake in order to reduce the SO NO_x.
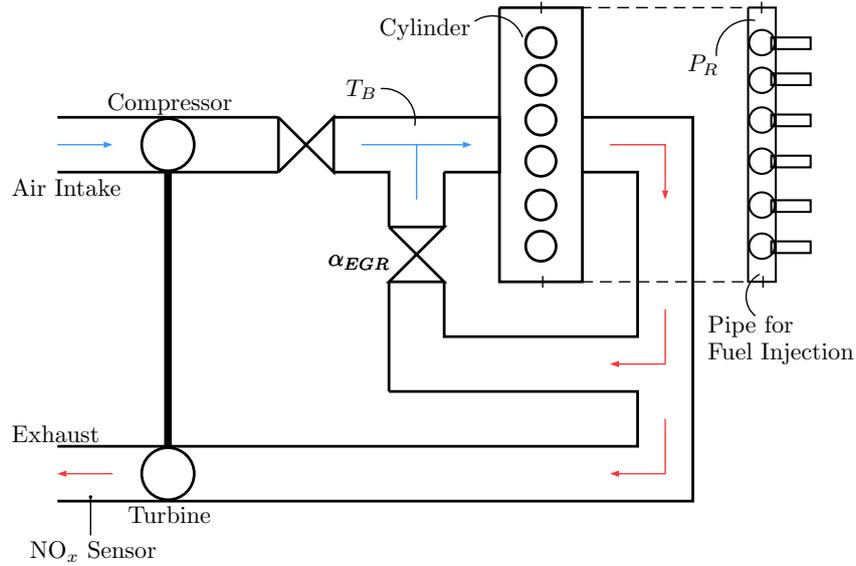
Figure 1.5: Overview of a diesel engine with six cylinders and EGR, the cut shows the injection pipe. The blue arrows mean that the gases are cold and the red ones means that the gases are warm.

There are two major types of engines that will be used in this thesis, one is an engine with extended SCR (eSCR) and the other one is a turbo compound (TC) engine. The difference between these are that the TC engine have an additional turbo after the turbine that have a mechanically connected crank shaft to recover exhaust waste energy. In addition to the extra turbo the TC engines also cool the EGR gases.

The engine is equipped with different sensors apart from the NO_x sensor. Some of these sensors are measuring:

- Rail pressure, P_R, which is the pressure in the pipe containing fuel for the cylinders.

- Boost temperature, T_B, which is the temperature of the compressed gases in the cylinders.

- Engine speed, ω_{Eng}, which is the speed of the engine in revolutions per minute (rpm).

- Engine torque, τ_{Eng}, which is the estimated torque from the engine.

- Air fuel ratio, AFR, which is the ratio of air and fuel in the cylinder charge.

- EGR-valve position, α_{EGR}, is the angle measuring how big part of the valve that is open to let gases recirculate.

There are also values calculated internally in the NO_x flow models in order to estimate the EO NO_x, here called "correction factors". Some of these are:

- Cylinder temperature corrections C_{corr} and C_{cyl} where C_{corr} is from the semi-physical model and C_{cyl} is from the radial basis model.

- Engine speed and torque correction, $C_{\omega\tau}$, from the radial basis model. It changes the EO NO$_x$ based on current value of ω_{Eng} and τ_{Eng}.

1.2 Purpose

The main purpose of the thesis is to design an EO NO$_x$ estimation observer that does not have as large delay in EO NO$_x$ estimation as today. The observer should primarily have a better dynamic behaviour and needs to remove and compensate for the sensor delay to achieve a quicker system. After that the static behaviour is going to be investigated which is the secondary purpose of the thesis and is also of high importance for the overall performance. An observer will be implemented in TargetLink in order to simulate a virtual control unit and a motor model.

If the estimation of EO NO$_x$ is done with reduced delay and more accurate than today, the dosing of urea can be done more precisely. This can lead to a lower amount of SO NO$_x$ and less consumption of urea in the trucks.

1.3 Objective

The objective of this thesis is to combine the NO$_x$ flow models and the existing sensor and investigate if this will result in less SO NO$_x$ in an engine test cell at Volvo. The research questions in this thesis are:

- Which models are preferable in which situations?

- How should the quality of the estimation of the NO$_x$ flow be analyzed?

- How should the models be weighted to utilize their respective benefits?

- Under what circumstances are the sensor better than the models?

- How should a sensor with delay be implemented in combination with the real time models to give as good estimation as possible?

- Which technique is most suitable for this task?

1.4 Scope

The scope of this thesis is to design an observer for the EO NO$_x$ that improve the accuracy compared to the presently used system. The observer should then be implemented in TargetLink.

During the project it is assumed that the EO NO$_x$ measurements from the Horiba system are the true values. The Horiba system is a measurement system that are used in the engine test cells to get accurate measurements of NO$_x$ as the conditions are well known. There are always errors in measurement systems but the Horiba system give the most accurate measurements that are at hand. So in the lack of ground truth data the Horiba measurements are assumed to be the true data.

Three different types of engines are considered when designing the observers. These engines are 480 hp eSCR, 540 hp eSCR and a 500 hp TC engine. These will henceforth be called low hp engine, high hp engine, and TC engine.

2
Theory

In this chapter, theory regarding the different areas of this thesis will be presented. These areas are stepwise linear regression, Cooks distance, linear Kalman filter, linear Kalman filtering with delayed measurements and an unknown input observer.

2.1 Stepwise Linear Regression

Stepwise linear regression is a method for estimating models from data. A linear regression model can be described as

$$m = c_0 + c_1 p_1 + ... + c_N p_N \tag{2.1}$$

where m is the model, c are coefficients, p are predictor variables and N is the number of predictor variables. A predictor variable can be either a single input to the model or combinations of several inputs. Inputs to the model are the different types of data available in the dataset used for training the linear regression model.

The stepwise linear regression algorithm recursively adds and removes predictor variables from an initial linear or generalized linear model and determines if the addition or removal improves the model. The linear regression algorithm often makes this evaluation by a series of F-tests or T-tests. An F-test is a statistical test with an F-distribution under the null hypothesis and a T-test is a statistical test that is used to make inferences about the coefficients. When the model, m, is no longer improved by adding or removing of predictor variables, the method will terminate and a final model is achieved.

2.1.1 Forward Regression and Backward Regression

There are two types of stepwise regressions, forward and backward. Forward regression adds predictor variables to an initial model, and backward regression removes predictor variables from an initial model.

In a forward regression step a predictor variable, p_n, is not included in the regression model, m, but is going to be evaluated if it should be included in m. p_n will be evaluated using the null hypothesis

$$H_0 : c_n = 0 \tag{2.2}$$

where H_0 is the null hypothesis and c_n is the coefficient multiplied with p_n if p_n is added to m. H_0 in (2.2) implies that p_n should not be included in m, because if p_n

is multiplied with zero, the information in p_n will not be used. If there are enough statistical evidence to reject H_0, p_n is added to m, otherwise H_0 is correct and p_n is discarded.

In a backward regression step the predictor variable, p_n, is on the other hand already included in m and reevaluated. In this case the null hypothesis is the same as in the forward step and shown in (2.2). If there are enough statistical evidence to reject H_0, p_n stays in m, otherwise H_0 is correct and it is removed from m.

2.1.2 Overfitting in System Identification

Stepwise linear regression is a method that can be used for system identification, and a general problem in linear regression is overfitting. This is a problem that is due to that the trained model is too specific for the training data and trained "too much" on the training data. It might get very good performance on the training data but when exposed to new data the performance is bad. This problem is illustrated in Figure 2.1.

Overfit   Appropriate Fit
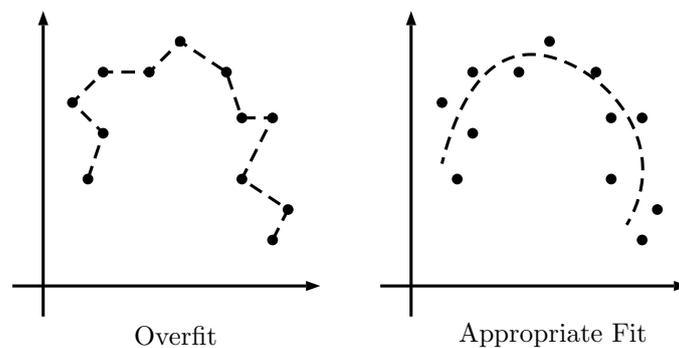
Figure 2.1: Example of when a trained model have achieved an appropriate fit to data and when it have problem with overfitting.

In Figure 2.1 the model is trying to follow the dots with the dashed line. First the model fits the data very well and can reach each dot exactly. But if this model would be exposed to new data, it would probably perform poorly. In the second case the model is not fitted exactly to the data but is more generalized and would probably perform better on new data.

2.1.3 Performance Measures for Linear Regression Models

A measure of how well a regression model performs is the mean squared error (MSE). This is calculated by

$$\text{MSE} = \frac{1}{n} \sum_{i=1}^{n} \left(x(i) - \hat{x}(i) \right)^2 \tag{2.3}$$

where x is the true value and $\hat{x}$ is the estimated value. This measure of error will punish samples which are far away from the real value due to the square of the error.

Another way to measure the quality of a regression model is the normalized root mean square error (NRMSE). This is a measure of how good the regression model is

in comparison to the trivial solution. The trivial solution in this case is a constant value equal to the mean of true values. The NRMSE is calculated with

$$\text{NRMSE} = 1 - \frac{\sum_{i=1}^{n} \left(x(i) - \hat{x}(i) \right)^2}{\sum_{i=1}^{n} \left(x(i) - \bar{x} \right)^2} \tag{2.4}$$

where x are the true value, $\hat{x}$ the estimated value and $\bar{x}$ the mean of the true values. The value of NRMSE is

$$-\infty \leq \text{NRMSE} \leq 1 \tag{2.5}$$

where $-\infty$ indicates that the fit of the model is bad, 0 if the model is equally good as the trivial solution, and 1 indicates that the model is a perfect fit.

2.1.4 Cooks Distance for Analyzing Regression Model Data

Cooks distance is a measure in regression analysis to decide the influence of data and is a combination of leverage and the residual for each training sample. Leverage is a measure of how much a training sample is affecting the model and residual is the error between the model value and the training data. This means that if the Cooks distance is high, the training sample is both far from the fitted model and has a large impact on the model. This is a bad training sample which should be removed from the dataset used for training the model. It is mathematically described as

$$D_i = \frac{\sum_{j=1}^{n} \left(\hat{y}_j - \hat{y}_{j(i)} \right)^2}{p \cdot \text{MSE}}, \tag{2.6}$$

where y_j is the fitted value for the model, $y_{j(i)}$ is the fitted value for the model without the measurement observed in i, p is the number of coefficients and the MSE is the mean squared error of the regression model. An example of Cooks distance is shown in Figure 2.2.
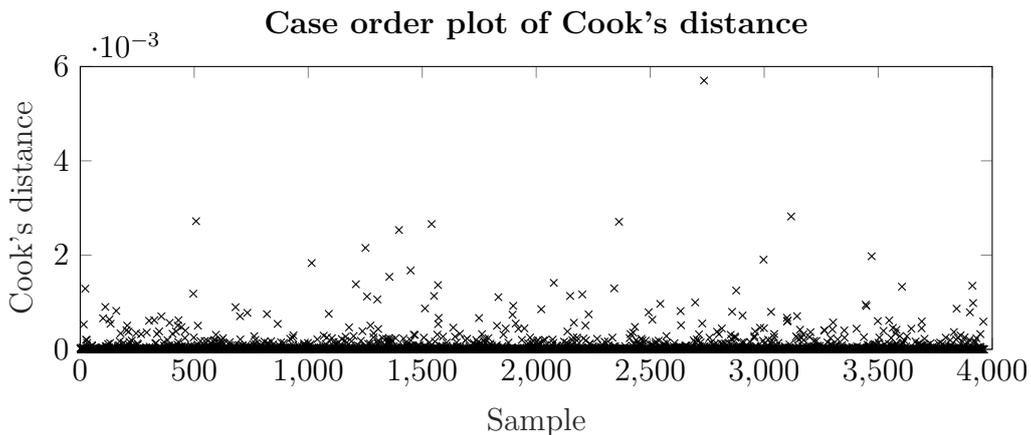
Figure 2.2: Example of Cooks Distance, one point is way higher than the others (between sample 2500 and 3000) and should be removed when training a model.

The procedure of using Cooks Distance for regression models can be formulated as:

1. Train a model on all training samples

2. Analyze the developed model by finding training samples with large Cooks distance

3. Remove the found training samples from the dataset

4. Train a new model based on the dataset with removed training samples

2.2   State Estimation with Kalman Filtering

The Kalman filter is an observer that estimates states in a dynamical system based on noisy measurements. This can be useful when, for example, estimating position and velocity of an object. The linear Kalman filter is applicable on linear systems on the state-space form

$$
\begin{aligned}
x_{k+1} &= A_{k+1}x_k + B_{k+1}u_{k+1} + q_{k+1} \\
y_k &= C_k x_k + D_k u_k + r_k
\end{aligned}
\tag{2.7}
$$

at the discrete time $k+1$ and where q and r are Gaussian distributed independent noises,

$$
q \sim \mathcal{N}(\mu_q, Q) \qquad r \sim \mathcal{N}(\mu_r, R).
\tag{2.8}
$$

The equation describing x_{k+1} in (2.7) is called a state-transition equation and the equation describing y_k is called the measurement equation. In the state-transition equation, A is the state transition matrix that describes how the states changes for each time step and B is the control input model. C in the measurement equation is the measurement transition equation which describes how the measurements from each source relates to each state. D specify how the control input affects the measurement signal.

   The estimated state of a Kalman filter is described as a Gaussian distribution with mean $\hat{x}_{k+1|k+1}$ and covariance $P_{k+1|k+1}$. This normal distribution is recursively calculated by the linear Kalman filter algorithm which is done in two distinct steps; the prediction step and the update step. In the prediction step the calculations

$$
\begin{aligned}
\hat{x}_{k+1|k} &= A_{k+1}\hat{x}_{k|k} + B_{k+1}u_{k+1} \\
P_{k+1|k} &= A_{k+1}P_{k|k}A_{k+1}^T + Q_{k+1}
\end{aligned}
\tag{2.9}
$$

are executed. Here is a prediction of the state mean, $\hat{x}_{k+1}$, and covariance, P_{k+1}, done based on the information at time k. The second step of the linear Kalman filter is to make an update to this prediction based on information received from the measurement at time $k+1$. The update step is done by first calculating

$$
\begin{aligned}
K_{k+1} &= P_{k+1|k}C_{k+1}^T S_{k+1}^{-1} \\
V_{k+1} &= y_{k+1} - C_{k+1}\hat{x}_{k+1|k} \\
S_{k+1} &= C_{k+1}P_{k+1|k}C_{k+1}^T + R_{k+1}
\end{aligned}
\tag{2.10}
$$

and then doing the actual update according to

$$\hat{x}_{k+1|k+1} = \hat{x}_{k+1|k} + K_{k+1}V_{k+1}$$
$$P_{k+1|k+1} = P_{k+1|k} - K_{k+1}S_{k+1}K_{k+1}^T$$

(2.11)

where K_{k+1} is the Kalman gain, V_{k+1} the innovation and S_{k+1} the innovation covariance. The Kalman gain describes how much of the innovation that will be used when estimating the state based on the new measurements at time $k+1$ and the innovation captures how much the prediction differs from the measurement.

that is used to determine how much of the innovation that will be used when estimating the state.

2.3   Kalman Filtering With Delayed Measurements

If a measurement delay is present in the system it is not optimal to fuse the measurement value received at time k with non delayed values at time k. An extended approach of the method described in section 2.2 is instead preferable where the delay of the delayed measurements are taken into consideration.

Let s denote the time where the delayed measurement was taken, and k denote the time when the measurement was received. Let this delayed measurement be denoted as y_k^*. The measurement have then been delayed d samples. This can be described as

$$k = s + d$$

(2.12)

and this delay is shown in Figure 2.3.
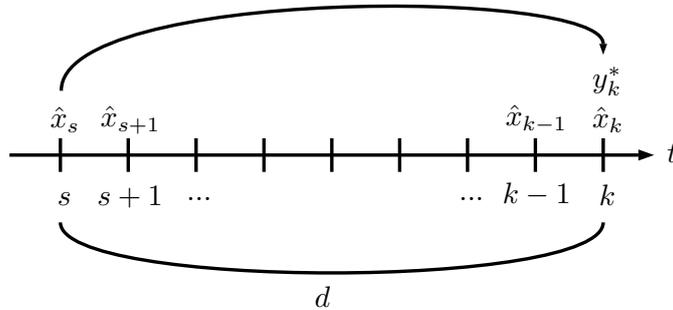
Figure 2.3: Timeline over when the delayed measurement y_k^* was taken at time s and when it was received at time k.

One approach for handling this delay is to continuously run a Kalman filter as described in Section 2.2 without the delayed measurement until time k when a delayed measurement from time s is received. If the system contains a total of N sources of information, including the delayed sensor, only $N-1$ of them are used until time k. These measurements are denoted as y_k. Here the measurement equation

$$y_k = \begin{bmatrix} C_k \end{bmatrix} x_k + r_k$$

(2.13)

will be used, where C_k are measurement equations for y_k.

When the delayed measurement y_k^* are received at time k the state $\hat{x}_s$ is re-estimated using both y_k and y_k^*. This will lead to the measurement equation

$$y_s = \begin{bmatrix} C_s \\ C_s^* \end{bmatrix} x_s + r_s^* \tag{2.14}$$

where C_s^* describes the time delayed measurement and r_s^* describes the measurement noise for all sources. Similarly to the linear Kalman filter without measurement delay, the measurement are Gaussian noises and distributed as

$$r \sim \mathcal{N}(\mu_r, R) \qquad r^* \sim \mathcal{N}(\mu_{r^*}, R^*). \tag{2.15}$$

After estimating $\hat{x}_s$, $\hat{x}_{s+1:k}$ are reestimated using $y_{s+1:k}$ with the measurement model (2.13), updating the values based on the new information from the delayed sensor.

The steps in the delayed Kalman filter (DKF) can be summarized as:

1. Run Kalman filter without the delayed sensor until delay sensor values, y_k^*, are received at time k

2. Recalculate $\hat{x}_s$ with both y_s and y_k^*

3. Recalculate $\hat{x}_{s+1:k}$ with $y_{s+1:k}$ as before

4. Run Kalman using y_k until next delayed sensor value arrives

For this method to work, the delay must be known for each delayed measurement. Otherwise there is no information where s is located in time and no reestimation can be done.

2.4 Unknown Input Observer

The unknown input observer (UIO) is a method where a linear Kalman filter can be used in the presence of unknown inputs to the system. The idea is based on adding an extra state variable per unknown input. The extra state will be used to estimate the unknown input. Instead of having the unknown state as an input to the system, integrated white noise is added. The integrator transforms the white noise to a random walk which will be used to model the unknown input. Figure 2.4 shows the system with an unknown input and Figure 2.5 shows the system with the noise added instead of the unknown input.
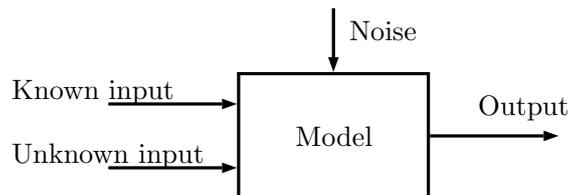
Figure 2.4: A system with an unknown input.
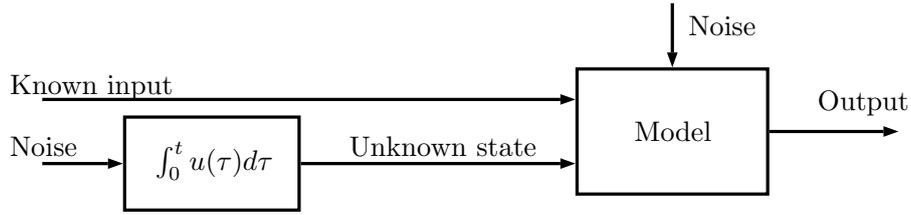
Figure 2.5: A system where an unknown input is modeled as integrated white noise.

The addition of an extra state will effect the state space model of the system. Let a state space model be described as

$$
\begin{aligned}
x_{k+1} &= A_k x_k + B_k u_k \\
y_k &= C_k x_k + D_k u_k.
\end{aligned}
\tag{2.16}
$$

Now let the input vector, u, be split in to a known and an unknown input vector, v and d. This will expand (2.16) to

$$
\begin{aligned}
x_{k+1} &= A_k x_k + B_k^v v_k + B_k^d d_k \\
y_k &= C_k x_k + D_k^v v_k + D_k^d d_k
\end{aligned}
\tag{2.17}
$$

where now B^v and D^v only contains the known input equations and B^d and D^d the unknown inputs. Now let the state x^d that correspond to the unknown inputs be added. The expressions for x^d are

$$
\begin{aligned}
x_{k+1}^d &= A^d x_k^d \\
d_k &= C_k^d x_k^d.
\end{aligned}
\tag{2.18}
$$

This will make (2.17)

$$
\begin{aligned}
x_{k+1} &= A_k x_k + B_k^v v_k + B_k^d C_k^d x_k^d \\
y_k &= C_k x_k + D_k^v v_k + D_k^d C_k^d x_k^d.
\end{aligned}
\tag{2.19}
$$

This can now in combination with (2.18) be expressed on matrix form as

$$
\begin{aligned}
\begin{bmatrix} x_{k+1} \\ x_{k+1}^d \end{bmatrix} &= \begin{bmatrix} A_k & B_k^d C_k^d \\ 0 & A_k^d \end{bmatrix} \cdot \begin{bmatrix} x_k \\ x_k^d \end{bmatrix} + \begin{bmatrix} B_k^v \\ 0 \end{bmatrix} v_k \\
y_k &= \begin{bmatrix} C_k & D_k^d C_k^d \end{bmatrix} \cdot \begin{bmatrix} x_k \\ x_k^d \end{bmatrix} + \begin{bmatrix} D_k^v \\ 0 \end{bmatrix} v_k
\end{aligned}
\tag{2.20}
$$

which is the augmented state space model for unknown inputs. By using the augmented state space model an UIO can now be created by designing a linear Kalman filter from that model.

2.5   Steady State Kalman Filters

The Kalman filter recursively finds a Kalman filter gain, K, that is used to determine how much of the innovation that will be used when estimating the state. K converge if the current system is time-invariant and R is a constant matrix, or converges to a constant matrix. The converged K, further on called $\bar{K}$, can be used instead of recursively determine K for each time step when estimating the states. This is cheaper in a computational aspect and might give as good results as the recursive solution to the Kalman filter. Note that the first estimations wont be optimal with this approach. Due to that $\bar{K}$ is known on beforehand, the Kalman filter equations will be simplified to

$$\hat{x}_{k+1|k} = A\hat{x}_{k|k}$$
$$V_{k+1} = y_{k+1} - C_k\hat{x}_{k+1|k} \tag{2.21}$$

and

$$\hat{x}_{k+1|k+1} = \hat{x}_{k+1|k} + \bar{K}V_{k+1}. \tag{2.22}$$

3
Methods

In this chapter the different techniques used in the thesis will be described. First are models describing when the two NO_x flow models are accurate or not going to be concluded. To achieve this, variables that affect this accuracy must be found and then these variables are used in a stepwise linear regression problem. When models describing the accuracy of the two NO_x flow models are concluded this information can be used in different observer strategies to weigh the information from the two NO_x flow models.

First in this chapter important terms will be defined, then the preprocessing of data used for linear regression will be presented, thereafter the identification of important variables for both models will be presented, and then the sensor fusion and observer strategies will be shown.

3.1 Identification of Significant Variables

One crucial part of the thesis is to conclude when to trust either one of the NO_x flow models, the sensor, or a combination of these. To do this, information about when the respective NO_x model are accurate or not is important. The information about model accuracy are here described by two functions, E_{SP} for the semiphysical model and E_{RB} for the empirical radial basis model. Each model also have unique calibrations based on different engine hardware, low hp eSCR, high hp eSCR, and 500 hp TC. In total there are six different accuracy functions named

$$E_{SP}^{low} \quad E_{SP}^{high} \quad E_{SP}^{TC} \quad E_{RB}^{low} \quad E_{RB}^{high} \quad E_{RB}^{TC}. \tag{3.1}$$

The low and the high hp engines have the same system description but there are differences in for example the turbine specifications.

The accuracy function, E, is based on how far the NO_x flow model value is from the Horiba measurement. Horiba measurements are measured in the engine test cells and are considered to be the true NO_x values. If the difference between model value and Horiba is high the accuracy E will be low, and when the difference between Horiba and model is low, the accuracy E will be high. The magnitude of the difference between any model and Horiba are denoted with D and is calculated as

$$D = |M - H_N| \tag{3.2}$$

where H_N is the Horiba and M is any model. From this, the accuracy function is calculated as

$$E = \frac{1}{D}. \tag{3.3}$$

The difference, D, is shown in Figure 3.1
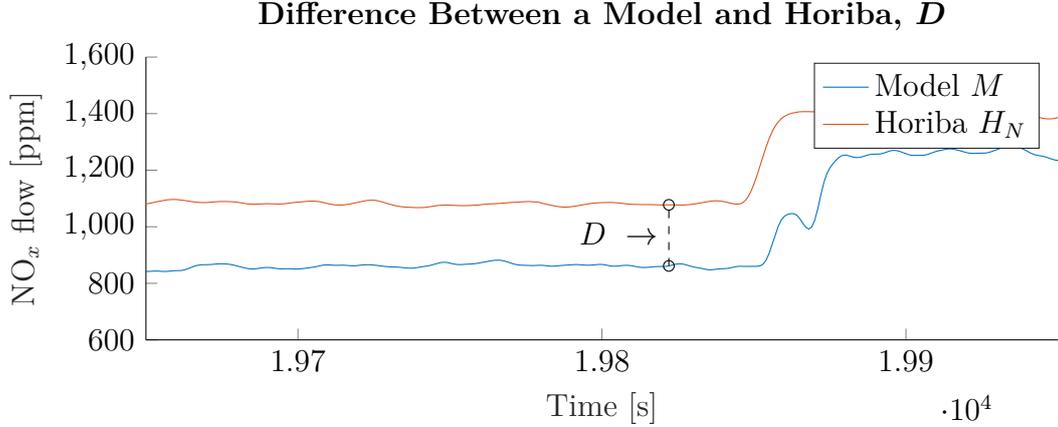
Figure 3.1: Difference in EO NO_x between a model, M, and Horiba, H_N, denoted as D.

The problem that will be solved in this section is that D is not available in the truck for any of the engine types. Horiba measurements are only available in the test cells and therefore D cannot be used directly. The solution to this is to build functions that are as similar to D as possible, named $\widehat{D}$. $\widehat{D}$ will be calculated by setting up a stepwise linear regression problem that will be solved in MATLAB©. $\widehat{D}$ will in this case correspond to the model m described in Section 2.1. With the two different NO_x flow models and the three different engine types there are six different functions

$$\widehat{D}_{SP}^{low} \quad \widehat{D}_{SP}^{high} \quad \widehat{D}_{SP}^{TC} \quad \widehat{D}_{RB}^{low} \quad \widehat{D}_{RB}^{high} \quad \widehat{D}_{RB}^{TC} \tag{3.4}$$

that will be estimated. E will then be based on $\widehat{D}$ instead of D and be approximated with

$$E \approx \frac{1}{\widehat{D}}. \tag{3.5}$$

To do this, the variables that influence D are going to be found. These variables are then going to be used in a linear regression problem to construct $\widehat{D}$.

When comparing separate variables with model accuracy it is important to know that these separate variables alone never will represent the whole error. The system is very complex and there exists no variable that on its own can explain it as a whole.

The following sections will first describe the preprocessing of the data and then the process of designing the functions in (3.4).

3.1.1 Data Preprocessing

When identifying $\widehat{D}$, data representing normal truck driving is important. The used datasets contain values from cycles representing driving on highway and inner city driving. This data is acquired in the engine test cells where these cycles were simulated. The reason for the different types of driving is that the observer should work everywhere and not only for a certain type of drive cycle. The datasets contains

for example values from one of the NO$_x$ flow models, engine speed, torque, readings from the NO$_x$ sensor and much more. Such datasets are used for all three engines.

There are several preprocessing steps that are necessary to apply on the used datasets to use them for identifying $\widehat{D}$. These steps are:

1. Low pass filter the model values

2. Shift the Horiba measurements

3. Crop the datasets to remove idle parts

4. Split the data into training datasets and validation datasets

The first part of the preprocessing is to make the model values as similar to the Horiba measurements as possible. The model values have an issue with high frequency ripple and therefore a low pass filter is applied to the model output. The low pass filter applied have the order of 6 and the time constant 10. The result from this filtering is shown in Figure 3.2. After filtering the ripple is removed and the model values have similar smoothness as the Horiba measurements.

Original and Filtered Signal From One Model
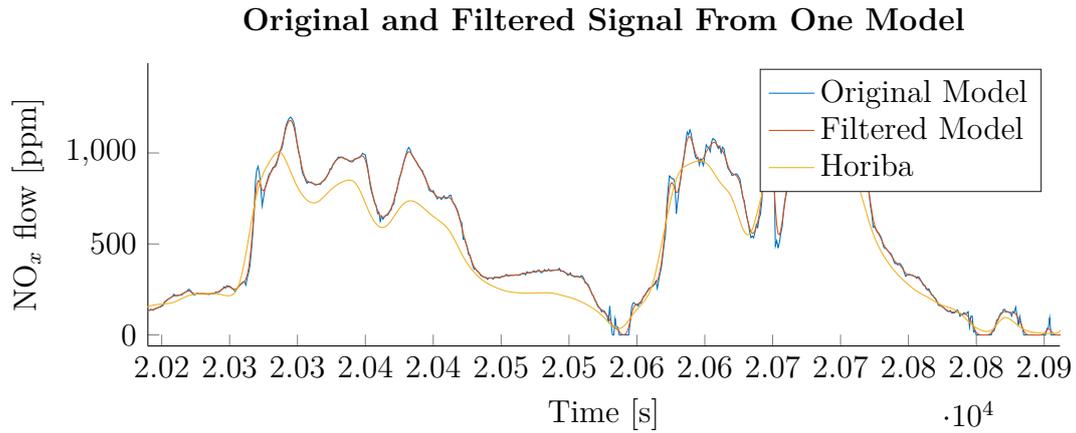
Figure 3.2: Filtered and unfiltered values from one of the models in comparison to the Horiba measurements.

The second part of the preprocessing is to time shift the Horiba measurements. Compared to the data from the models the Horiba measurements are shifted. The Horiba measurements are calculated through a system to get an exact result and the calculation takes time, therefore it needs to be shifted to line up with the models to be able to correctly evaluate the models. This is done by hand for each dataset to get it as accurate as possible. However it is still not exactly precise because the delay is varying through the cycle. The delay are varying between one to five samples trough out the dataset. This is a small variation but will still effect the performance.

In Figure 3.3 and 3.4 a comparison between a filtered and shifted dataset and a non filtered and non shifted dataset is shown.
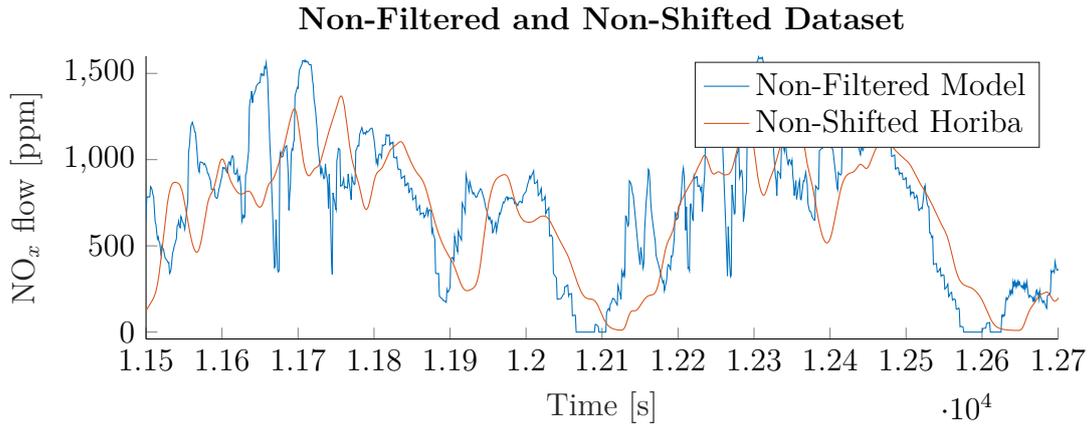
Figure 3.3: Dataset before filtering and shifting. The blue line is the Horiba measurement and the orange is the NO_x flow estimated by one of the NO_x flow models.
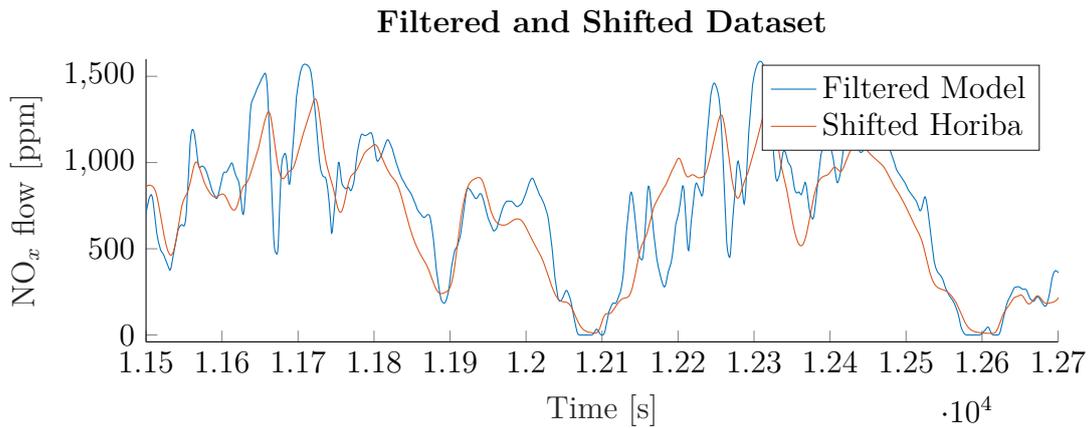
Figure 3.4: Dataset after filtering and shifting. The goal here is that the two signals should have as similar characteristics as possible. The blue line is the Horiba measurement and the orange is the NO_x flow estimated by one of the NO_x flow models.

As shown in Figure 3.3 and 3.4 the filtering and shifting need to be done in order to get the two signals to align in time and have similar smoothness. This is very important later when an observer will be designed.

The main goal of this thesis is to create an observer that have better transient behavior than the NO_x sensor. The behavior of $\widehat{D}$ will influence the behaviour of the final observer. The datasets used in the thesis have parts where the engine is idle driving. This can be seen in Figure 3.5 where the idle driving is occurring when the engine speed is around 500 rpm for long periods of time.
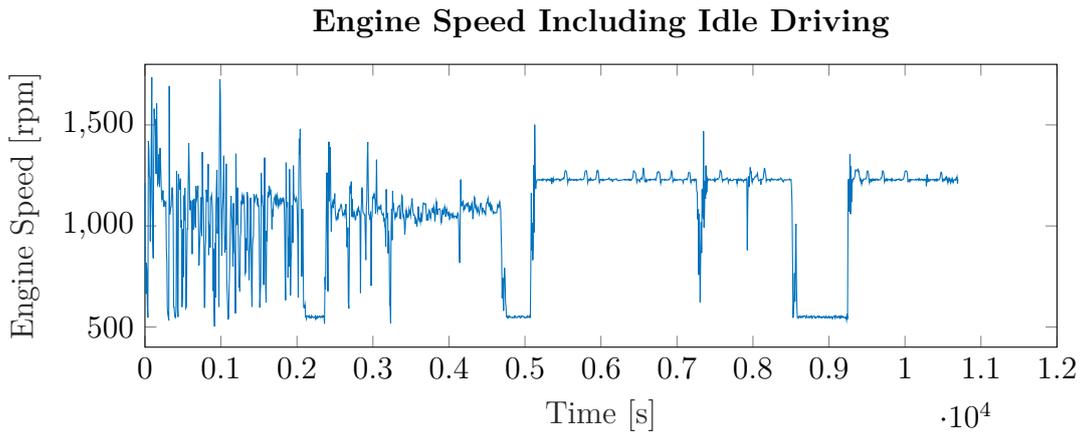
Figure 3.5: The engine speed over a whole dataset. Note here that the idle parts of the dataset, where the engine speed is around 500 rpm, will get a disproportional impact while calculating $\widehat{D}$ as the transient behaviour is prioritized.

The idle parts should in some extent be reduced to a much smaller length than what is represented in the datasets. The reason for this is that the idle parts will influence $\widehat{D}$ in a disproportional way. They will force $\widehat{D}$ to adjust to the idle parts more and potentially lose accuracy in the transient parts. Since the main goal of the thesis is to make the system perform well in transients these parts should be reduced or even removed. This is shown in Figure 3.6 where the idle parts from Figure 3.5 are removed.
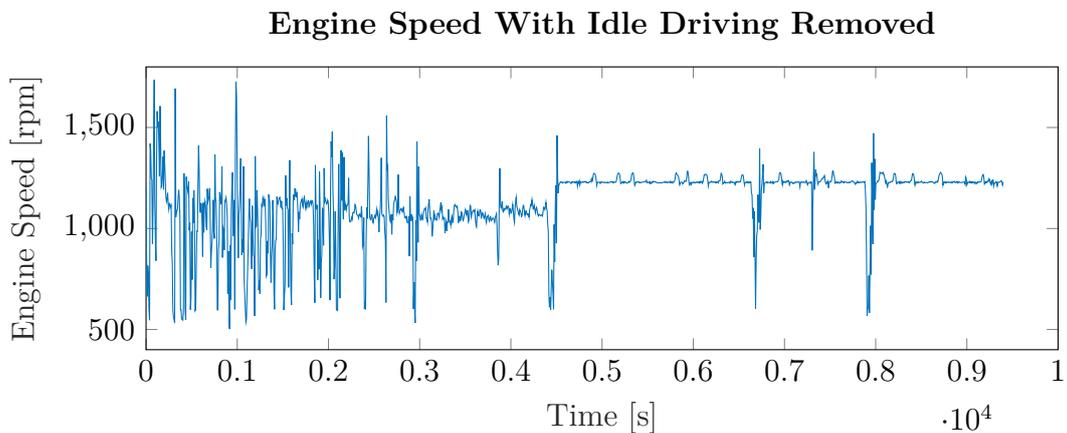
Figure 3.6: The engine speed over a whole dataset where idle parts have been removed.

The resulting error between the two NO_x flow models and the Horiba measurements after shifting, filtering, and cropping the data are shown for low hp eSCR in Figure 3.7. However, it is not the absolute value of the errors that are shown here as the difference between the NO_x flow models and the Horiba measurements can be both positive and negative. The non-absolute valued errors are hence called $\bar{D}$ instead.
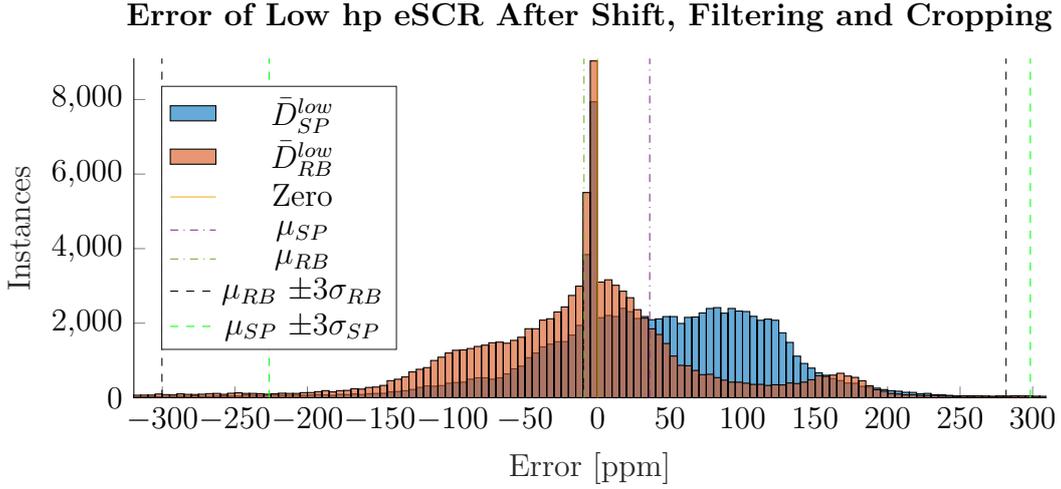
Figure 3.7: The error for the low hp eSCR engine for both the semiphysical and radial basis models where σ is the standard deviation and μ is the mean.

The reason for the spike around zero in Figure 3.7 is due to that the model values correctly shows zero NO_x flow when the Horiba measurements does the same. This means that the model is predicting the correct value which leads to no error. The small hump at 175 is due to that when the NO_x flow is stationary, $\bar{D}_{RB}^{low}$ is often overestimating the NO_x flow with approximately 175 ppm for this engine model.

The resulting error between the two NO_x flow models and the Horiba after shifting, filtering and cropping the data are shown for high hp eSCR in Figure 3.8.
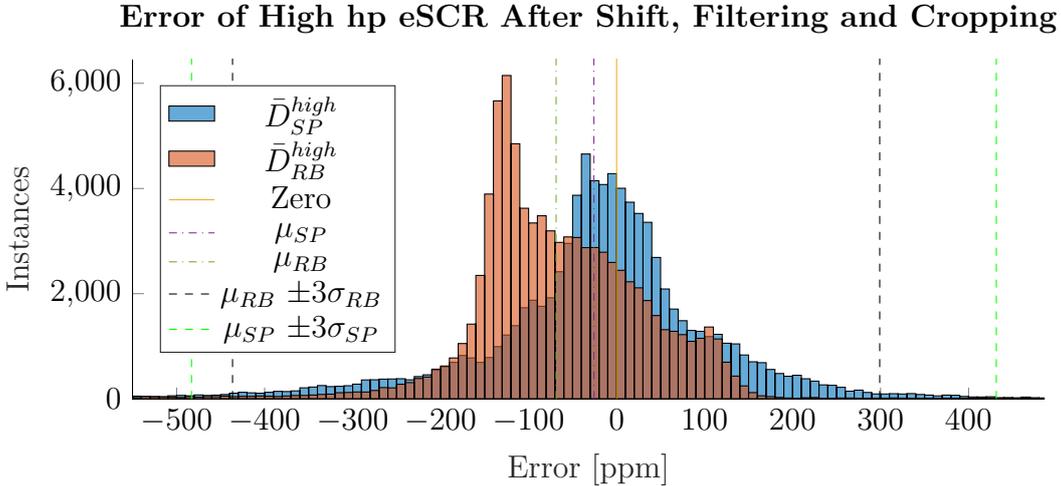
Figure 3.8: The error for the high hp eSCR engine for both the semiphysical and radial basis models where σ is the standard deviation and μ is the mean.

In Figure 3.8 $\bar{D}_{SP}^{high}$ has a good distribution. It could preferably have a smaller standard deviation but that behaviour is a side effect of the model bias and variance. $\bar{D}_{RB}^{high}$ on the other hand have a large peak at approximately -130. The reason for this is that $\bar{D}_{RB}^{high}$ often underestimates the NO_x flow but not in all points.

The resulting error between the two NO_x flow models and the Horiba after shifting, filtering and cropping the data are shown for TC in Figure 3.9.

Error of TC After Shift, Filtering and Cropping
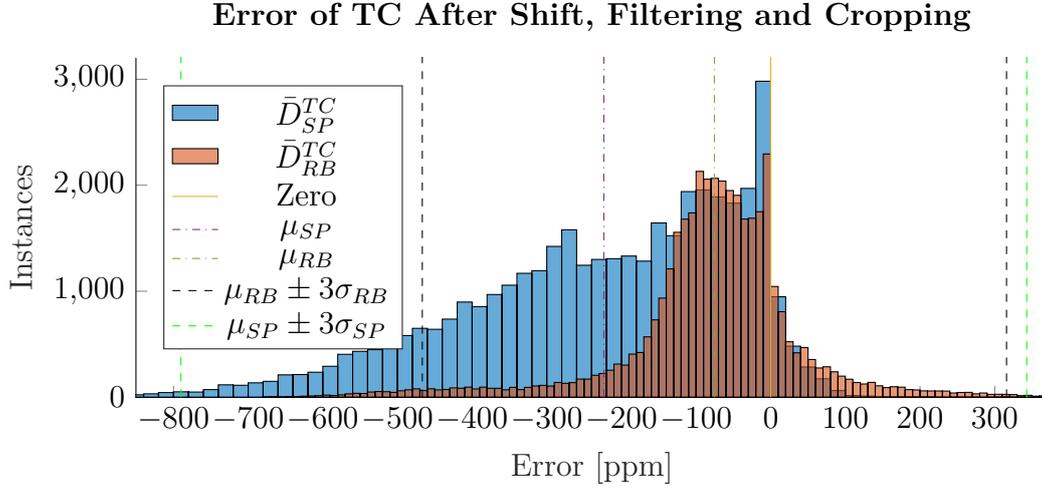
Figure 3.9: The error for the TC engine for both the semiphysical and radial basis models where σ is the standard deviation and μ is the mean.

In Figure 3.9 $\bar{D}_{RB}^{TC}$ somewhat follow a left-shifted normal distribution. This indicates that the model continuously underestimates the NO_x flow values. This behaviour is also seen for $\bar{D}_{SP}^{TC}$ but here the standard deviation is much wider. It is known that M_{SP} was not developed for handling TC engines. Due to this, the performance of $\bar{D}_{SP}^{TC}$ on this engine is therefore not as good as for the low hp or high hp eSCR engines. M_{SP} is often overestimating the NO_x flow and this creates the right skewed form of $\bar{D}_{SP}^{TC}$ in Figure 3.9.

The results from Figure 3.7-3.9 are shown in Table 3.1.

Table 3.1: Standard deviations and mean values for the two NO_x flow models in comparison to the corresponding Horiba measurement for all engines.

		Engine Models		
		Low	High	TC
$\bar{D}_{SP}$	μ_{SP}	36.00	-25.79	-223.55
	$3\sigma_{SP}$	262.43	457.31	566.46
$\bar{D}_{RB}$	μ_{RB}	-9.28	-68.72	-75.43
	$3\sigma_{RB}$	291.00	367.75	391.30

From Table 3.1 it can be seen that values for three standard deviations are high based on that they are measured as NO_x flow in ppm. For a normal distribution $\sim 99.7\%$ of all points are included inside the span defined by

$$\mu - 3\sigma \leq X \leq \mu + 3\sigma \tag{3.6}$$

where μ is the mean of the distribution and σ is the standard deviation. Table 3.1

can be further expressed as the expressions

$$-226.43 \leq \bar{D}_{SP}^{low} \leq 298.43 \qquad -300.26 \leq \bar{D}_{RB}^{low} \leq 281.70$$
$$-483.09 \leq \bar{D}_{SP}^{high} \leq 431.52 \qquad -436,47 \leq \bar{D}_{RB}^{high} \leq 299.02 \qquad (3.7)$$
$$-790.01 \leq \bar{D}_{SP}^{TC} \leq 342.91 \qquad -466.73 \leq \bar{D}_{RB}^{TC} \leq 315.87.$$

The ideal case is that $\sigma = 0$ and $\mu = 0$. This would mean that the model value is equal to the Horiba measurement all the time but this is, as seen in (3.7) not the case.

The performance of D for the two NO_x models will influence all upcoming steps and set an upper limit of how good the future developed system can be. For example if D_{SP}^{TC} have bad performance $\widehat{D}_{SP}^{TC}$ will also have bad performance because its purpose is to be as similar to D_{SP}^{TC} as possible. This will affect the observer that will use $\widehat{D}_{SP}^{TC}$ and the performance of it will be bad. If the performance for D_{SP}^{TC} is improved the final result could be improved as well.

Later in the process a regression problem to find $\widehat{D}$ is going to be done. The regression is done on one dataset per engine-model combination and consist of approximately 100 000 samples in each. These datasets are all split in two parts where 80% were used for training and 20% were used for validation. This split is visualized in Figure 3.10. The reason for this is that the validation of the function performance should be evaluated on data not used in the optimization to avoid overfitting.
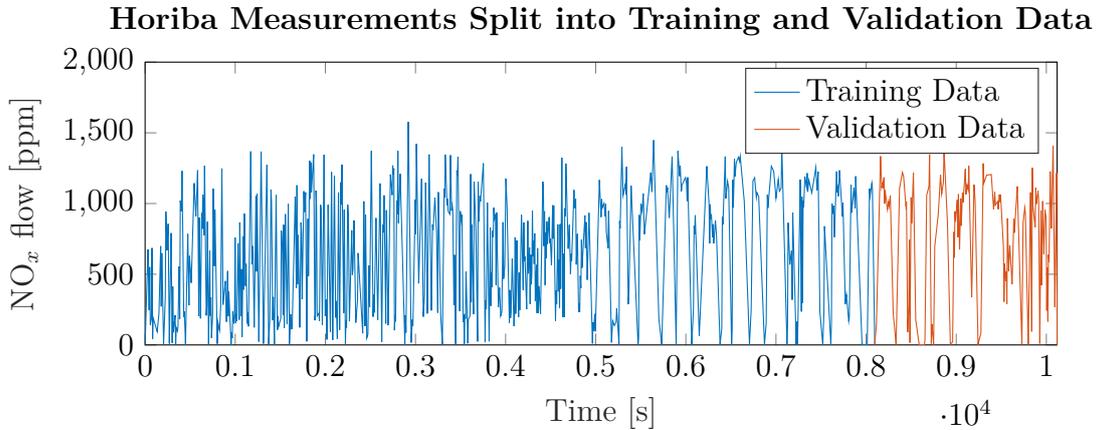
Figure 3.10: Horiba measurements that have been splitted into a training and validation dataset.

Now the preprocessing of the data is done and it is ready to be used in a future regression problem. First, variables that have any kind of relation with D must be determined and this is going to be done in the two upcoming sections.

3.1.2 Semiphysical Model

In this section the semiphysical model will be investigated. A function that describes the accuracy of the model as accurate as possible will be developed. Only the low hp eSCR will be presented in this section, i.e., $\widehat{D}_{SP}^{low}$ which from here on is refereed

to as $\widehat{D}_{SP}$. The reason for this is that the other engines have similar behaviour as the low hp eSCR and the same reasoning are applied to the other engines.

In the two upcoming sections different variables will be investigated to see if there is any relation to D_{SP}. When the variables that effect D_{SP} have been found these can be used to find $\widehat{D}_{SP}$ through stepwise linear regression.

One feature of the semiphysical model that will be exploited is that it is constructed for different reference points for some of the inputs. For example, engine speed have an "usual" speed which the model is constructed around. An offset from the reference points should be an indicator of the performance of the model where a high offset would mean a high error in D_{SP}. Some of the variables with high offset are therefore used when constructing $\widehat{D}_{SP}$ as they most likely affect the performance.

The two first variables, ω_{Eng} and P_R, are both constructed for stationary values. These two variables are here used as percentage offsets from their stationary values. These stationary values are here denoted as $\widetilde{\omega}_{Eng}$ and $\widetilde{P}_R$. In Figure 3.11 the fraction between the measured values and the reference values are shown.
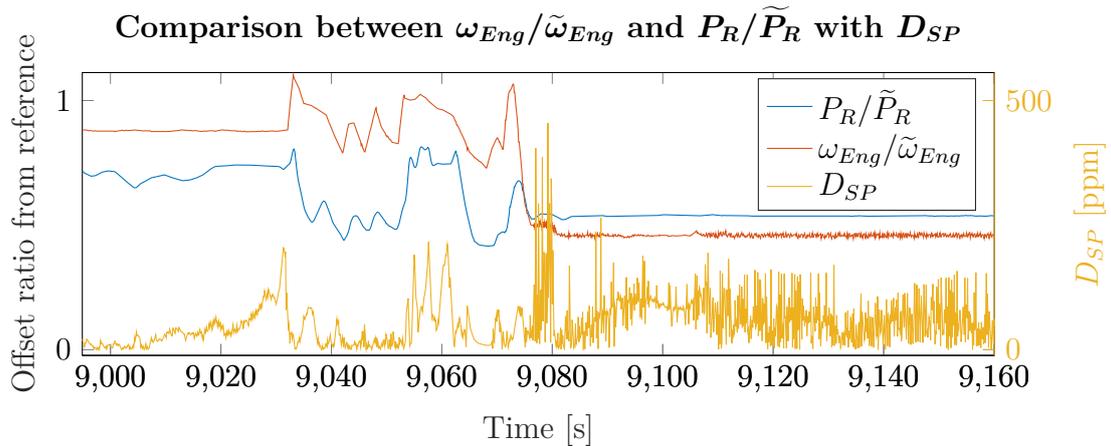
Figure 3.11: Comparison of the percentage offset for ω_{Eng} and P_R and the difference between the Horiba values and estimated values from the semiphysical model.

As seen in Figure 3.11, D_{SP} is higher and more noisy when the offset is further away from the reference point at 1 on the left y-axis for both ω_{Eng} and P_R. This is the reason for including them in $\widehat{D}_{SP}$.

Another variable that is not related to a reference point but still contain useful information about D_{SP} is the cylinder wall temperature correction factor, C_{corr}. This is a variable that is internally used in the semiphysical model and have an inverse relation to the model error D_{SP}. This relation is shown in Figure 3.12.
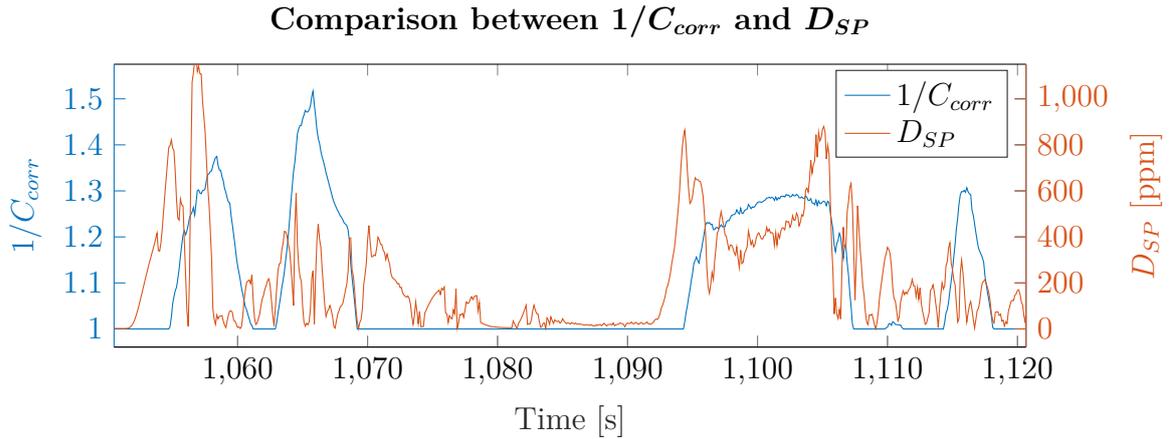
Figure 3.12: Comparison of $1/C_{corr}$ and the difference between the Horiba values and estimated values from the semiphysical model.

It can be seen in Figure 3.12 that the inverse of C_{corr} somewhat follows the same behaviour as D_{SP}. This indicates that information about D_{SP} is present in C_{corr} and should therefore be taken into consideration when calculating $\widehat{D}_{SP}$ as well.

Another variable that can be seen to have a relation with D_{SP} is the air fuel ratio, AFR. The AFR does not have a relation that is dependent on a reference point either but it has a property which is useful when creating $\widehat{D}_{SP}$. The comparison between D_{SP} and the AFR is shown in Figure 3.13.
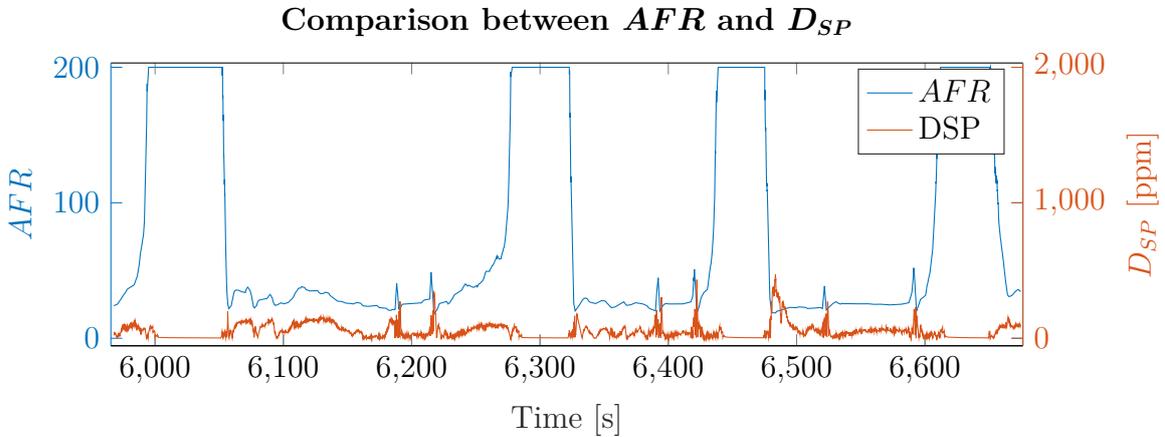
Figure 3.13: Comparison of AFR and the difference between the Horiba values and estimated values from the semiphysical model.

It can be seen in Figure 3.13 that D_{SP} goes towards zero when AFR goes to its saturation value as described in Section 1.1.5. In practice this means that when no fuel is injected there is no engine out NO_x and therefore there will be no difference in D_{SP}.

When adding AFR to the regression model it was, due to the saturation, set to be $200 - AFR$. The reason for this is that the desired dynamic from it was that D_{SP} was low when AFR equals 200. In these cases it is desirable to cancel out terms to get as low value as possible and therefore should $200 - AFR$ be added.

The torque, τ_{Eng}, was also found to affect D_{SP}. The comparison between them is shown in Figure 3.14.

Comparison between τ_{Eng} and D_{SP}
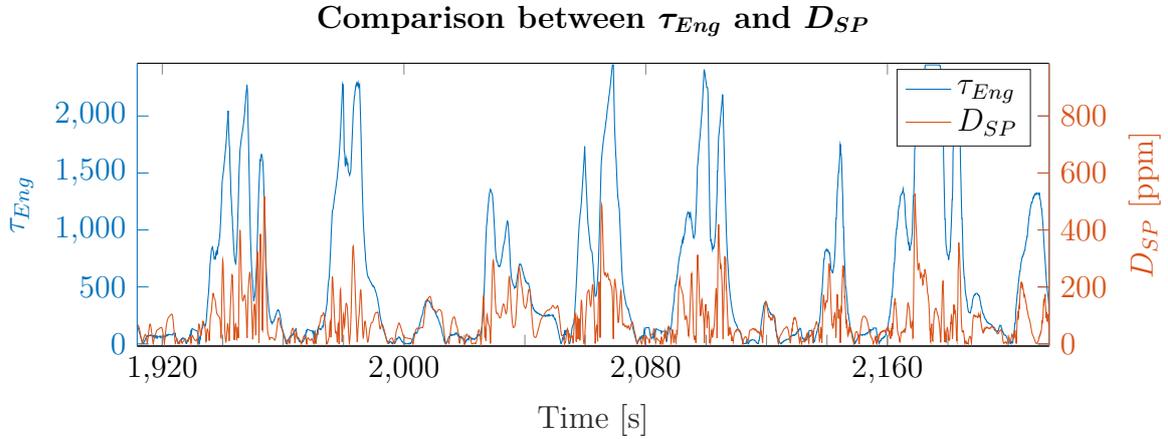
Figure 3.14: Comparison of τ_{Eng} and the difference between the Horiba values and estimated values from the semiphysical model.

It can be seen in Figure 3.14 that D_{SP} tends to go higher every time that τ_{Eng} gets a higher value. τ_{Eng} is therefore also good to use when estimating $\widehat{D}_{SP}$.

The last variable that was visually found to have an impact on D_{SP} is the EGR valve angle, α_{EGR}. This is a valve that controls the EGR flow with α_{EGR} which is a percentage value where 100 means that the valve is fully opened and 0 is fully closed. The relation between the position of α_{EGR} and D_{SP} is shown in Figure 3.15. The sign of α_{EGR} in Figure 3.15 is changed to easier see the relation.

Comparison between α_{EGR} and D_{SP}
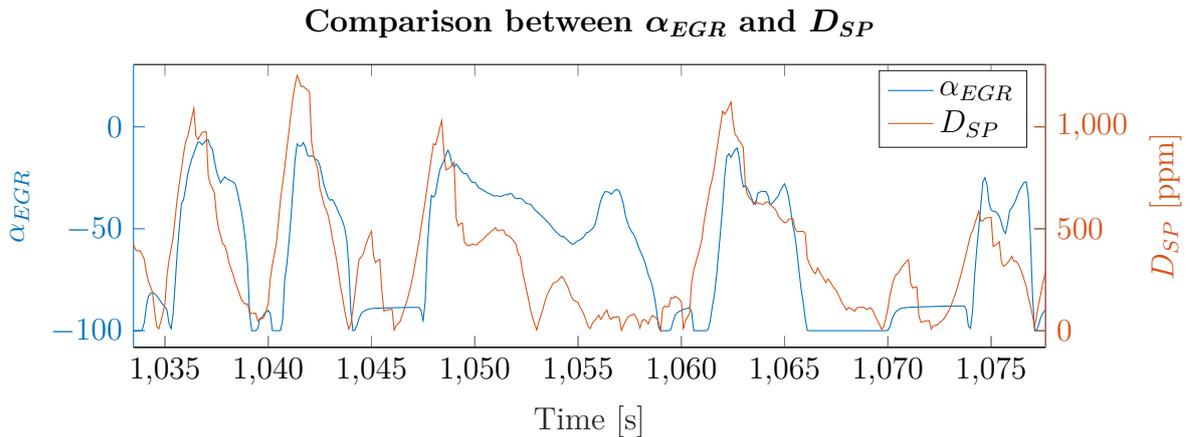
Figure 3.15: Comparison of α_{EGR} and the difference between the Horiba values and estimated values from the semiphysical model. The position is sign changed to easier see the relation.

Similar to C_{corr}, the valve position have somewhat the same behaviour as D_{SP} and should be used when concluding $\widehat{D}_{SP}$.

In addition to the variables that have been found to have a connection to D_{SP}, other variables are also added to the model such as the inverse of an already added variable and the boost temperature T_B. The added variables are

- $\widetilde{T}_B - T_B$

- $1/(\widetilde{T}_B - T_B)$

- $\widetilde{P}_R/P_R$

The boost temperature, T_B, is dependent on a reference value in the semiphysical model. This variable is in the model dependent on the reference temperature, $\widetilde{T}_B$, and the variable are therefore set to $\widetilde{T}_B - T_B$. No graphical relation can be found for the boost temperature but it provides a positive impact on making $\widehat{D}_{SP}$ as similar to D_{SP} as possible. The same goes for $\widetilde{P}_R/P_R$ and $1/(\widetilde{T}_B - T_B)$. The reason to why these variables were considered are because the product between variables can have a relation as well, meaning that the variable itself does not necessary provide a positive impact.

Hundreds of variables are measured and calculated in the ECU and they might affect D_{SP} more or less. However, the ones used for $\widehat{D}_{SP}$ have connection to M_{SP} in one way or another and the other ones from the ECU were never considered.

3.1.3 Radial Basis Empirical Model

The second model used in this thesis uses radial basis functions and are empirically developed. Another difference from the semiphysical model is that this model does not depend on such reference points. Instead, every estimation from the model is calculated from a set of stationary points that describes the behaviour of the NO_x flow as good as possible.

In the same way as for the semiphysical model, only the function for low hp eSCR engine is going to be presented in this section, i.e. $\widehat{D}_{RB}^{low}$. The reason for this is once again that the same behaviour is seen for all engines and the same reasoning is applied on them.

The first variable that is used for the empirical model is P_R, i.e. the rail pressure. The relation between P_R and D_{RB} is shown in Figure 3.16.

28
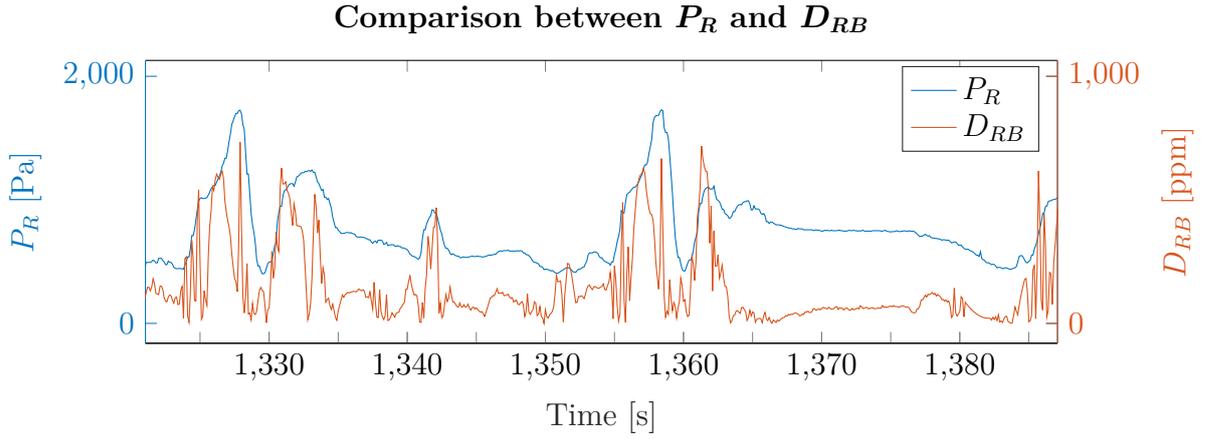
Figure 3.16: Comparison of P_R and the difference between the Horiba values and estimated values from the empirical model.

In Figure 3.16 a relation between the error and P_R can be seen. P_R tends to have the same behaviour as D_{RB}, when D_{RB} is high is P_R high as well. This is enough information to include P_R when making $\widehat{D}_{RB}$.

For the radial basis model the cylinder wall temperature correction factor, C_{corr}, is considered as well. Just as for the semiphysical model, it is inverted in this context. The relation between $1/C_{corr}$ and D_{RB} is shown in Figure 3.17.
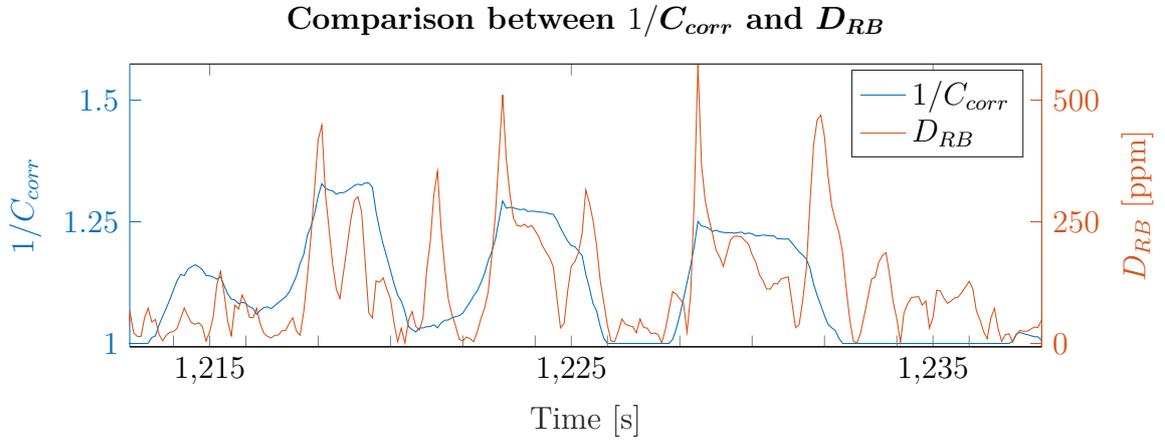
Figure 3.17: Comparison of $1/C_{corr}$ and the difference between the Horiba values and estimated values from the empirical model.

It can be seen in Figure 3.17 that when $1/C_{corr}$ is high D_{RB} is high as well. This is an indicator that there is a relation between the variables and therefore $1/C_{corr}$ should be included when calculating $\widehat{D}_{RB}$.

The EGR valve angle, α_{EGR}, is considered for the radial basis model too. Just as for the semiphysical model α_{EGR} is sign changed when searching for the relation between it and D_{RB}. This is shown in Figure 3.18.

29
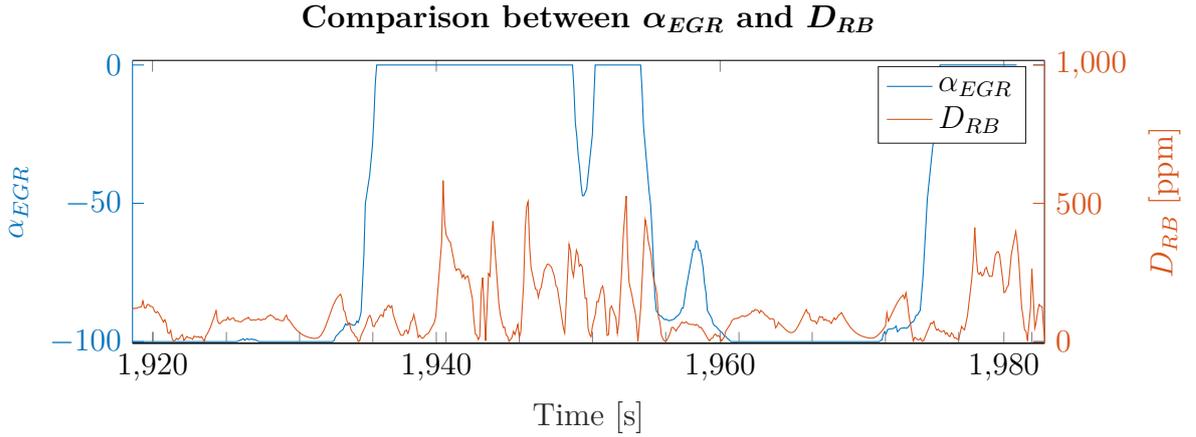
Figure 3.18: Comparison of α_{EGR} and the difference between the Horiba values and estimated values from the empirical model.

As seen in Figure 3.18 D_{RB} tends to be higher when α_{EGR} is close to zero. This is enough information to include α_{EGR} when calculating $\widehat{D}_{RB}$.

The air fuel ratio, AFR is included for the radial basis model as well as for the semiphysical model. The same arguments is valid here as for the semiphysical model where the AFR saturates at its max and at that point D_{RB} is low.
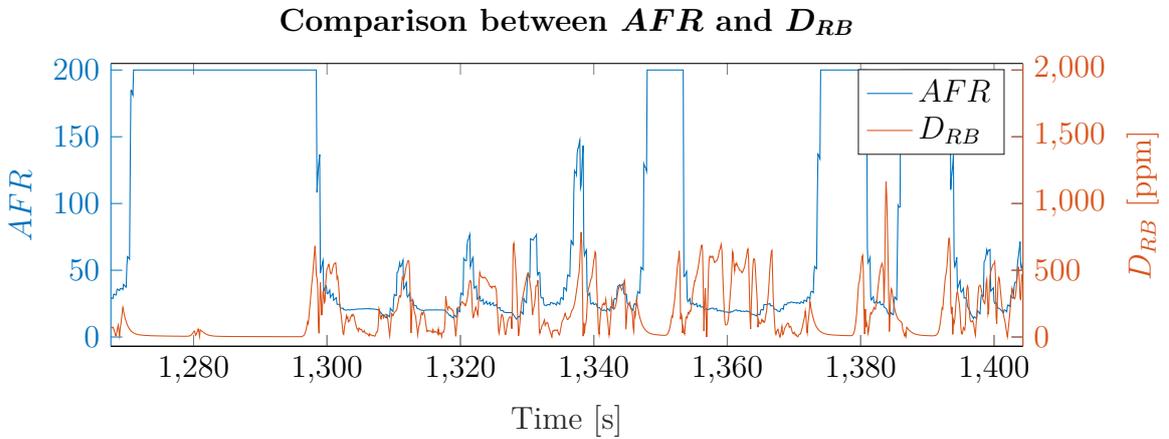
Figure 3.19: Comparison of AFR and the difference between the Horiba values and estimated values from the empirical model.

In Figure 3.19 it can be seen that there is information in the AFR that affects the performance of D_{RB}. Therefore, AFR should be included when calculating $\widehat{D}_{RB}$. In the same way as for the semiphysical case the AFR is here added to the regression as $200 - AFR$ to capture the behaviour that D_{RB} is low when the AFR saturates.

Another variable that is shown to have a relation to D_{RB} is the torque from the engine, τ_{Eng}. The relation between these two is shown in Figure 3.20.
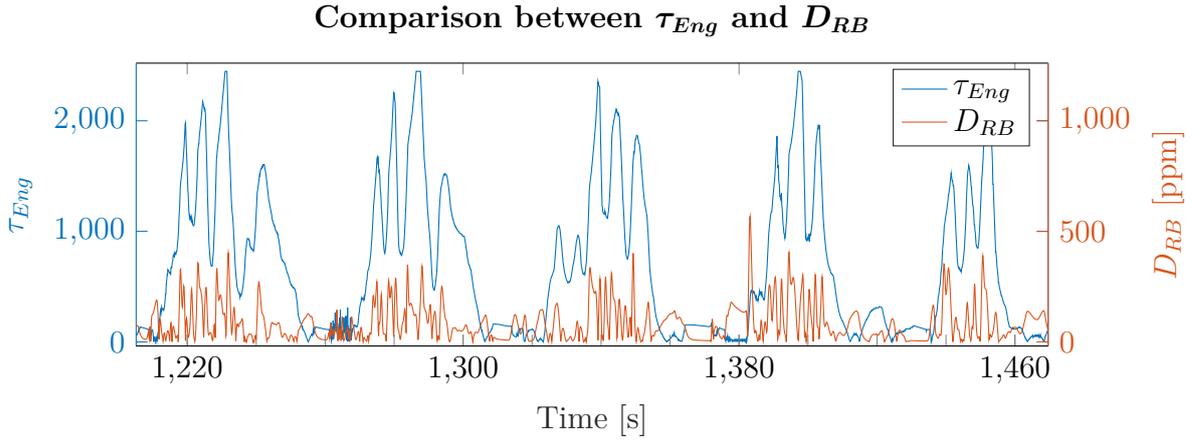
Figure 3.20: Comparison of τ_{Eng} and the difference between the Horiba values and estimated values from the empirical model.

It can be seen in Figure 3.20 that when τ_{Eng} increases in amplitude D_{RB} is increasing as well. The percentage they increase do not match but there is still information that can be useful when constructing $\widehat{D}_{RB}$.

The final variable that is used for $\widehat{D}_{RB}$ is the speed and torque correction variable, $C_{\omega\tau}$. The relation between $C_{\omega\tau}$ and D_{RB} is shown in Figure 3.21.
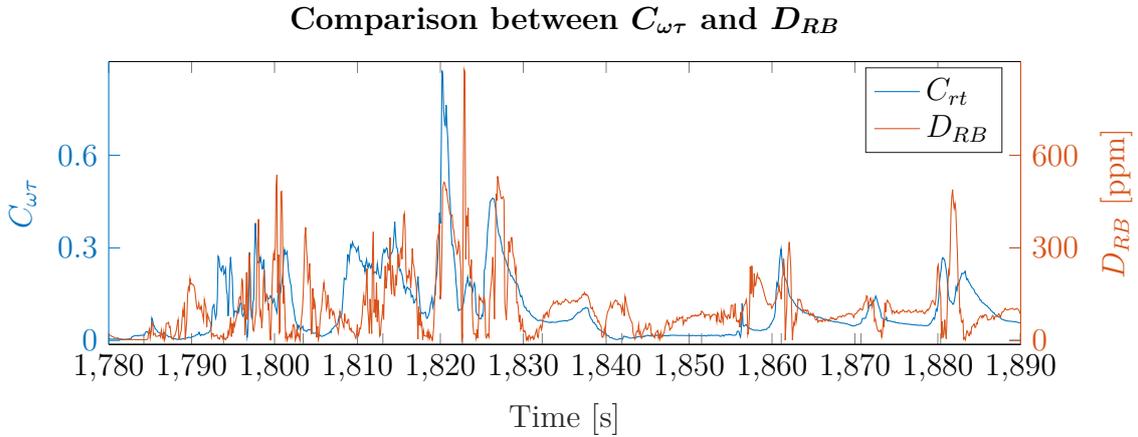
Figure 3.21: Comparison of $C_{\omega\tau}$ and the difference between the Horiba values and estimated values from the empirical model.

In Figure 3.21 a relation between the error and can $C_{\omega\tau}$ be seen. Also here the behaviour between $C_{\omega\tau}$ and D_{RB} is similar and is therefore included in $\widehat{D}_{RB}$ as well.

In addition to the variables that have been found to have a relation to D_{RB}, there are also other variables added to the model such as the inverse of an already existing variable and the boost temperature T_B. The added variables are

- $1/P_R$

- T_B

- $1/T_B$

These variables were proven to have a positive impact on making $\widehat{D}_{RB}$ as similar to D_{RB} as possible.

For the boost temperature T_B no graphical relation can be found but it provides a positive impact on making $\widehat{D}_{RB}$ as similar to D_{RB} as possible. The same goes for $1/P_R$ and $1/T_B$.

Just like the semiphysical model many more variables than the ones used are available but they were never considered.

3.1.4 Equation for Model Accuracy Estimation

The goal of the function $\widehat{D}$ is to be as close to D as possible in terms of engine out NO_x. The creation of $\widehat{D}$ is done using `stepwiselm` in MatLab$^{\copyright}$ which creates a linear regression model using both forward and backward stepwise regression. For this method different types of inputs to $\widehat{D}$ were tested based on the observations done in section 3.1.2 and 3.1.3 to get $\widehat{D}$ to represent D as good as possible. The final set of inputs that were used for $\widehat{D}_{SP}^{low}$ was

$$
\begin{bmatrix}
\tau_{Eng} \\
\omega_{Eng}/\widetilde{\omega}_{Eng} \\
(\widetilde{T}_B - T_B) \\
1/(\widetilde{T}_B - T_B) \\
P_R/\widetilde{P}_R \\
\widetilde{P}_R/P_R \\
1/C_{corr} \\
\alpha EGR \\
(200 - AFR)
\end{bmatrix}
\tag{3.8}
$$

and for $\widehat{D}_{RB}^{low}$

$$
\begin{bmatrix}
\tau_{Eng} \\
T_B \\
1/T_B \\
P_R \\
1/P_R \\
1/C_{corr} \\
\alpha EGR \\
(200 - AFR) \\
C_{\omega\tau}
\end{bmatrix}.
\tag{3.9}
$$

The inputs for all other engines are shown in Appendix A.1.1. These inputs were then used in `stepwiselm` to get the first version of $\widehat{D}_{SP}^{low}$ and $\widehat{D}_{RB}^{low}$. The stepwise regression model calculated a model which uses both the inputs as well as combinations of them to build $\widehat{D}_{SP}^{low}$ and $\widehat{D}_{RB}^{low}$. The number of predictor variables in $\widehat{D}_{SP}^{low}$ and $\widehat{D}_{RB}^{low}$ will therefore most likely exceed the number of inputs to the model.

After this first set of $\widehat{D}_{SP}^{low}$ and $\widehat{D}_{RB}^{low}$, the models were analyzed to improve the performance. This was done by recursively remove those coefficients with the the highest p-value, i.e the ones that have the weakest evidence for rejecting the null hypothesis. This was done to reduce the computational effort of the models. In

Figure 3.22 the calculation time it takes for $\widehat{D}_{SP}^{low}$ to predict an estimate is shown. Here only the method for $\widehat{D}_{SP}^{low}$ is going to be presented but the exact same reasoning is applied on the other combinations of models and engines.
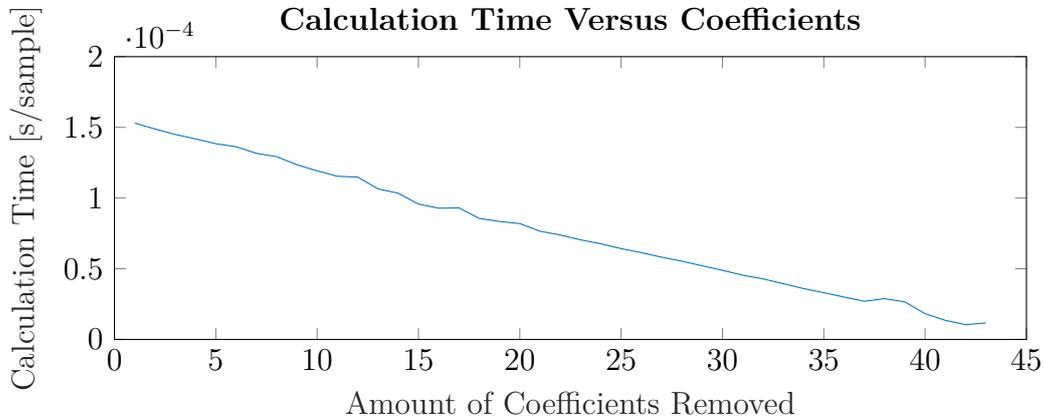
Figure 3.22: The relation between amount of samples in the model and the computational time it takes to predict a sample for $\widehat{D}_{SP}^{low}$. In 0 no coefficients are removed from $\widehat{D}_{SP}^{low}$ and at 43 all coefficients except for the constant term are removed.

It can be seen in Figure 3.22 that the computational time per predicted sample relates linearly to the amount of coefficients in the model. It is therefore wanted to have as few coefficients as possible in the model. The small humps are probably there because of other running tasks on the computer.

To be able to determine how many and which coefficients to remove in the model MSE, NRMSE, and R^2 are measured for different amount of removed coefficients. R^2 is a measurement between 0 and 1 of how good the model can predict data where a high number is preferable. The outcome for NRMSE and R^2 applied on validation data are shown in Figure 3.23. The reason why MSE is not shown is that it follows the inverted behaviour of NRMSE and therefore it is sufficient to only analyze NRMSE.
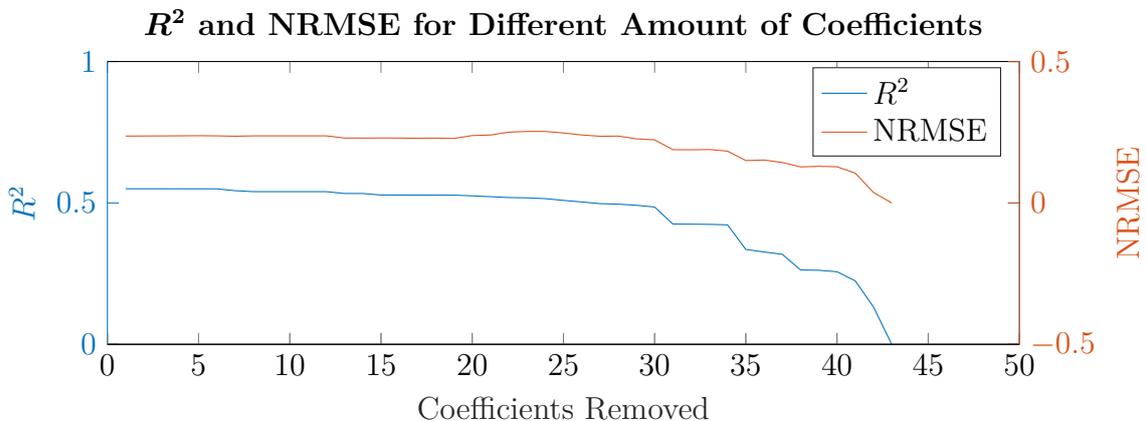
Figure 3.23: R^2 and NRMSE values for different amount of coefficients removed from $\widehat{D}_{SP}^{low}$. At 0 no coefficients are removed from $\widehat{D}_{SP}^{low}$ and at 43 all coefficients except for the constant term are removed.

From Figure 3.23 a trade-off between speed and performance can be done by choosing how many variables to remove. The curves are in the beginning quite flat and removal of these variables with highest p-statistic will not effect the performance much. But if one choose to remove approximately 35 coefficients in this case the performance will be effected negatively but the calculation time will at the same time improve by decreasing the number of coefficients. For this case the amount of removed coefficients was chosen to be six. This number was chosen because the first decline in R^2 was spotted there and a decrease in performance is not wanted.

The method of removing coefficients with the the highest p-value can also be done by changing the acceptance level of the p-value in `stepwiselm` and the retrain the models. But due to that the performance of the models was considered to be good enough, and the training process is very slow, were the models not retrained but instead edited. If the models were retrained, a slight performance improvement in comparison to the edited models are expected. But as mentioned was already the performance considered to be good enough.

To get the final equation for $\widehat{D}_{SP}^{low}$ and $\widehat{D}_{RB}^{low}$ the Cooks distance for all training samples was also analyzed as described in Section 2.1.4. Here outliers were removed and a more accurate model was achieved.

The final equations for $\widehat{D}_{SP}^{low}$ and $\widehat{D}_{RB}^{low}$ are shown in Appendix A.1.2.

3.2   Sensor Fusion

Sensor fusion is the process when different sources of information are combined to make a prediction with less uncertainty than the separate measurements individually. The used sources for this thesis are the sensor and five observer estimations from five different methods that will be described in the following sections. Those five are; FE, DKF, UIO, SSUIO and ESV. No specific engine model is going to be mentioned because the exact reasoning and method presented here is applicable on all engines.

3.2.1   State Observation Using Fraction Estimator

A first estimation of the current NO_x flow, NO_x^{est}, can be done using E_{SP} and E_{RB} in a so called fraction estimator (FE). Depending on the magnitude of E_{SP} and E_{RB} they can separately be chosen to utilize fractions of the estimated NO_x values from the two NO_x flow models. To do this, E_{SP} and E_{RB} are first normalized to better be able to compare them. This is done by dividing E_{SP} and E_{RB} separately with $E_{SP} + E_{RB}$. Then a combined estimation using M_{SP} and M_{RB} can be calculated with

$$NO_x^{est} = \frac{E_{SP}M_{SP} + E_{RB}M_{RB}}{E_{SP} + E_{RB}}. \tag{3.10}$$

One signal that is used to extend equation (3.10) for an even better estimation is the cylinder wall temperature correction factor, C_{corr}. C_{corr} has a value between 1, when it is in steady state, and between 0 and 1 when the system is in a transient phase. Since the NO_x sensor is assumed to be accurate in steady state, and less

accurate when the NO$_x$ flow have transient behaviour, C_{corr} is a good signal to use. The correction is done in a multiplicative manner and in steady state the temperature is multiplied with 1 because no correction is needed. However its value never reaches 0 and therefore a slack γ is added to get a better balance of the weighting between the sensor and the combined estimation NO_x^{est}. The values of the slack variables for the three different engines are seen in Appendix A.2.1. The sensor and NO_x^{est} are then fused together with C_{corr} to provide an as good estimate as possible. How this is done is illustrated in Figure 3.24.
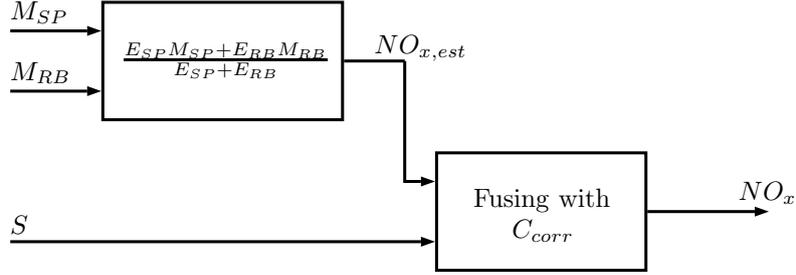
Figure 3.24: Description of how the estimation of engine out NO$_x$ is done using the two models and the sensor. The model values are first weighted, corresponding to how accurate they are, and then the combined model estimation, NO_x^{est}, is fused with the sensor value using C_{corr}.

From Figure 3.24 with equation (3.10) a first observer can be formulated as

$$NO_x = (1 - C_{corr} + \gamma)\left(\frac{E_{SP}M_{SP} + E_{RB}M_{RB}}{E_{SP} + E_{RB}}\right) + (C_{corr} - \gamma)S. \qquad (3.11)$$

A problem with this approach is that the NO$_x$ estimation can become unstable if E_{SP} or E_{RB} are negative. In reality E_{SP} and E_{RB} are always positive but this information is not known by the regression models even if all the training data are strictly larger than zero. The operating points where the models return negative number occurs when the true values of E_{SP} or E_{RB} are small. Therefore these values are saturated to a small value, which in this case is set to 10 to prevent the unstable behaviour. This saturation is also used in the upcoming observers where (3.10) is used.

Another approach of handling the negative values of E_{SP} or E_{RB} could be to use $|E_{SP}|$ or $|E_{RB}|$. The cases when E_{SP} or E_{RB} become negative was when the true values were close to zero. The negative values was in this points a small spike below zero. If the absolute value was applied on this signal would a positive spike be appearing rather than zero which is the true value. E_{SP} or E_{RB} was then set to a small value instead of using the absolute value because this should be closer to the true value.

3.2.2 State Observation Using Linear Kalman Filter With Delayed Measurements

When designing an observer, a state space model is first implemented. For this observer, the system is represented as a random walk model where the state vector,

x, is just one state; the current NO_x flow. The reason for this is that there is no memory in the observed state. The current NO_x flow is only dependent on the current operating point and not on the earlier NO_x values. Therefore there is no reason to extend the system model to a constant velocity or a constant acceleration model. Following this the state space system is formulated as

$$\begin{aligned} x_{k+1} &= x_k + q_k \\ y_k &= C_k x_k + r_k \end{aligned} \tag{3.12}$$

at time k and where q and r are the independent Gaussian noises

$$q_k \sim \mathcal{N}(0, 1) \qquad r_k \sim \mathcal{N}(0, R_k). \tag{3.13}$$

In the same way as C will change size depending on number of inputs the covariance matrix R will do that as well. R is used when the sensor value is not available and R^* is used when the sensor value is available as described in equation (2.13) and (2.14). There will be no control inputs to the model and because of that B and D from equation (2.7) are zero and removed from the equation.

The linear Kalman filter with delayed measurements will be using the random walk model, but with the method described in Section 2.3 instead. In this strategy all sources of information are fused in one estimator simultaneously. This strategy is shown in Figure 3.25.
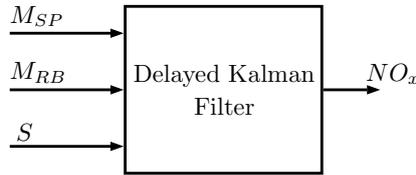
Figure 3.25: Fusion of information using one estimator where the models and the sensor are fused together simultaneously.

For this observer the measurement noise covariances are defined as

$$R_k = \begin{bmatrix} a \cdot D_{SP,k}^2 & 0 \\ 0 & b \cdot D_{RB,k}^2 \end{bmatrix} \qquad R_k^* = \begin{bmatrix} a \cdot D_{SP,k}^2 & 0 & 0 \\ 0 & b \cdot D_{RB,k}^2 & 0 \\ 0 & 0 & c \cdot F_{S,k} \end{bmatrix} \tag{3.14}$$

$D_{SP,k}$ and $D_{RB,k}$ are squared due to that they are the expected value of the error and not the variance of the error. $F_{S,k}$ is the error function for the sensor that is tuned to match the error functions for the models. It is set to

$$F_{S,k} = \frac{1}{C_{corr}}. \tag{3.15}$$

a, b and c are tuning variables that are tuned to achieve as good fit to validation data as possible. The reason for that a and b are the same in R_k and R_k^* is because these reflect the relation between the accuracy of the models. This accuracy relationship should be the same independent of the presence of a sensor in the measurement equation.

As earlier stated, C_{corr} is a good measurement of steady state behaviour which means that it is also a good estimation for the sensor covariance as the uncertainty will be higher with a lower value of C_{corr}. However, it does not have the same magnitude as $D_{RB,k}$ and $D_{SP,k}$ and needs to be tuned accordingly by tuning c.

The reason for having the tuning parameters a and b and not use $D_{RB,k}$ and $D_{SP,k}$ directly is that the models are not perfect. They are both trained on data that have uncertainty and this will decrease the performance of the models. By tuning the parameters a and b the impact of bad models can be decreased because the uncertainty of the models will be captured in a and b.

When the Kalman filter is updating $\hat{x}_s$ with information from the sensor in time k, the measurement model in (3.12) is extended to

$$y_s = \begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix} x_s + r_k^* \tag{3.16}$$

but as mentioned it will use

$$y_k = \begin{bmatrix} 1 \\ 1 \end{bmatrix} x_k + r_k \tag{3.17}$$

in all other calculation steps.

3.2.3 State Observation Using an Unknown Input Observer

In this section an UIO as explained in Section 2.4 in combination with parts from the FE is developed. The strategy of this observer is shown in Figure 3.26.
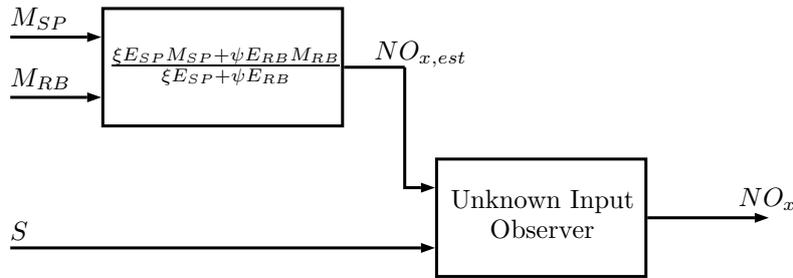
Figure 3.26: Description of how the estimation of engine out NO_x is done using the two models and the sensor. A first estimation is done with the FE and then it is fused with the sensor value using an unknown input observer.

The model values that are used in the UIO are the fused estimates from M_{SP} and M_{RB}. This is done in the same manner as in (3.10) and only a single estimate NO_x^{est} from the models is used. In the same way as a, b and c were implemented for tuning of the DKF, ξ and ψ are here introduced to further capture the difference between the model performance. These tuning parameters extends (3.10) to

$$NO_x^{est} = \frac{\xi(E_{SP} M_{SP}) + \psi(E_{RB} M_{RB})}{\xi E_{SP} + \psi E_{RB}}. \tag{3.18}$$

37

These new values will then be fused with the sensor values in the UIO, as shown in Figure 3.26. One major reason for this is to enable the possibility to have a steady state solution to the UIO. It is possible to do a steady state solution to the Kalman filter if $E_{RB,k}$ and $E_{SP,k}$ are excluded in the Kalman filter calculations and instead used for fusing model values. By doing these two things, the pros with $E_{RB,k}$ and $E_{SP,k}$ are kept and at the same time is a steady state Kalman filter possible.

The representation of the delayed sensor is an important aspect of this observer. The sensor dynamics will here be modeled as

$$H^*(s) = \frac{1}{\tau s + 1}, \tag{3.19}$$

in other words, a low pass filter where τ is the time constant and s is the Laplace variable. A low pass filter is a simple way to model the sensor delay and that is the reason for the usage here. This can be rewritten as

$$H^*(s) = \frac{\frac{1}{\tau}}{s + \frac{1}{\tau}}. \tag{3.20}$$

For the filter to be applicable in this thesis a corresponding discrete time filter must be calculated. The pole of $H^*(s)$ is $s = -\frac{1}{\tau}$ and the discrete filter should have the same pole placement. This implies that the discrete system should have a pole in $z = e^{-\frac{Ts}{\tau}}$ where Ts is the sampling rate. This means that the discrete filter transfer function is

$$H(z) = \frac{G}{z - e^{-\frac{Ts}{\tau}}} \tag{3.21}$$

where G is still to be concluded. To determine G the low frequency behaviour of both $H(s)$ and $H(z)$ are analyzed. They are both transformed to the frequency domain and becomes

$$H^*(j\omega) = \frac{\frac{1}{\tau}}{j\omega + \frac{1}{\tau}}$$
$$H(j\omega) = \frac{G}{e^{j\omega} - e^{-\frac{Ts}{\tau}}}. \tag{3.22}$$

This makes

$$\left. \begin{array}{l} \lim_{\omega \to 0} H^*(j\omega) = 1 \\ \lim_{\omega \to 0} H(j\omega) = \frac{G}{1 - e^{-\frac{Ts}{\tau}}} \end{array} \right\} \implies 1 = \frac{G}{1 - e^{-\frac{Ts}{\tau}}} \iff G = 1 - e^{-\frac{Ts}{\tau}} \tag{3.23}$$

The final discrete time low pass filter is then formulated as

$$H(z) = \frac{1 - e^{-\frac{Ts}{\tau}}}{z - e^{-\frac{Ts}{\tau}}}. \tag{3.24}$$

For the UIO a state space model with unknown inputs needs to be created. For each unknown input an extra state is added and a system description of this is seen in Figure 3.27.
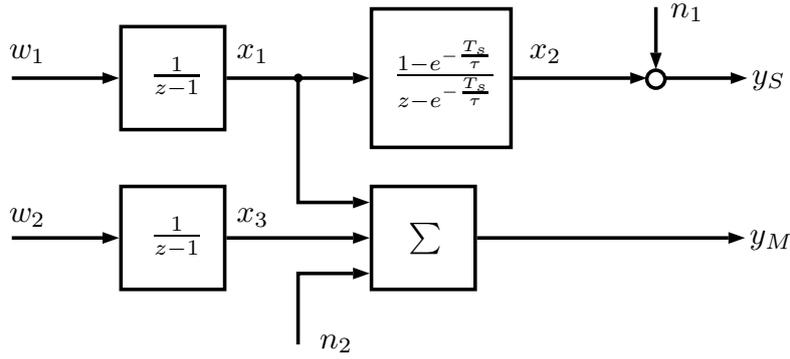
Figure 3.27: System description of the unknown input observer.

There are three states in the system where x_1 is the true NO_x flow, x_2 is the delayed sensor value without noise and x_3 is the bias of the models. Both x_1 and x_3 are driven by white process noises w_1 and w_2 that are integrated. This will change them to random walk models that are used to model the unknown input which in this case are the true state and the true model values without bias and noise. n_1 and n_2 are sensor noise and model noise respectively.

The state space model for the UIO can now be formulated as

$$x_{k+1} = \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix}_{k+1} = \begin{bmatrix} 1 & 0 & 0 \\ 1-e^{-\frac{T_s}{\tau}} & e^{-\frac{T_s}{\tau}} & 0 \\ 0 & 0 & 1 \end{bmatrix} x_k + \begin{bmatrix} 1 & 0 \\ 0 & 0 \\ 0 & 1 \end{bmatrix} \cdot \begin{bmatrix} w_1 \\ w_2 \end{bmatrix}$$

$$y_k = \begin{bmatrix} y_S \\ y_M \end{bmatrix} = \begin{bmatrix} 0 & 1 & 0 \\ 1 & 0 & 1 \end{bmatrix} x_k + \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \cdot \begin{bmatrix} n_1 \\ n_2 \end{bmatrix}$$

(3.25)

where x_1 will be the tracked state.

3.2.4 Steady State Solution of a Unknown Input Observer

Only one of the developed observers will be implemented in TargetLink and this will be the UIO described in Section 3.2.3. The reason for this is that the UIO have been seen to have good performance in comparison to the other observers and is not computational demanding. The metrics behind this decision is shown in Chapter 4. It is once again a trade off between speed and performance that have to be done. The observer needs to be feasible in the Volvo software and therefore the steady state solution of the UIO (SSUIO) is wanted. The steady state solution is computationally cheaper than the iterative solution to the filter. Let the value of the steady state Kalman gain be $\bar{K}$. This value will now be used instead of the iterative solution.

3.2.5 State Observation Using an Extended State Vector

The final observer strategy is to extend the state vector (ESV) to include delayed states from k to s. An argument for this is that the delayed sensor measurement

can observe a state directly and the current state will be affected by it through the system dynamics. After the expansion, the state vector expanded to

$$\mathbb{X}_k = \begin{bmatrix} x_k \\ x_{k-1} \\ \vdots \\ x_{s+1} \\ x_s \end{bmatrix} \tag{3.26}$$

where the delayed sensor value is measured in time s. The state space, using this state vector is

$$\begin{aligned} \mathbb{X}_{k+1} &= A\mathbb{X}_k + q_{k+1} \\ y_k &= C\mathbb{X}_k + r_k. \end{aligned} \tag{3.27}$$

In the same way as for the UIO the Kalman filter used is an ordinary Kalman filter and the system description is instead the feature investigated here.

The strategy of this observer is shown in Figure 3.28.
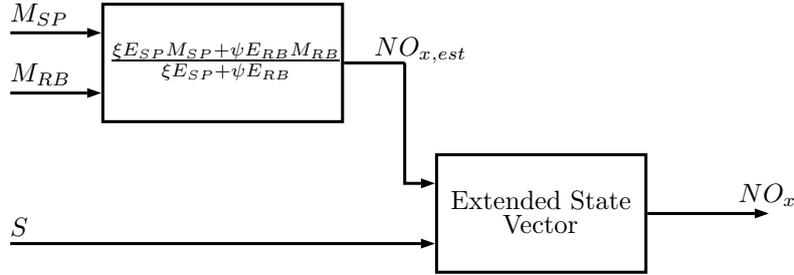
Figure 3.28: Description of how the estimation of engine out NO_x is done using the two models and the sensor. A first estimation is done with the FE and then it is fused with the sensor value using an extended state vector description in a Kalman filter.

Following the description in Figure 3.28 only two measurements are going to be used in the Kalman filter, the combined model value and the sensor measurement. A and C in (3.27) can now be formulated as

$$A = \begin{bmatrix} 0 & 1 & 0 & 0 & & 0 \\ 0 & 0 & 1 & 0 & \cdots & 0 \\ 0 & 0 & 0 & 1 & & 0 \\ & \vdots & & & \ddots & \vdots \\ 0 & 0 & 0 & 0 & \cdots & 1 \\ 0 & 0 & 0 & 0 & \cdots & 0 \end{bmatrix} \qquad C = \begin{bmatrix} 1 & 0 & \cdots & 0 \\ 0 & 0 & \cdots & 1 \end{bmatrix}. \tag{3.28}$$

The noise q is tuned as an increasing diagonal with the highest value in the bottom right element. This is due to that the uncertainty should increase for each step back in time.

The noise covariance R in r is constructed as

$$R = \begin{bmatrix} \frac{d}{C_{corr}} \cdot a & 0 \\ 0 & b \end{bmatrix} \tag{3.29}$$

where d is the delay, a and b are tuning variables. C_{corr} is used to weigh the sensor more or less depending on if the system is in steady state or not. The uncertainty of the sensor should also be increased when the delay is increased and that is why d is included as well.

4
Results

In this chapter the results of the thesis are presented. First all six functions (three engine types times two NO_x flow models) are evaluated against their respectively true NO_x flow. After that the results from the sensor fusion are presented, first with a developed fraction estimator, then with Kalman filters with delayed measurements, then with the unknown input observer both for iterative and steady state Kalman gain and lastly the Kalman filter with extended state vector.

All results are evaluated and compared to the validation data of the datasets and it have been executed on a laptop with Intel i7 processor in MATLAB© 2015.

4.1 Evaluation of the Estimated Difference Between Models and Horiba

In this section the performance of both $\widehat{D}_{SP}$ and $\widehat{D}_{RB}$ will be shown for all engines. The evaluation will be in terms of MSE in comparison to D_{SP} and D_{RB} and NRMSE. The combined performance of the models will then be presented where the amount of times the correct model was chosen are shown.

4.1.1 Evaluation of the Semiphysical Model

As mentioned in Section 3.1.1 cropping of non representative data might be good when finding the MSE over datasets. Therefore three crop sizes were tested. One for 100 % of the idle driving remaining, one for $\sim$ 20 % remaining and one for 0 % remaining idle driving. How the MSE is changed based on how much of the non representative data that is cropped is seen in Table 4.1 for all three engines.

Table 4.1: MSE for different amounts of cropping and engines for $\widehat{D}_{SP}$, measured in [ppm^2].

	$\widehat{D}_{SP}^{low}$	$\widehat{D}_{SP}^{high}$	$\widehat{D}_{SP}^{TC}$
100%	2218	7447	5006
20%	2332	6241	5574
0%	2340	6144	5617

Even though the MSE was decreased for 100% for two out of three engines the one with 0% will be used. This is because it do not compensate for the fact that the model is trained on idle behaviour. The reason why it does not increase or decrease

for all three engines is because the idle behaviour was caught very well for $\widehat{D}_{SP}^{low}$ and $\widehat{D}_{SP}^{TC}$ while there was a constant offset in the idle parts between $\widehat{D}_{SP}^{high}$ and D_{SP}^{high}. This makes the MSE higher than for the other two where the offset did not exist.

The following three Figures 4.1-4.3 compares the calculated function $\widehat{D}_{SP}$ with D_{SP} for each engine model. In Figure 4.1 the comparison between $\widehat{D}_{SP}^{low}$ and D_{SP}^{low} is shown.

Comparison Between $\widehat{D}_{SP}^{low}$ and D_{SP}^{low}
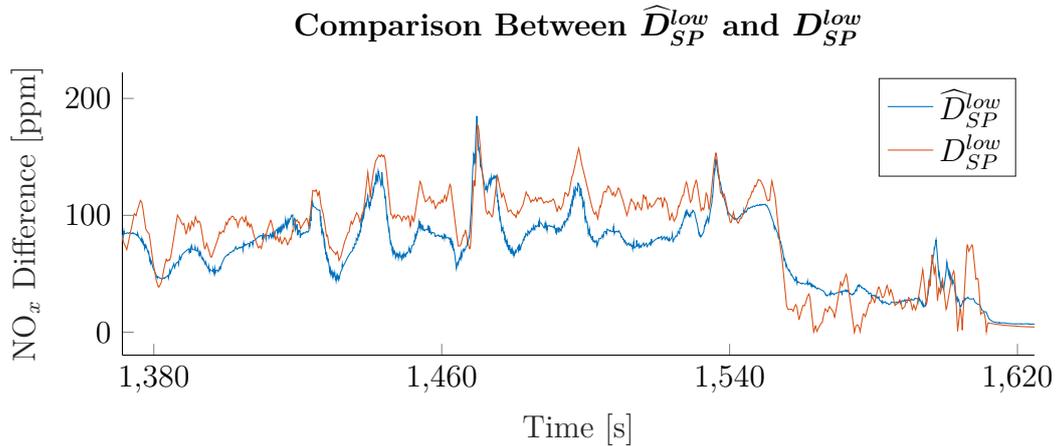
Figure 4.1: Comparison between $\widehat{D}_{SP}^{low}$ and D_{SP}^{low}. $\widehat{D}_{SP}^{low}$ manages to describe the behaviour of D_{SP}^{low} rather well. They have the same characteristics in both amplitude and in transients.

In Figure 4.2 the comparison between $\widehat{D}_{SP}^{high}$ and D_{SP}^{high} is shown.

Comparison Between $\widehat{D}_{SP}^{high}$ and D_{SP}^{high}
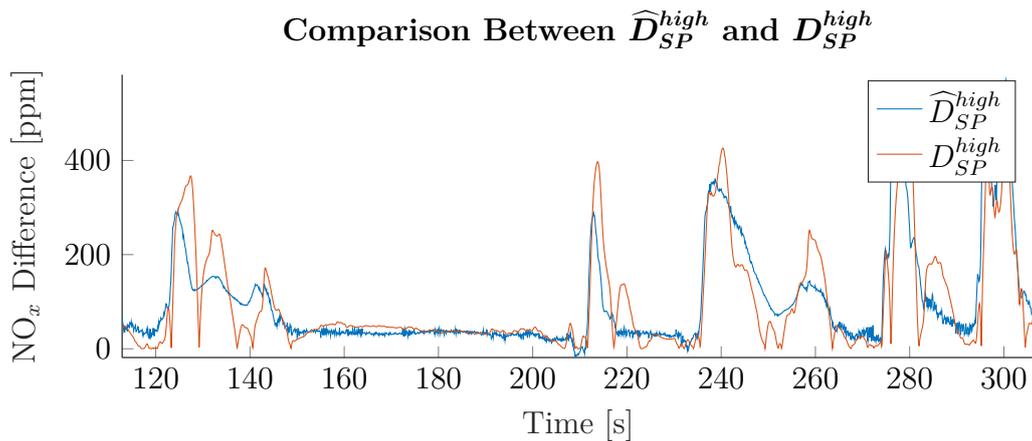
Figure 4.2: Comparison between $\widehat{D}_{SP}^{high}$ and D_{SP}^{high}. $\widehat{D}_{SP}^{high}$ manages to describe the behaviour of D_{SP}^{high} well. They have the same characteristics in both amplitude and in transients.

In Figure 4.3 the comparison between $\widehat{D}_{SP}^{TC}$ and D_{SP}^{TC} is shown.
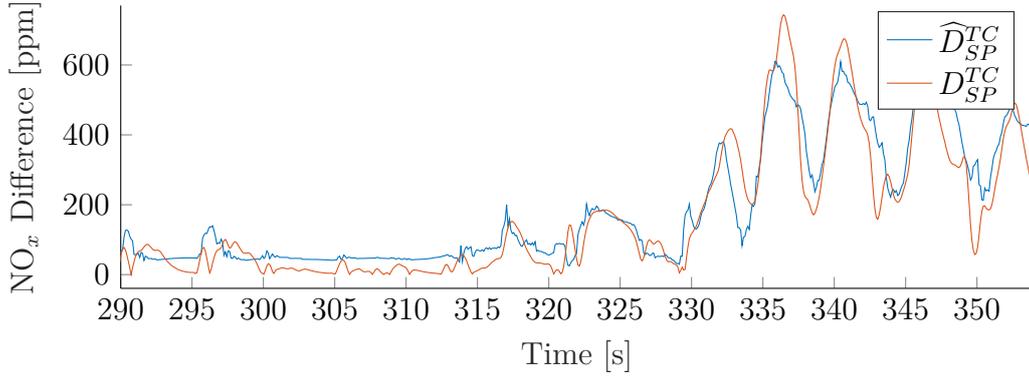
Figure 4.3: Comparison between $\widehat{D}_{SP}^{TC}$ and D_{SP}^{TC}. $\widehat{D}_{SP}^{TC}$ manages to describe the behaviour of D_{SP}^{TC} rather well. They have the same characteristics in both amplitude and in transients.

As seen in Figure 4.1-4.3 the calculated functions for $\widehat{D}_{SP}$ manages to mimic D_{SP} rather well based on the complexity the system have for all three engine models. The same characteristics are seen both for amplitude and in transients.

The MSE against D_{SP} and how far on average $\widehat{D}_{SP}$ deviates from D_{SP} are measured for the developed models. These metrics are shown in Table 4.2.

Table 4.2: Metrics describing the performance of $\widehat{D}_{SP}$.

	$\widehat{D}_{SP}^{low}$	$\widehat{D}_{SP}^{high}$	$\widehat{D}_{SP}^{TC}$
MSE [ppm^2]	2456	3398	8607
Average Deviation [ppm]	27.0	37.4	64.7
Average Deviation [%]	19.2	24.4	42.4

From Table 4.2 it is seen that the metrics for $\widehat{D}_{SP}^{TC}$ are highest and this is also confirmed from Figure 4.1 - 4.3 where the y-axis for 4.3 has a higher amplitude than the other two which leads to a higher MSE and average deviation.

4.1.2 Evaluation of the Radial Basis Model

The data was cropped for the radial basis model as well as for the semiphysical model. Also here the same three crop sizes were tested. The resulting MSE for each engine model and how much data removed are seen in Table 4.3.

Table 4.3: MSE for different amounts of cropping and engines for $\widehat{D}_{RB}$, measured in [ppm^2].

	$\widehat{D}_{RB}^{low}$	$\widehat{D}_{RB}^{high}$	$\widehat{D}_{RB}^{TC}$
100%	1944	2402	4321
20%	1909	2197	4389
0%	1856	2160	4418

As seen in Table 4.3 the MSE is decreased when all of the idle driving is removed for two out of three engines and therefore it is excluded from the datasets when training the best $\widehat{D}_{RB}$ as well.

The following three Figures 4.4-4.6 compares the calculated function $\widehat{D}_{RB}$ with D_{RB} for each engine model. In Figure 4.4 the comparison between $\widehat{D}_{RB}^{low}$ and D_{RB}^{low} is shown.
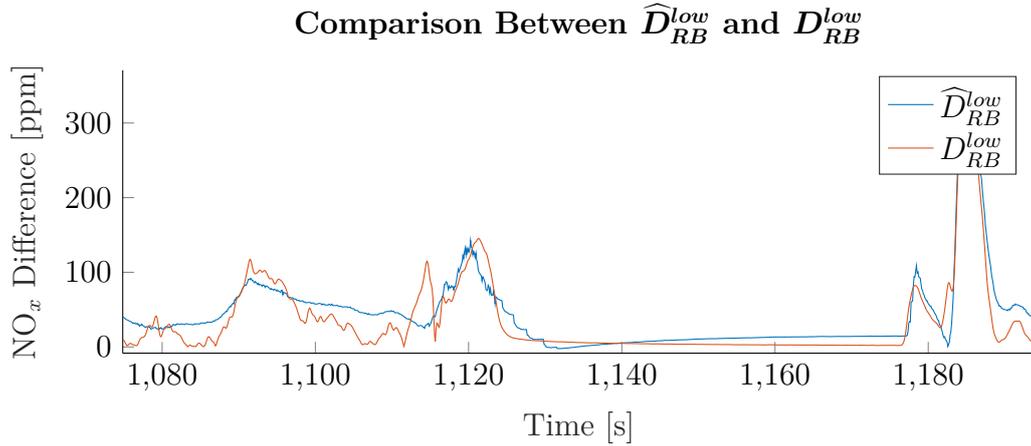
Figure 4.4: Comparison between $\widehat{D}_{RB}^{low}$ and D_{RB}^{low}. $\widehat{D}_{RB}^{low}$ manages to describe the behaviour of D_{RB}^{low} rather well. They have the same characteristics in both amplitude and in transients.

In Figure 4.5 the comparison between $\widehat{D}_{RB}^{low}$ and D_{RB}^{low} is shown.
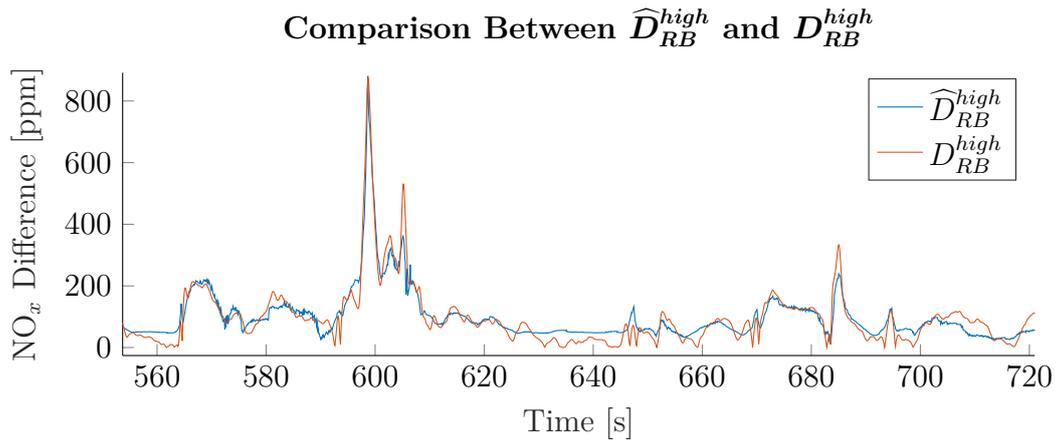
Figure 4.5: Comparison between $\widehat{D}_{RB}^{high}$ and D_{RB}^{high}. $\widehat{D}_{RB}^{high}$ manages to describe the behaviour of D_{RB}^{high} rather well. They have the same characteristics in both amplitude and in transients.

In Figure 4.6 the comparison between $\widehat{D}_{RB}^{TC}$ and D_{RB}^{TC} is shown.
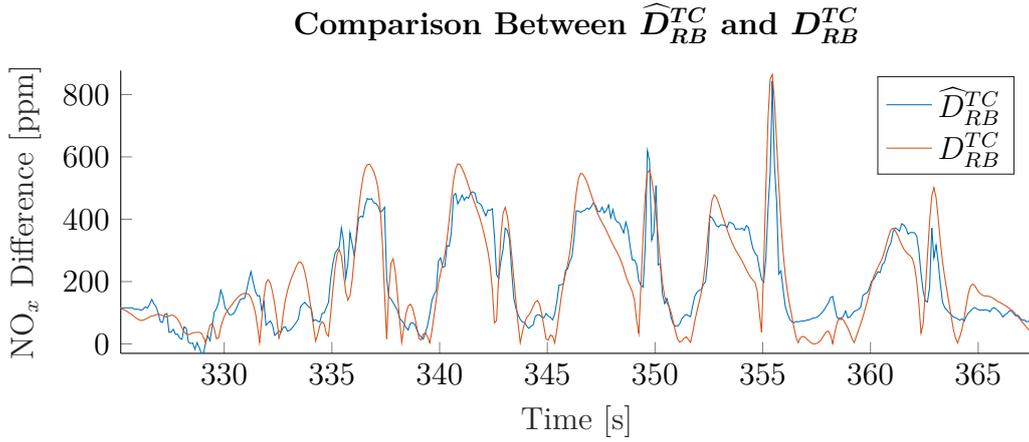
Figure 4.6: Comparison between $\widehat{D}_{RB}^{TC}$ and D_{RB}^{TC}. $\widehat{D}_{RB}^{TC}$ manages to describe the behaviour of D_{RB}^{TC} rather well. They have the same characteristics in both amplitude and in transients.

As seen in Figure 4.4-4.6 the calculated functions for $\widehat{D}_{RB}$ manages to mimic D_{RB} rather well based on the complexity the system have for all three engine models. The same characteristics are seen both for amplitude and in transients.

The metrics collected for these developed models are both the MSE against D_{RB} and how far on average $\widehat{D}_{RB}$ deviates from D_{RB}. This data is shown in Table 4.4.

Table 4.4: Metrics describing the performance of $\widehat{D}_{RB}$.

	$\widehat{D}_{RB}^{low}$	$\widehat{D}_{RB}^{high}$	$\widehat{D}_{RB}^{TC}$
MSE [ppm^2]	2149	2052	8461
Average Deviation [ppm]	24.9	25.0	56.8
Average Deviation [%]	16.5	16.9	36.9

From Table 4.4 it is seen that the metrics for $\widehat{D}_{RB}^{TC}$ are highest and this is also confirmed from Figure 4.4 - 4.6 where the y-axis for Table 4.4 has a lower amplitude than the other two and $\widehat{D}_{RB}^{high}$ follows D_{RB}^{high} more precise than how $\widehat{D}_{RB}^{TC}$ follows D_{RB}^{TC}. This leads to a higher MSE and average deviation.

4.1.3 Evaluation of the Developed Models Using Normalized Root Mean Squares

To further evaluate the performance of the models the NRMSE can be used. The corresponding NRMSE for each model is shown in Table 4.5.

Table 4.5: The NRMSE for all engine models and both developed models. The values are all over 0 and this indicates that the developed models are better than the trivial solution.

	NRMSE	
	$\widehat{D}_{SP}$	$\widehat{D}_{RB}$
Low hp	0.2189	0.3326
High hp	0.3566	0.4679
TC	0.5795	0.2908

The values in Table 4.5 shows that all models are better than the trivial solution, a constant line fitted to the mean of the true values. They are not a perfect fit but that was expected due to that the inputs to the models do not perfectly match D, as seen in Section 3.1.2 and 3.1.3. The result was considered to be good enough to use in the upcoming observers because all NRMSE values was larger than zero. The reason to why the NRMSE for the semiphysical model on TC is higher than the other two is because of that D_{SP} performs poorly and have a high difference. This means that even if $\widehat{D}_{SP}$ performs averagely it will be better than the mean of D_{SP} and the NRMSE will be closer to one than to zero.

4.1.4 Evaluation of the Combined Performance of the Developed Models

Two evaluation methods are used when evaluating the combined performance of $\widehat{D}_{SP}$ and $\widehat{D}_{RB}$ for respective engine. One is by comparing D with $\widehat{D}$ to see how often the right model is chosen to represent the difference. The second method is to evaluate how often the correct model was weighted most in those cases where it matters the most, i.e. where $\widehat{D}_{SP}$ or $\widehat{D}_{RB}$ have high values.

The cooperation between the two models developed for each engine is very important. In some cases one of the models could be inaccurate and in that case the observers should rely more on the other model. When to listen to which NO_x flow model can be described with

$$\zeta = D_{SP} - D_{RB}. \tag{4.1}$$

When ζ is larger than 0 D_{SP} is larger than D_{RB}, which means that the semiphysical model differ more from the Horiba measurements than the radial basis model. In that case it is desirable to take the radial basis model more into account in the observer. The same behaviour of ζ is desirable when the estimated models $\widehat{D}_{SP}$ and $\widehat{D}_{RB}$ are used according to

$$\widehat{\zeta} = \widehat{D}_{SP} - \widehat{D}_{RB}. \tag{4.2}$$

A metric for the performance of $\widehat{D}_{SP}$ and $\widehat{D}_{RB}$ can then be described as the function when ζ and $\widehat{\zeta}$ suggest to listen to the same model the most. This condition can be described by

$$sign\left(\zeta\right) = sign\left(\widehat{\zeta}\right). \tag{4.3}$$

48

In Figure 4.7 ζ and $\widehat{\zeta}$ are shown respectively for the low hp eSCR engine. The goal is to make the two signals as similar as possible and at least on the same side of the dashed red line.
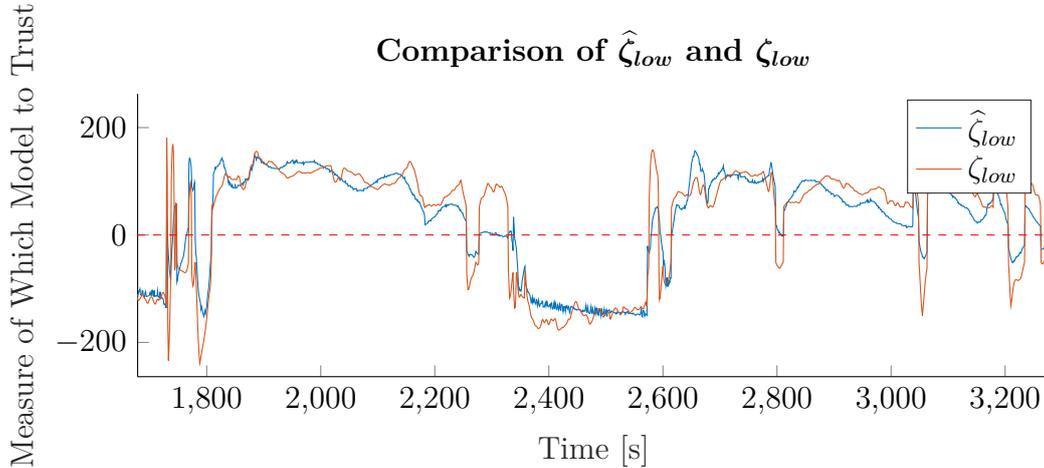
Figure 4.7: Comparison of the choice of model that $\widehat{\zeta}_{low}$ and ζ_{low} indicate. The goal is to make the two signals as similar as possible and at least on the same side of the dashed red line. If they are on different sides the $\widehat{\zeta}_{low}$ and ζ_{low} are suggesting different models.

It can be seen in Figure 4.7 that $\widehat{\zeta}_{low}$ is following ζ_{low} well. One of the most important things here is that they are located on the same side of the dashed red line. If they are on different sides $\widehat{\zeta}_{low}$ and ζ_{low} are suggesting to listen to different models.

In Figure 4.8 ζ and $\widehat{\zeta}$ are shown respectively for the high hp eSCR engine.
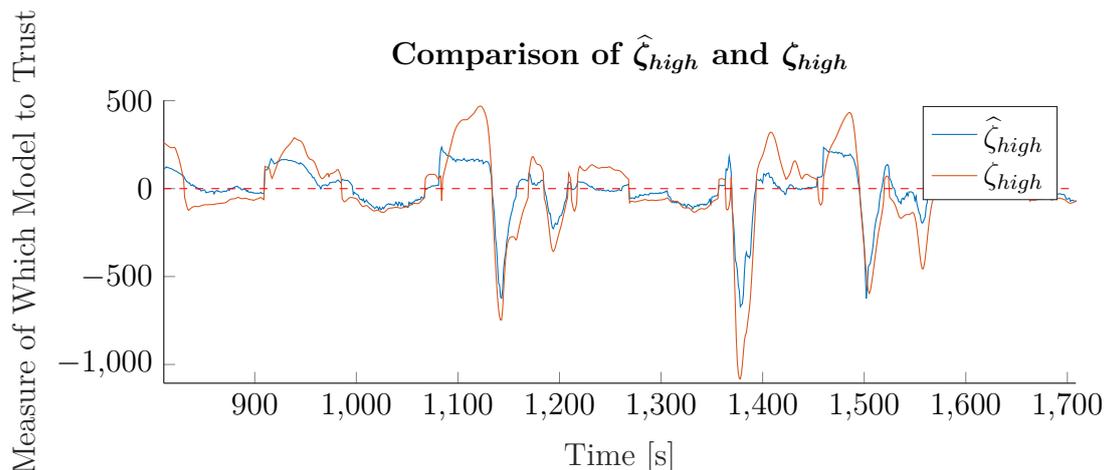
Figure 4.8: Comparison of the choice of model that the developed $\widehat{\zeta}_{high}$ and ζ_{high} indicates.

It can be seen in Figure 4.8 that $\widehat{\zeta}_{high}$ is following ζ_{high} well. They are located

on the same side of the dashed red line most of the time and have the same characteristics but $\widehat{\zeta}_{high}$ does not quite reach the peaks of ζ_{high}.

In Figure 4.9 ζ and $\widehat{\zeta}$ are shown respectively for the 500 hp TC engine. Note that the values are almost always over zero here which means that D_{RB} is estimating the NO_x flow better for this engine model.
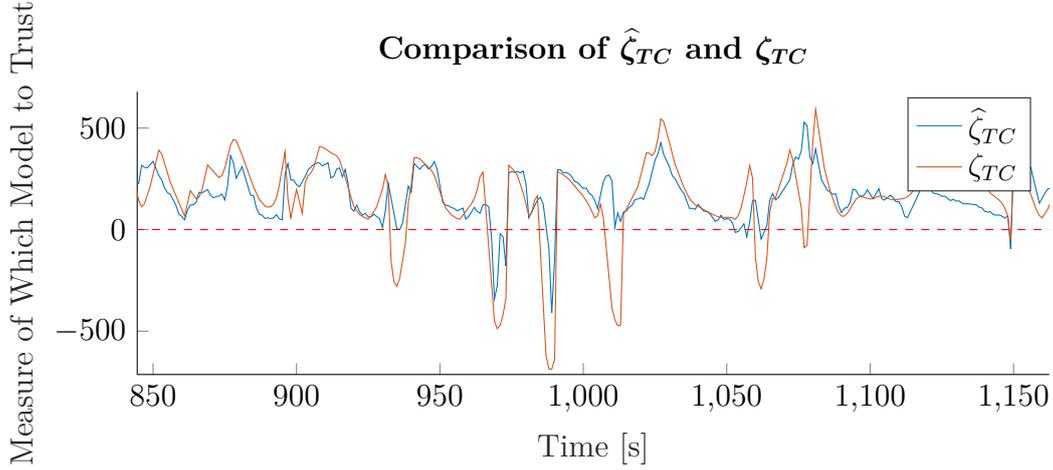
Figure 4.9: Comparison of the choice of model that the developed $\widehat{\zeta}_{TC}$ and ζ_{TC} indicate.

It can be seen in Figure 4.9 that $\widehat{\zeta}_{TC}$ is following ζ_{TC} well. They are located on the same side of the dashed red line most of the time and have the same characteristics expect for on a few of the peaks.

In Table 4.6 is the percentage of how often $\widehat{\zeta}$ and ζ wants to listen to the same model, described by (4.3), for respective engine.

Table 4.6: Percentage of how $\widehat{\zeta}$ and ζ wants to listen to the same NO_x flow model for each engine model.

	low hp	high hp	TC
Accuracy	93.59%	84.56%	94.94%

The TC data can here be a bit misleading because one model is clearly better than the other. In Figure 4.9 both ζ_{TC} and $\widehat{\zeta}_{TC}$ most of the time are positive which means that D_{SP}^{TC} is larger than D_{RB}^{TC}. This means that the precision of M_{SP} is lower than M_{RB} for the most part. Even with bad $\widehat{D}_{SP}^{TC}$ and $\widehat{D}_{RB}^{TC}$ that suggest to pick M_{RB} in all points will a good percentage for TC be achieved when one model has much less accuracy than the other. This is not a problem here though because the performance can be validated through Figure 4.3 and Figure 4.6 where $\widehat{D}^{TC}$ seems to follow well.

The most crucial part of the cooperation between $\widehat{\zeta}$ and ζ are that they are on the same side for large values of ζ. These points are where one of the NO_x flow models is much better than the other because the subtraction in (4.2) results in a large value. To illustrate this a lower limit of ζ is implemented. All values in ζ

that are lower then the limit are removed and not included in the calculations. The calculations describes how many times $\widehat{\zeta}$ is on the same side of 0 as ζ and fulfills the condition in (4.3). This can be expressed by an extension of the criterion (4.3) to

$$sign\left(\zeta\right) = sign\left(\widehat{\zeta}\right) \quad \wedge \quad \zeta > lower\ limit. \tag{4.4}$$

The lower limit in (4.4) was then slided from 0 to 1100 and for each new lower limit more points were removed. The result from increasing the lower limit for ζ and the three engines are shown in Figure 4.10.
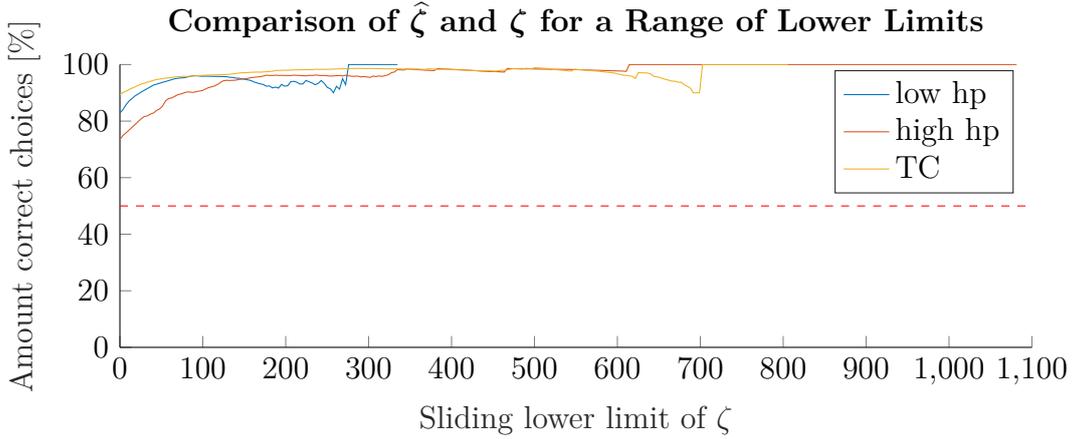
Figure 4.10: Amount of time $\widehat{\zeta}$ and ζ have chosen to listen to the same NO_x flow model for different lower limits. The desirable behaviour here is that the lines converge to 100 %, preferably as quickly as possible. This means that the correct model was chosen when it was most important.

For low limits the choice of model is not crucial. If ζ is small an inaccurate choice will not have a large impact on the performance. For high limits the choice is very important. In these points one model is clearly better than the other and wrong choice here could lead to large decrease in performance.

The reason why it starts low and then increase is that it is hard to predict which model to listen to when ζ is small. $\widehat{\zeta}$ must be very accurate to catch this behavior.

The different engine models do here have different lengths in Figure 4.10. The reason for this is that if the model is good, then the D value will be small because the model will be accurate. It will then not have any values left when the lower limit is approaching high values. The line will then be shorter than for models which have large values in D. It can then be seen in the figure that high hp and TC have larger values, which means that the models are performing better on low hp than high hp and TC.

4.2 Sensor Fusion Performance with Different Estimators

Earlier in the thesis five different estimators have been presented. These are:

1. FE

2. DKF

3. UIO

4. SSUIO

5. ESV

The performance of these observers will in this section first be presented separately and then compared against each other. Only the results for the low hp engine are going to be shown in this section. The corresponding results for the high hp and TC engine are shown in Appendix A.2.7.

4.2.1   Performance of Fraction Estimator

In this section the performance of the fraction estimator described in Section 3.2.1 will be presented. The estimated NO_x flow from this observer, NO_x sensor values, and Horiba measurements for eSCR low hp are shown in Figure 4.11.

Comparison of Horiba, NO$_x$ Sensor and FE for 480 hp eSCR

Figure 4.11: Sequence from the validation dataset of eSCR low hp that shows that FE behaves like the Horiba measurements well. It has the same amplitude and keeps up in the transient parts. The content in the dashed square is shown in the plot in the left corner. Here it can be seen that the transient behaviour is better for the FE than for the sensor.

It can be seen in Figure 4.11 that the estimator is behaving like the Horiba measurements quite well. They have both the same characteristics and about the same amplitude. The left corner of Figure 4.11 shows the content from the dashed square. It can be seen that FE have better transient behaviour then the NO_x sensor. FE responses to both a positive and negative step faster than the NO_x sensor and this behaviour was consistent through out the dataset.

The performance of the FE is evaluated for both MSE and calculation time per sample for all engine models. This evaluation is shown in Table 4.7. It is the fraction between the MSE for the FE and the sensor who are presented and not the pure MSE. The reason for this is that the important performance in this part is the observer MSE in relation to the sensor MSE and not the actual MSE value of the observer. It is wanted to have as low MSE quota as possible but less than one implies that an improvement from the sensor is achieved.

Table 4.7: Metrics for the fraction estimator performance.

	low hp	high hp	TC
MSE (FE/Sensor)	0.31469	0.56405	0.55190
Calculation Time per Sample [ms]	0.23033	0.41810	0.12326

4.2.2 Performance of Delayed Linear Kalman Filter

In this section the performance of the linear Kalman filter with delayed measurements, described in Section 3.2.2, is presented. The delay that were used when calculating the delayed Kalman filter was 15 samples which correspond to 1.5 seconds. As described earlier, this filter must be tuned and the values that have been concluded for a, b and c are shown in Table A.1 in Appendix A.2.2.

The estimated NO_x flow from this observer, NO_x sensor values, and Horiba measurements for eSCR low hp are shown in Figure 4.12.

Comparison of Horiba, NO_x Sensor and DKF for 480 hp eSCR
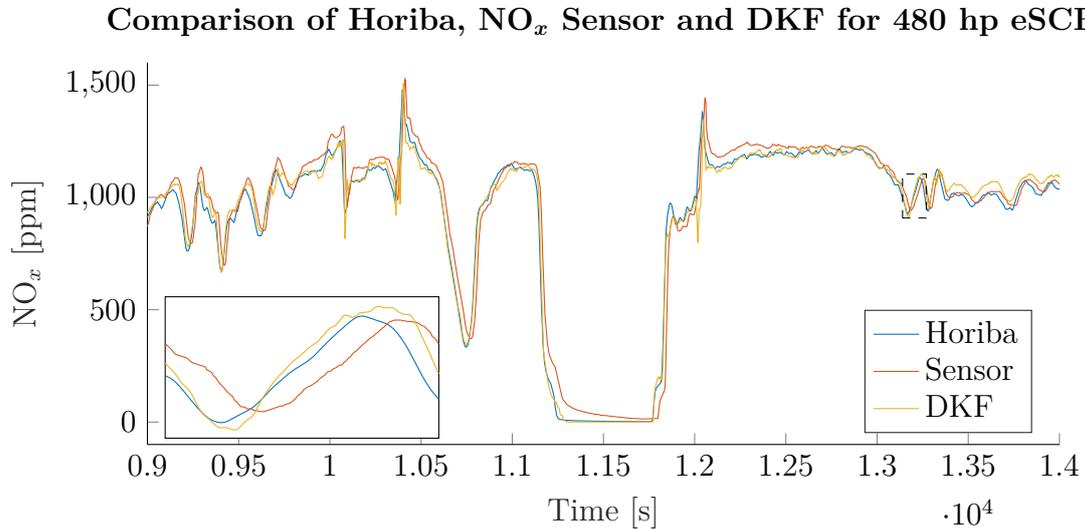
Figure 4.12: Sequence from the validation dataset of eSCR low hp which shows that DKF follows the Horiba measurements well. It have the same amplitude and keeps up in the transient parts. The contents in the dashed square are shown in the plot in the left corner. Here it can be seen that the transient behaviour is better for the DKF than for the sensor.

It can be seen in Figure 4.12 that the DKF estimator is behaving like the Horiba measurement well. They have both the same characteristics and about the same

amplitude. The transient behaviour, shown in the zoomed in part in Figure 4.12, is also better than the sensor where it responded to both a positive and negative step faster then the NO_x sensor and this behaviour was consistent throughout the dataset.

The metrics of the observer performance for the different engine models are shown in Table 4.8.

Table 4.8: Metrics for the delayed linear Kalman filter performance.

	low hp	high hp	TC
MSE (DKF/Sensor)	0.30751	0.64489	0.71552
Calculation Time per Sample [ms]	6.1377	9.2503	4.3487

4.2.3 Performance of the Unknown Input Observer

In this section the performance of the unknown input observer, described in Section 3.2.3, is presented. The tuning of the covariances in this filter that was found to perform well for each engine is shown in Table A.3 in Appendix A.2.4 and were used as

$$Q \sim \mathcal{N}(0, \text{diag}(\sigma_x)) \qquad R \sim \mathcal{N}(0, \text{diag}(\sigma_y)). \qquad (4.5)$$

The estimated NO_x flow from this observer, NO_x sensor values, and Horiba measurements for eSCR low hp are shown in Figure 4.13.

Comparison of Horiba, NO_x Sensor and UIO for 480 hp eSCR
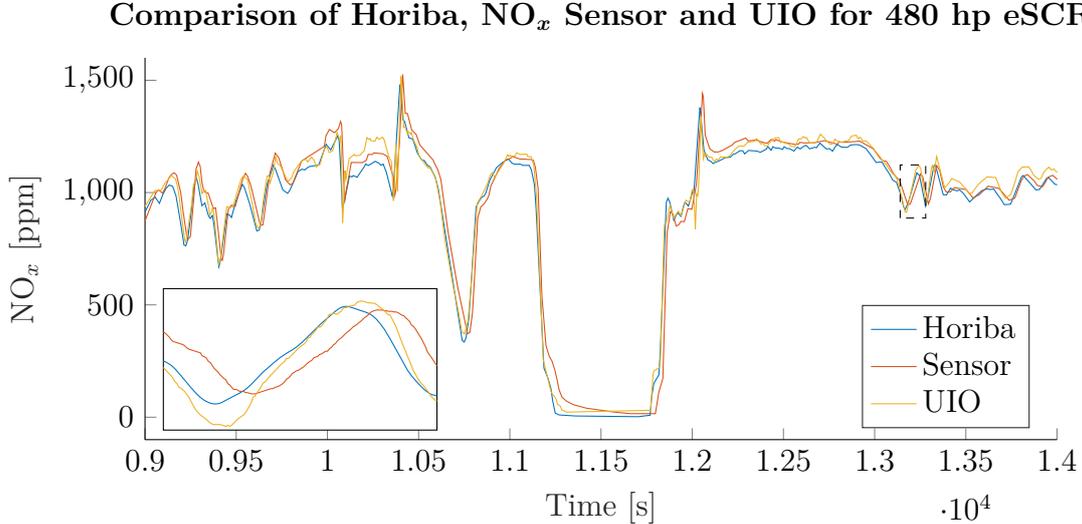
Figure 4.13: Sequence from the validation dataset of eSCR low hp which shows that UIO follows the Horiba measurement well. It have the same amplitude and keeps up in the transient parts. The contents in the dashed square are shown in the plot in the left corner. Here it can be seen that the transient behaviour is better for the UIO than for the sensor.

The methods for evaluating the UIO performance are the calculation time and the fraction between observer MSE and sensor MSE. These metrics are shown in Table 4.9 for all three engines.

Table 4.9: Metrics for the UIO performance.

	low hp	high hp	TC
MSE (UIO/Sensor)	0.34475	0.47945	0.40624
Calculation Time per Sample [ms]	0.31257	0.54464	0.19609

4.2.4 Performance of the Unknown Input Observer with Steady State Kalman Gain

In this section the performance of the unknown input observer with steady state solution from Section 3.2.4 is presented. The values which the Kalman gain converges to are shown in equation (A.12) in Appendix A.2.5.

The reason for that the Kalman gain is close to just ones and zeros for $\bar{K}^{low}$ is because of that the models themselves perform better than the sensor and for this engine the model has chosen to almost exclude the sensor for the estimation. The estimated NO_x flow from this observer, NO_x sensor and Horiba measurements for eSCR low hp are shown in Figure 4.14.
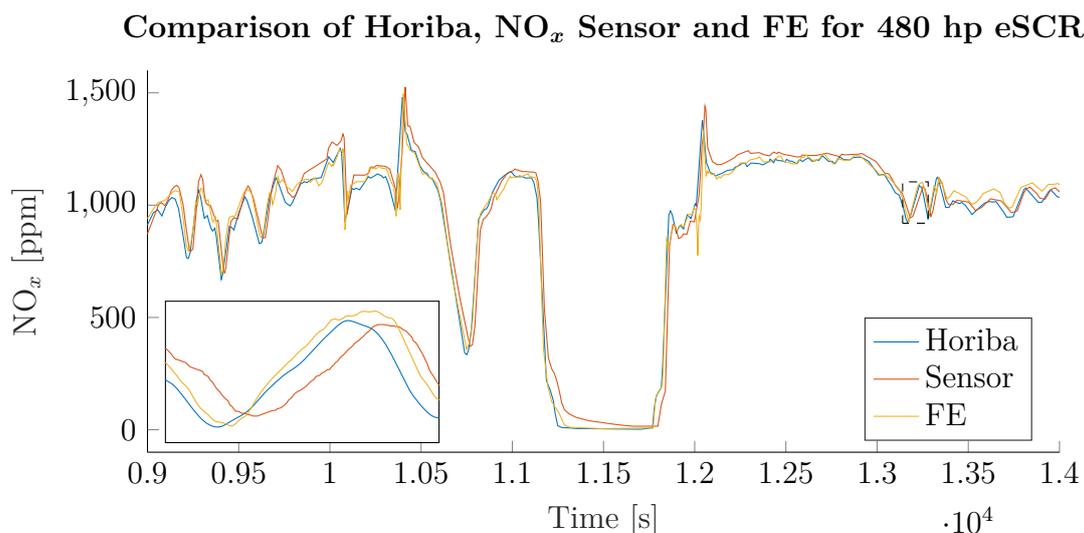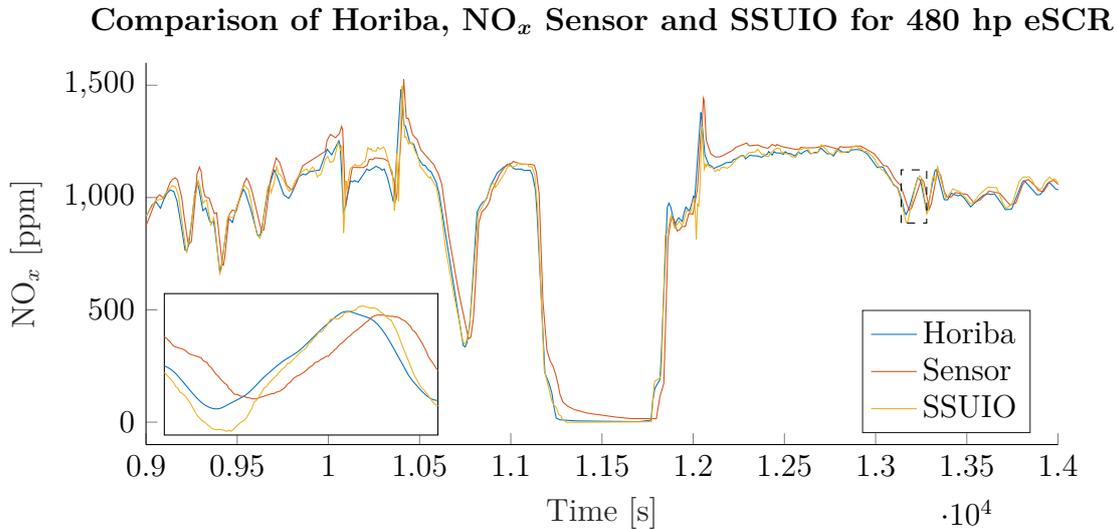

Figure 4.14: Sequence from the validation dataset of eSCR low hp which shows that SSUIO follows the Horiba measurements well. It have the same amplitude and keeps up in the transient parts. The contents in the dashed square are shown in the plot in the left corner. Here it can be seen that the transient behaviour is better for the SSUIO than for the sensor.

As seen in Figure 4.14 it behaves almost exactly as the UIO with iterative calculation of the Kalman gain. To further investigate the performance, calculation time and the fraction between the observer MSE and sensor MSE are analyzed. The results are shown in Table 4.10.

Table 4.10: Metrics for the SSUIO performance.

	low hp	high hp	TC
MSE (SSUIO/Sensor)	0.34459	0.49974	0.41685
Calculation Time per Sample [ms]	0.24107	0.44515	0.13603

4.2.5 Performance of the Extended State Vector Observer

In this section the performance of the Kalman filter with extended state vector from Section 3.2.5 is presented. Here the covariances were set constant in order to get the observer running and further tuning can be done to see if this observer is comparable to the other ones. The estimated NO_x flow from this observer, NO_x sensor and Horiba measurements for eSCR low hp are shown in Figure 4.15.
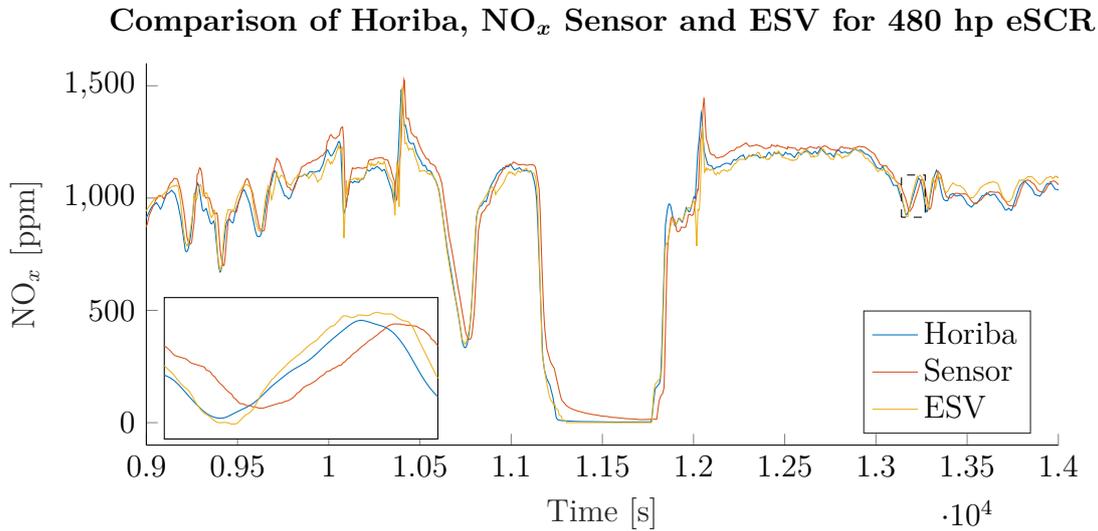
Figure 4.15: Sequence from the validation dataset of eSCR low hp which shows that ESV follows the Horiba measurements well. It have the same amplitude and keeps up in the transient parts. The contents in the dashed square are shown in the plot in the left corner. Here it can be seen that the transient behaviour is better for the ESV than for the sensor.

To further investigate the performance, calculation time and the fraction between the observer MSE and sensor MSE are analyzed. The results are shown in Table 4.11.

Table 4.11: Metrics for the ESV performance.

	low hp	high hp	TC
MSE (SSUIO/Sensor)	0.29783	0.87144	0.77563
Calculation Time per Sample [ms]	0.29432	0.47777	0.18426

4.2.6 Comparison of the Developed Estimator Performance on each Engine Model

In this section the results from Section 4.2.1-4.2.2 will be compared for each engine model. The low hp eSCR engine is shown in Table 4.12, the high hp eSCR in Table 4.13 and the TC engine in Table 4.14.

Table 4.12: Comparison of metrics for all observers applied on the low hp eSCR engine.

	MSE Observer/Sensor	Calculation Time Per Sample [ms]
FE	0.31469	0.23033
DKF	0.28460	6.1377
UIO	0.34475	0.31257
SSUIO	0.34459	0.24107
ESV	0.29783	0.29432

Table 4.13: Comparison of metrics for all observers applied on the high hp eSCR engine.

	MSE Observer/Sensor	Calculation Time Per Sample [ms]
FE	0.56405	0.41810
DKF	0.64489	9.2503
UIO	0.47945	0.54464
SSUIO	0.49974	0.44515
ESV	0.87144	0.47777

Table 4.14: Comparison of metrics for all observers applied on the TC engine.

	MSE Observer/Sensor	Calculation Time Per Sample [ms]
FE	0.55190	0.12326
DKF	0.42371	4.3487
UIO	0.40624	0.19609
SSUIO	0.41685	0.13603
ESV	0.77563	0.18426

From Table 4.12-4.14 several patterns can be seen. The DKF clearly have the largest computational time and the SSUIO have slightly higher MSE than the iterative UIO. The calculation time decreased from UIO to the SSUIO. For low hp was it decreased with 22%, 16% for high hp and with 26% for TC.

5

Discussion

In this section the methods and results presented in this thesis will be discussed. First different aspects of the developed models will be presented, followed by a discussion about the developed observers and at last future work will be discussed.

5.1 Discussion of the Developed Models

In this section different aspects of the developed models will be discussed. This includes inconsistent Horiba measurement delay and the model performance on the TC engine.

5.1.1 Side Effect due to Inconsistent Horiba Measurement Delay

The delay of the Horiba measurements are inconsistent and it is not possible to perfectly shift it so it aligns with the other signals. This will give raise to different artifacts in $\widehat{D}$ that are unwanted. In a positive step the Horiba measurements can for example be behind the NO_x flow models and this will give raise to a spike in D. This spike is not present in the true system but when training $\widehat{D}$ these spikes will be present. This will add uncertainty to the models and decrease their performance because they are trained on non representative data.

5.1.2 Quality of Model performance on Turbo Compound Engine

One issue in this thesis is that the performance of the NO_x flow models on the TC engine is not good. The semiphysical model is not optimized for TC engines and is constantly underestimating the NO_x flow values. The performance of the radial basis model is not good either but not as bad as the semiphysical model. Because the semiphysical model is not performing well it might be an idea to exclude it from the observers as it is increasing the computational effort without bringing that much value. This is a trade off between speed and accuracy and the choice will differ depending on how good the semiphysical model performs.

5.2 Discussion of the Developed Observers

In this section different aspects of the developed observers will be discussed. This includes aspects of the delayed Kalman filter and exclusion of observers.

5.2.1 Steady State Solution to the Delayed Kalman Filter

If the methods that are developed in this thesis are going to be implemented on a real ECU the computational power is an important factor. This implies that the steady state solution of a Kalman filter is wanted, which means that the DKF in the current execution is not possible to use in the ECU available today. In the DKF R and R^* are depending on dynamic functions. This implies that the Kalman gain, in the most cases, will not converge and no steady state solution is possible.

The steady state Kalman gain for the low hp engine, shown in Table **??** Appendix A.2.5, seems to exclude the sensor measurements when estimating the NO_x flow. This is due to that the models are performing very good on the low hp engine. This is not true for the high hp engine or the TC and in that case they are more evenly weighted.

For all engines UIO is more accurate in terms of MSE than the SSUIO. This might be a effect of that the Kalman gain is not optimal in the first samples of the SSUIO. It will therefore have lower accuracy initially and increase the MSE of the SSUIO. If, on the other hand, the iterative solution is used, the Kalman filter gain is optimal from the start and this will improve the performance. Depending on how off the Kalman filter gain is from optimal in the beginning the MSE will be affected differently.

5.2.2 Drawbacks of the Delayed Kalman Filter

Using the DKF, many points are calculated several times due to the update of states when a new measurement arrives. This will increase the calculation time per sample and might not be feasible in systems where the calculation time is crucial, computing power limited or when the sensor delay is large.

Another disadvantage of this method is that values must be stored to some extent. To be able to do the recalculation of $\hat{x}_{s+1:k}$, $y_{s+1:k}$ must be known. It must therefore be stored in memory which in some cases are not possible. This disadvantage can be limited by setting a maximum delay that describes the maximum amount of values from y that are saved. If the delay is from a time before the limit it could then be discarded.

It is possible to use the same method for avoiding dynamic covariance matrices in the DKF as used in the UIO. This method would fuse the model values before using a combined measurement in the filter using equation (3.10). This will not solve the problem with large computational time but a steady state solution will be possible. The reason for that the computational time still will be a problem, even if it will be slightly reduced with a steady state solution, is that the majority of the calculation time comes from recalculating old estimates. The recalculations will still be needed even if the steady state solution is used.

5.2.3 Exclusion of Observers

It should be emphasized that no observers in this thesis should be excluded from being implemented in the future. All observers are using $\widehat{D}$ that are based on training data. If $\widehat{D}$ would be improved this can change which observer that is performing the best in terms of accuracy. As seen in Section 3.1.1 the training data is not optimal. This data might also affect the final accuracy of the observers.

5.3 Future Work

In this section potential future work will be presented. This includes improvements of the developed models, dynamic time delay and merging of models.

5.3.1 Improvements of the Developed Models

There are some ways that $\widehat{D}_{SP}$ and $\widehat{D}_{RB}$ could potentially be improved. One thing that could be tested is if the training and validation data would be split in an another way. In the end of the used datasets are the engine speed often, more or less, constant which can be seen in Figure 3.6. This behaviour will end up as a majority of the data that are used for validation. This might not be a problem due to that it is the transient behaviour of the NO_x flow that is important. One can see in Figure 3.10 that the NO_x flow is still transient even though the engine speed is somewhat stationary.

Another improvement that could have been done is to investigate more inputs to $\widehat{D}_{SP}$ and $\widehat{D}_{RB}$. There might be more signals that affect the models in a positive way that are not included in the current models. A large portion of time have been spent on finding the current parameters in the models and it is always possible to spend more time and find more variables. This is encouraged to do in the future due to the time restrictions of this thesis.

5.3.2 Dynamic Time Delay in Filters

An interesting next step for the DKF and UIO is the development of a dynamic time delay. Now the delay was set to a fixed value of 1.5 seconds when evaluating the observers. The delay of the sensor is not constant but it was put to a typical delay of the sensor. The delay of the sensor can potentially be estimated using the same methods that are used for finding $\widehat{D}$ in this thesis. This means, finding signals that have correlation with the delay and then apply linear regression to find a method that can predict the sensor delay. A more accurate delay can improve the performance in the DKF and the UIO where they both assume that the delay given is the true delay.

5.3.3 Possibility of One Model for All Engine Models

Initially in the project a single $\widehat{D}$ was developed for all three engine models. It turned out that that engine models were so different from each other that the performance

of the model was bad. It was inaccurate for some engines and accurate for others. More investigation can be put into trying to develop a single model for all engine types.

5.3.4 Improvements of the Developed Observers

There is always a possibility to improve the observers by fine tune them even more. The tuning process was done on the current observers until the performance was considered to be good enough, or until the performance was unable to improve by tuning. To further improve the observers are therefore the reader, or implementer, encouraged to tune the observers more.

6
Conclusion

In this thesis five different observers have been developed for estimating the EO NO_x flow. These are

- FE

- DKF

- UIO

- SSUIO

- ESV

The main goal of the observers was to have a better transient behaviour than the sensor used today for NO_x flow estimation. All five designed observers have faster, that will say better, transient behaviour than the sensor. The observers differ in both accuracy and computational effort which are the two metrics used in this thesis. If the observers would be ranked by their performed MSE, the list would be:

1. UIO - 1.23044 total MSE

2. SSUIO - 1.26118 total MSE

3. DKF - 1.3532 total MSE

4. FE - 1.43064 total MSE

5. ESV - 1.9449 total MSE

In this list the MSE values for all engine model are added together for each observer. The corresponding list for the calculation time is:

1. FE - 0.77169 ms total calculation time

2. SSUIO - 0.82225 ms total calculation time

3. ESV - 0.95635 ms total calculation time

4. UIO - 1.0533 ms total calculation time

5. DKF - 19.7367 ms total calculation time

The observer which had the best accuracy over all three engines was the UIO. However, as the calculation time has an important role the SSUIO was made the best candidate as it have less computational effort and have still a good performance in comparison to the other observers. These decisions led to that the SSUIO was chosen to be implemented in TargetLink and the observer which the authors suggest Volvo should use to improve the EO NO_x estimation.

Bibliography

[1] T. Jacobs, D. N. Assanis, and Z. Filipi, "The impact of exhaust gas recirculation on performance and emissions of a heavy-duty diesel engine," SAE Technical paper, Tech. Rep., 2003.

[2] C. P. Cho, Y. D. Pyo, J. Y. Jang, G. C. Kim, and Y. J. Shin, "No x reduction and n 2 o emissions in a diesel engine exhaust using fe-zeolite and vanadium based scr catalysts," *Applied Thermal Engineering*, vol. 110, pp. 18–24, 2017.

[3] M. Kampa and E. Castanas, "Human health effects of air pollution," *Environmental pollution*, vol. 151, no. 2, pp. 362–367, 2008.

[4] R. Westerholm and K.-E. Egebäck, "Exhaust emissions from light-and heavy-duty vehicles: chemical composition, impact of exhaust after treatment, and fuel parameters." *Environmental health perspectives*, vol. 102, no. Suppl 4, p. 13, 1994.

[5] W. Knecht, "Diesel engine development in view of reduced emission standards," *Energy*, vol. 33, no. 2, pp. 264–271, 2008.

[6] S. Lion, C. N. Michos, I. Vlaskos, C. Rouaud, and R. Taccani, "A review of waste heat recovery and organic rankine cycles (orc) in on-off highway vehicle heavy duty diesel engine applications," *Renewable and Sustainable Energy Reviews*, vol. 79, pp. 691–708, 2017.

[7] (2009, jun) Regulation (ec) no 595/2009 of the european parliament and of the council. [Online]. Available: http://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:32009R0595

[8] (2005, sep) Directive 2005/55/ec of the european parliament and of the council. [Online]. Available: http://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:32005L0055

[9] (2001, apr) Commission directive 2001/27/ec. [Online]. Available: http://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:32001L0027

[10] (1999, dec) Directive 1999/96/ec of the european parliament and of the council. [Online]. Available: http://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:31999L0096

[11] (1987, dec) Council directive of 3 december 1987 on the approximation of the laws of the member states relating to the measures to be taken against the emission of gaseous pollutants from diesel engines for use in vehicles (88/77/eec). [Online]. Available: http://eur-lex.europa.eu/legal-content/en/TXT/?uri=CELEX:31988L0077

[12] M. Koebel, M. Elsener, and T. Marti, "Nox-reduction in diesel exhaust gas with urea and selective catalytic reduction," *Combustion science and technology*, vol. 121, no. 1-6, pp. 85–102, 1996.

[13] Y.-Y. Wang, Y. Sun, C.-F. Chang, and Y. Hu, "Model-based fault detection and fault-tolerant control of scr urea injection systems," *IEEE Transactions on Vehicular Technology*, vol. 65, no. 6, pp. 4645–4654, 2016.

[14] S. Yao, M. Shost, J.-H. Yoo, D. Cabush, D. Racine, R. Cloudt, F. Willems, *et al.*, "Ammonia sensor for closed-loop scr control," *SAE international journal of passenger cars-electronic and electrical systems*, vol. 1, no. 2008-01-0919, pp. 323–333, 2008.

[15] R. Suarez-Bertoa, A. Zardini, and C. Astorga, "Ammonia exhaust emissions from spark ignition vehicles over the new european driving cycle," *Atmospheric Environment*, vol. 97, pp. 43 – 53, 2014. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S1352231014005809

[16] J. Gieshoff, A. Schäfer-Sindlinger, P. Spurk, J. Van Den Tillaart, and G. Garr, "Improved scr systems for heavy duty applications," SAE Technical Paper, Tech. Rep., 2000.

[17] B. Henriksson, "Calculation report: A new type of model for nox emissions and exhaust temperature has been developed. it is intended for the vtr with the purpose of increasing the robustness., Engineering Report" oct 2015, internal Volvo report. Not public.

A

Appendix

A.1 Model Details

A.1.1 Inputs to Models

The models have different inputs for $\widehat{D}_{SP}$ and $\widehat{D}_{RB}$ and they are also different for the different engines.

For $\widehat{D}_{SP}^{high}$ and $\widehat{D}_{RB}^{high}$ the equations are

$$
\begin{bmatrix}
\tau_{Eng} \\
1/\tau_{Eng} \\
\omega_{Eng}/\widetilde{\omega}_{Eng} \\
1/\omega_{Eng} \\
(\widetilde{T}_B - T_B) \\
1/(\widetilde{T}_B - T_B) \\
P_R/\widetilde{P}_R \\
\widetilde{P}_R/P_R \\
1/C_{corr}^2 \\
1/C_{corr} \\
\alpha EGR \\
(200 - AFR) \\
AFR
\end{bmatrix}
\tag{A.1}
$$

and

$$
\begin{bmatrix}
\tau_{Eng} \\
\omega_{Eng} \\
1/\omega_{Eng} \\
T_B \\
1/T_B \\
P_R \\
1/P_R \\
1/C_{corr} \\
C_{corr} \\
\alpha EGR \\
(200 - AFR) \\
C_{cyl} \\
C_{\omega\tau}
\end{bmatrix}
.
\tag{A.2}
$$

For $\widehat{D}_{SP}^{TC}$ and $\widehat{D}_{RB}^{TC}$ the equations are

$$\begin{bmatrix} \tau_{Eng} \\ \omega_{Eng}/\widetilde{\omega}_{Eng} \\ (\widetilde{T}_B - T_B) \\ 1/(\widetilde{T}_B - T_B) \\ P_R/\widetilde{P}_R \\ \widetilde{P}_R/P_R \\ 1/C_{corr} \\ \alpha EGR \\ (200 - AFR) \end{bmatrix} \tag{A.3}$$

and

$$\begin{bmatrix} \tau_{Eng} \\ T_B \\ 1/T_B \\ P_R \\ 1/P_R \\ 1/C_{corr} \\ \alpha EGR \\ (200 - AFR) \\ C_{\omega\tau} \end{bmatrix}. \tag{A.4}$$

A.1.2   Equations from Linear Regression

The final equation for $\widehat{D}_{SP}^{high}$ and $\widehat{D}_{RB}^{high}$ was obtained as

$$\begin{aligned}
&1 + x_1 + x_3 + x_1 x_3 + x_1 x_4 + x_2 x_4 + x_1 + x_5 + x_1 x_5 + x_4 x_5 + x_1 x_6 \\
&+ x_3 x_6 + x_1 + x_7 + x_1 x_7 + x_2 + x_7 + x_2 x_7 + x_3 + x_7 + x_3 x_7 + x_4 x_7 \\
&+ x_5 + x_7 + x_5 x_7 + x_5 + x_8 + x_5 x_8 + x_6 x_8 + x_1 x_9 + x_4 x_9 + x_5 x_9 \\
&+ x_6 x_9 + x_7 x_9 + x_1 + x_{10} + x_1 x_{10} + x_3 + x_{10} + x_3 x_{10} + x_5 \\
&+ x_{10} + x_5 x_{10} + x_6 x_{10} + x_7 + x_{10} + x_7 x_{10} + x_8 + x_{10} + x_8 x_{10} \\
&+ x_1 + x_{11} + x_1 x_{11} + x_3 + x_{11} + x_3 x_{11} + x_4 x_{11} + x_5 + x_{11} \\
&+ x_5 x_{11} + x_6 x_{11} + x_7 + x_{11} + x_7 x_{11} + x_8 + x_{11} + x_8 x_{11} \\
&+ x_9 x_{11} + x_5 + x_{12} + x_5 x_{12} + x_{10} + x_{12} + x_{10} x_{12} + x_{11} \\
&+ x_{12} + x_{11} x_{12} + x_1 x_{13} + x_3 x_{13} + x_4 x_{13} + x_7 x_{13} \\
&+ x_8 x_{13} + x_9 x_{13} + x_3^2 + x_5^2 + x_7^2 + x_8^2 + x_{11}^2 + x_{13}^2
\end{aligned} \tag{A.5}$$

II

and

$$
\begin{aligned}
1 &+ x_1 + x_2 + x_1 x_2 + x_1 + x_3 + x_1 x_3 + x_1 + x_4 + x_1 x_4 + x_2 + x_4 + x_2 x_4 \\
&+ x_3 + x_4 + x_3 x_4 + x_1 + x_5 + x_1 x_5 + x_2 + x_5 + x_2 x_5 + x_1 + x_6 + x_1 x_6 + x_3 \\
&+ x_6 + x_3 x_6 + x_5 + x_6 + x_5 x_6 + x_1 + x_7 + x_1 x_7 + x_4 + x_7 + x_4 x_7 + x_1 x_8 \\
&+ x_2 x_8 + x_4 x_8 + x_6 x_8 + x_1 + x_9 + x_1 x_9 + x_2 + x_9 + x_2 x_9 + x_4 + x_9 \\
&+ x_4 x_9 + x_6 + x_9 + x_6 x_9 + x_1 + x_{10} + x_1 x_{10} + x_2 + x_{10} + x_2 x_{10} + x_3 \\
&+ x_{10} + x_3 x_{10} + x_6 + x_{10} + x_6 x_{10} + x_7 + x_{10} + x_7 x_{10} + x_8 x_{10} \\
&+ x_9 + x_{10} + x_9 x_{10} + x_1 x_{11} + x_3 x_{11} + x_6 x_{11} + x_7 x_{11} + x_8 x_{11} \\
&+ x_{10} x_{11} + x_1 x_{12} + x_2 x_{12} + x_3 x_{12} + x_6 x_{12} + x_7 x_{12} \\
&+ x_8 x_{12} + x_9 x_{12} + x_{10} x_{12} + x_{11} x_{12} + x_1 x_{13} + x_4 x_{13} + x_6 x_{13} \\
&+ x_8 x_{13} + x_9 x_{13} + x_{10} x_{13} + x_{11} x_{13} + x_{12} x_{13} + x_1^2 + x_2^2 + x_4^2 \\
&+ x_6^2 + x_9^2 + x_{10}^2
\end{aligned}
\tag{A.6}
$$

where $x_1 - x_{13}$ are the corresponding elements from top to bottom of A.1 for $\widehat{D}_{SP}^{high}$ and A.2 for $\widehat{D}_{RB}^{high}$.

The final equation for $\widehat{D}_{SP}^{low}$ and $\widehat{D}_{RB}^{low}$ was obtained as

$$
\begin{aligned}
1 &+ x_1 + x_2 + x_1 x_2 + x_2 + x_4 + x_2 x_4 + x_2 + x_5 + x_2 x_5 + x_3 x_5 + x_1 + x_6 \\
&+ x_1 x_6 + x_4 + x_6 + x_4 x_6 + x_1 + x_7 + x_1 x_7 + x_2 + x_7 + x_2 x_7 + x_3 x_7 \\
&+ x_4 + x_7 + x_4 x_7 + x_5 + x_7 + x_5 x_7 + x_6 + x_7 + x_6 x_7 + x_1 + x_8 + x_1 x_8 \\
&+ x_2 + x_8 + x_2 x_8 + x_3 x_8 + x_4 + x_8 + x_4 x_8 + x_5 + x_8 + x_5 x_8 + x_7 + x_8 \\
&+ x_7 x_8 + x_1 x_9 + x_2 x_9 + x_3 x_9 + x_4 x_9 + x_5 x_9 + x_8 x_9 + x_1^2 + x_2^2 + x_3^2 \\
&+ x_4^2 + x_5^2 + x_6^2 + x_7^2 + x_8^2
\end{aligned}
\tag{A.7}
$$

and

$$
\begin{aligned}
1 &+ x_1 x_2 + x_1 + x_3 + x_1 x_3 + x_1 + x_4 + x_1 x_4 + x_2 x_4 + x_2 x_5 + x_2 x_6 \\
&+ x_3 + x_6 + x_3 x_6 + x_5 x_7 + x_6 x_7 + x_1 + x_8 + x_1 x_8 + x_2 x_8 + x_5 + x_8 \\
&+ x_5 x_8 + x_7 x_8 + x_1 + x_9 + x_1 x_9 + x_4 + x_9 + x_4 x_9 + x_6 + x_9 + x_6 x_9 + x_8 \\
&+ x_9 + x_8 x_9 + x_6^2 + x_8^2 + x_9^2
\end{aligned}
\tag{A.8}
$$

where $x_1 - x_9$ are the corresponding elements from top to bottom of **??** for $\widehat{D}_{SP}^{low}$ and **??** for $\widehat{D}_{RB}^{low}$.

The final equation for $\widehat{D}_{SP}^{TC}$ and $\widehat{D}_{RB}^{TC}$ was obtained as

$$
1 + x_1 + x_6 + x_1 x_2 + x_1 x_7 + x_1 x_8 + x_1 x_9 + x_8 x_9 + x_1^2 + x_6^2 + x_7^2 + x_9^2
\tag{A.9}
$$

and

$$
\begin{aligned}
1 &+ x_1 + x_5 + x_1 x_5 + x_1 x_6 + x_4 x_6 + x_4 x_7 + x_5 + x_7 + x_5 x_7 \\
&+ x_6 x_7 + x_1 + x_8 + x_1 x_8 + x_3 x_8 + x_6 x_8 + x_4 x_9 + x_6 x_9 + x_8 + x_9 \\
&+ x_8 x_9 + x_1^2 + x_7^2 + x_8^2 + x_9^2
\end{aligned}
\tag{A.10}
$$

where $x_1 - x_9$ are the corresponding elements from top to bottom of A.3 for $\widehat{D}_{SP}^{TC}$ and A.4 for $\widehat{D}_{RB}^{TC}$.

A.2 Observer Details

A.2.1 Slack Variables for FE

The used slack values for the FE are

$$
\begin{aligned}
\gamma_{low} &= 0.84 \\
\gamma_{high} &= 0.42 \\
\gamma_{TC} &= 0.44.
\end{aligned}
\tag{A.11}
$$

A.2.2 Tuning Parameters of DKF

The parameters a, b and c used for tuning the DKF are shown in Table A.1.

Table A.1: The tuning variables a, b and c of DKF for all engine models.

	a	b	c
low hp	0.0001	0.0001	150
high hp	0.01	0.005	20
TC	0.01	0.0001	1

A.2.3 Correction of Fraction Estimator for UIO

The correction factors used for the UIO are shown in Table A.2.

Table A.2: Correction values to FE for UIO

	ξ	ψ
low hp	1.06	0.94
high hp	0.96	1.04
TC	0.76	1.24

A.2.4 Noise Parameters for UIO

In the UIO noise parameters are necessary and the ones used in this thesis are shown in Table A.3.

Table A.3: The covariances used in the UIO.

	σ_x	σ_y
low hp	$\begin{bmatrix} 89 & 0.00003 \end{bmatrix}$	$\begin{bmatrix} 4000 & 0.1 \end{bmatrix}$
high hp	$\begin{bmatrix} 12 & 0.002 \end{bmatrix}$	$\begin{bmatrix} 0.06 & 9 \end{bmatrix}$
TC	$\begin{bmatrix} 10 & 0.0001 \end{bmatrix}$	$\begin{bmatrix} 0.04 & 8 \end{bmatrix}$

A.2.5 Steady State Kalman Gain for the UIO

The converged Kalman filter gains used in this thesis are

$$
\bar{K}^{low} = \begin{bmatrix} 0.0002 & 0.9989 \\ 0.0002 & 0.0001 \\ -0.0002 & 0.0000 \end{bmatrix}
$$

$$
\bar{K}^{high} = \begin{bmatrix} 1.4703 & 0.6382 \\ 0.4639 & 0.0089 \\ -0.1323 & 0.0034 \end{bmatrix} \tag{A.12}
$$

$$
\bar{K}^{TC} = \begin{bmatrix} 1.7027 & 0.6233 \\ 0.5012 & 0.0083 \\ -0.0369 & 0.0010 \end{bmatrix}.
$$

A.2.6 Tuning Parameters for the ESV

The used tuning parameters for the ESV is shown in Table A.4 and A.5.

	start	stop
low hp	0.1	10
high hp	0.1	10
TC	0.1	10

Table A.4: Tuning variables for the measurement noise covariance, Q, used in the ESV for each engine. The values are distributed evenly on the diagonal with values between the start and stop value.

	a	b
low hp	1	1
high hp	0.0001	20
TC	0.0001	10

Table A.5: Tuning variables for the measurement noise covariance, R, used in the ESV for each engine.

A.2.7 Observer Result of High hp and TC Engine

In the result section only the low hp engine is shown. This section will therefore show the corresponding results for the high hp and the TC engine.

A.2.7.1 High hp Engine Observers

The performance of all developed observers for the high hp engine are shown in Figure A.1-A.4.

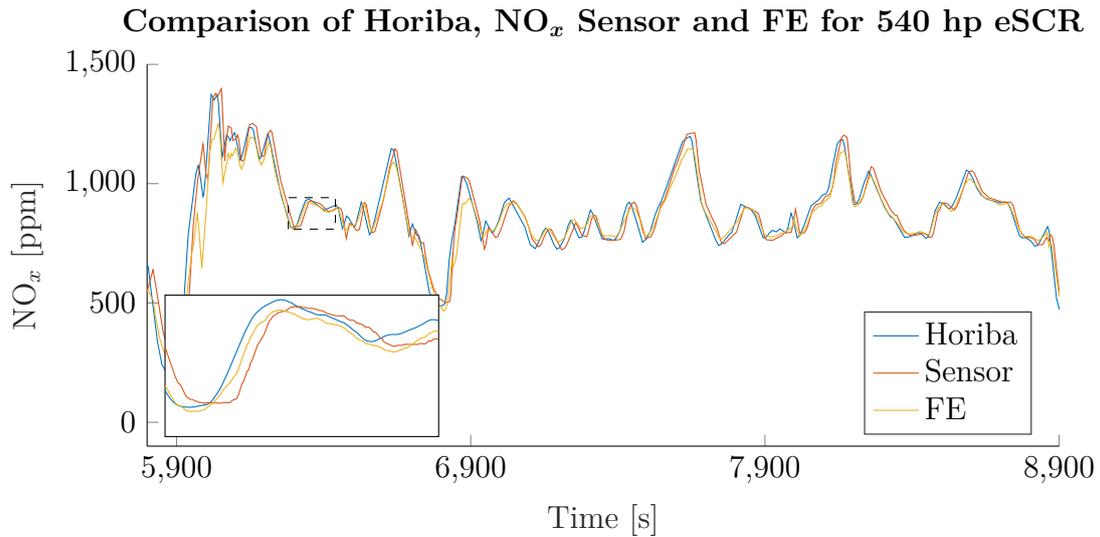**Comparison of Horiba, NO$_x$ Sensor and FE for 540 hp eSCR**

Figure A.1: Sequence from the validation dataset of the high hp engine that shows that FE follows the Horiba measurements well. It have the same amplitude and keeps up in the transient parts. The contents in the dashed square are shown in the plot in the left corner. Here it can be seen that the transient behaviour is better for the FE than for the sensor.
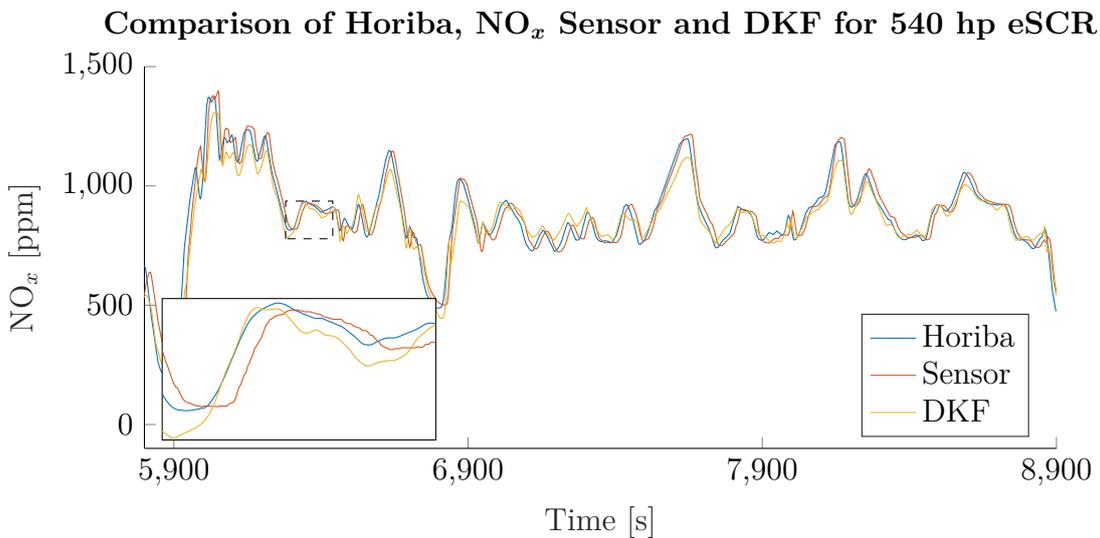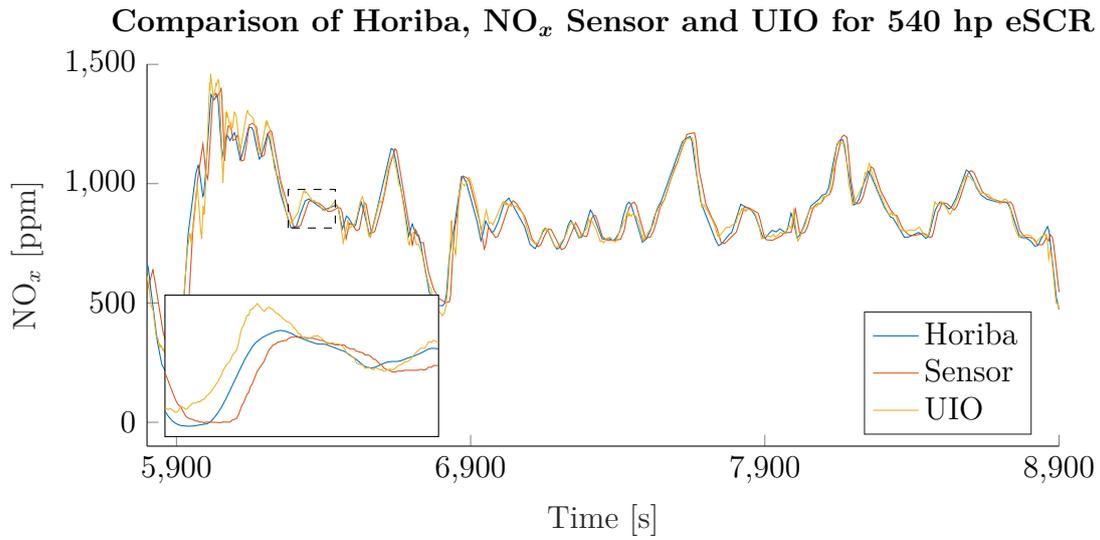
Comparison of Horiba, NO$_x$ Sensor and DKF for 540 hp eSCR

Figure A.2: Sequence from the validation dataset of the high hp engine that shows that DKF follows the Horiba measurements well. It have the same amplitude and keeps up in the transient parts. The contents in the dashed square are shown in the plot in the left corner. Here it can be seen that the transient behaviour is better for the DKF than for the sensor.
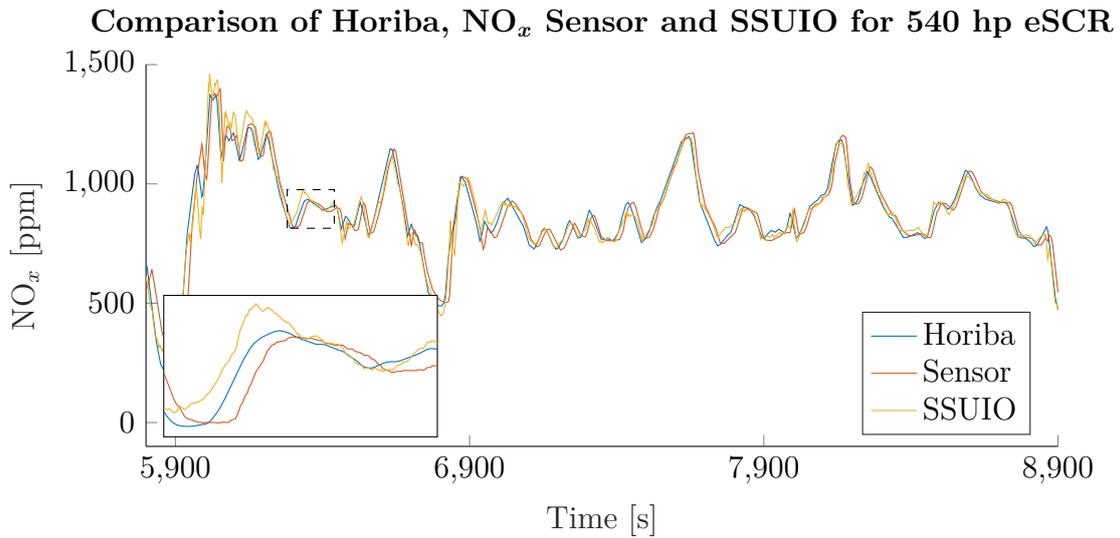
Figure A.3: Sequence from the validation dataset of the high hp engine that shows that UIO follows the Horiba measurements well. It have the same amplitude and keeps up in the transient parts. The contents in the dashed square are shown in the plot in the left corner. Here it can be seen that the transient behaviour is better for the UIO than for the sensor.

Figure A.4: Sequence from the validation dataset of the high hp engine that shows that SSUIO follows the Horiba measurements well. It have the same amplitude and keeps up in the transient parts. The contents in the dashed square are shown in the plot in the left corner. Here it can be seen that the transient behaviour is better for the SSUIO than for the sensor.
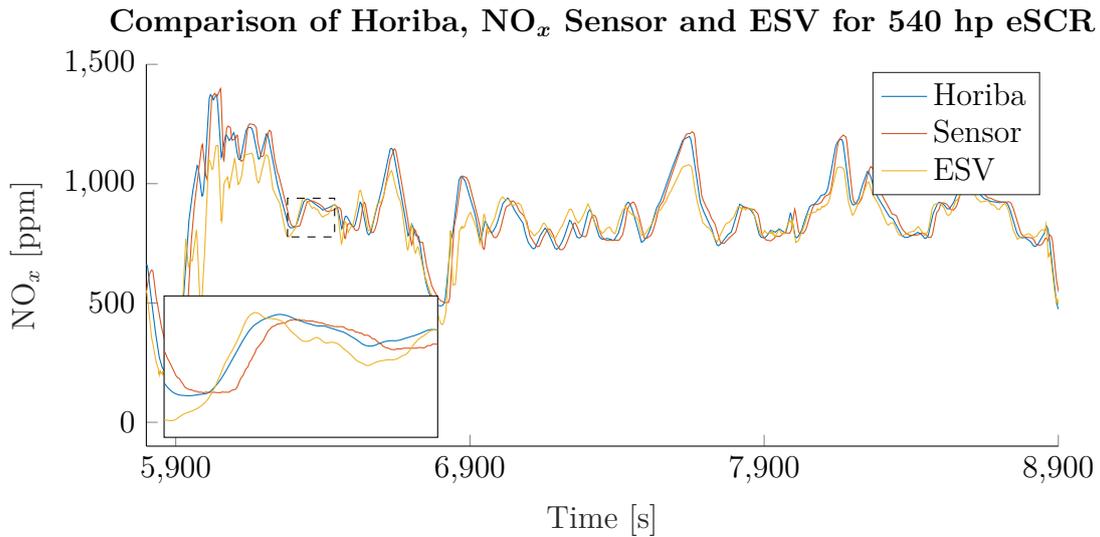
Figure A.5: Sequence from the validation dataset of the high hp engine that shows that ESV follows the Horiba measurements well. It have the same amplitude and keeps up in the transient parts. The contents in the dashed square are shown in the plot in the left corner. Here it can be seen that the transient behaviour is better for the ESV than for the sensor.

A.2.7.2 TC Engine Observers

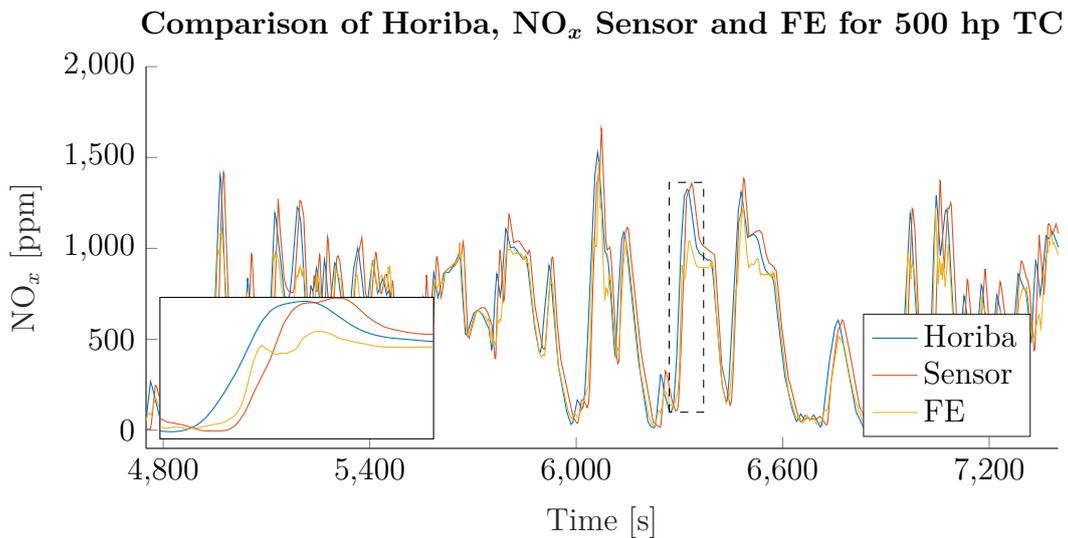The performance of all developed observers for the TC engine are shown in Figure A.6-A.9.

Figure A.6: Sequence from the validation dataset of the TC engine that shows that FE follows the Horiba measurements well. It have the same amplitude and keeps up in the transient parts. The contents in the dashed square are shown in the plot in the left corner. Here it can be seen that the transient behaviour is better for the FE than for the sensor.
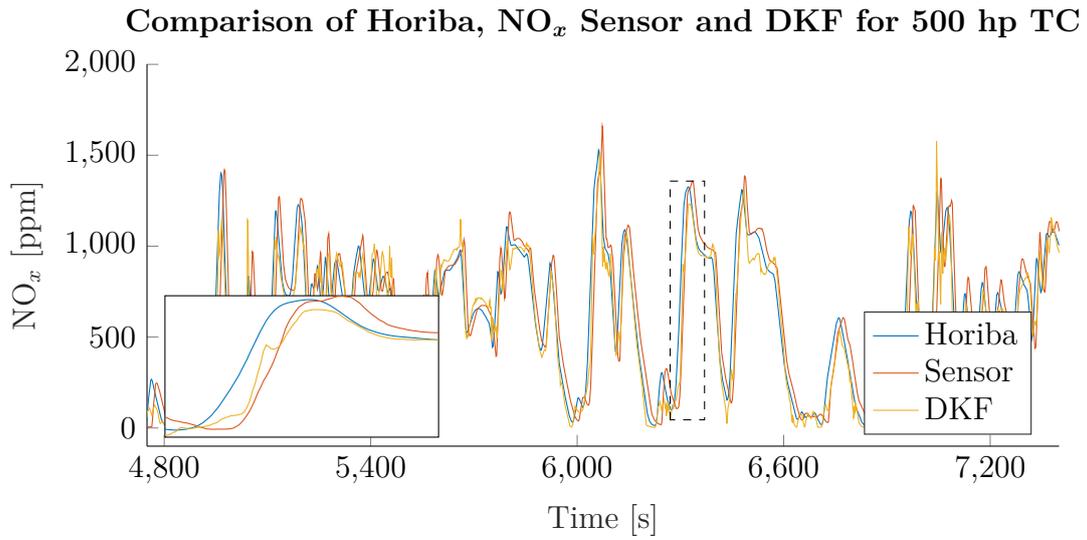
Figure A.7: Sequence from the validation dataset of the TC engine that shows that DKF follows the Horiba measurements well. It have the same amplitude and keeps up in the transient parts. The contents in the dashed square are shown in the plot in the left corner. Here it can be seen that the transient behaviour is better for the DKF than for the sensor.
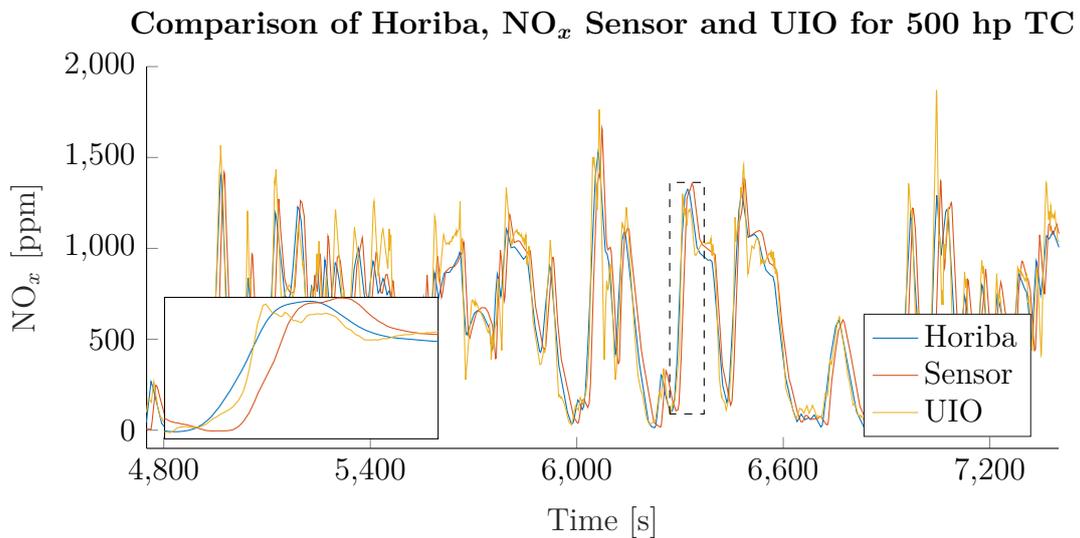

Figure A.8: Sequence from the validation dataset of the TC engine that shows that UIO follows the Horiba measurements well. It have the same amplitude and keeps up in the transient parts. The contents in the dashed square are shown in the plot in the left corner. Here it can be seen that the transient behaviour is better for the UIO than for the sensor.
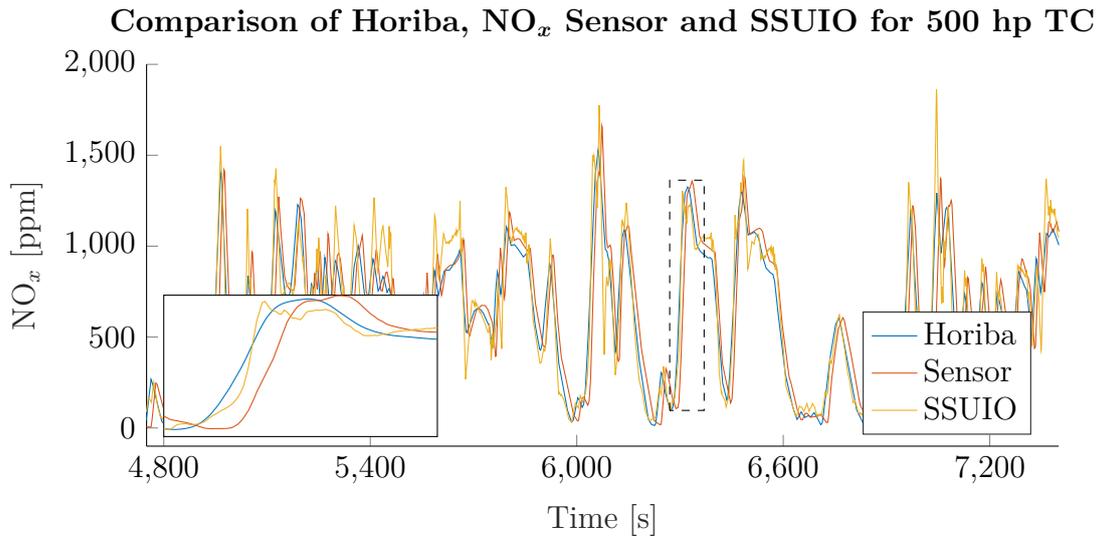
Comparison of Horiba, NO$_x$ Sensor and SSUIO for 500 hp TC

Figure A.9: Sequence from the validation dataset of the TC engine that shows that SSUIO follows the Horiba measurements well. It have the same amplitude and keeps up in the transient parts. The contents in the dashed square are shown in the plot in the left corner. Here it can be seen that the transient behaviour is better for the SSUIO than for the sensor.
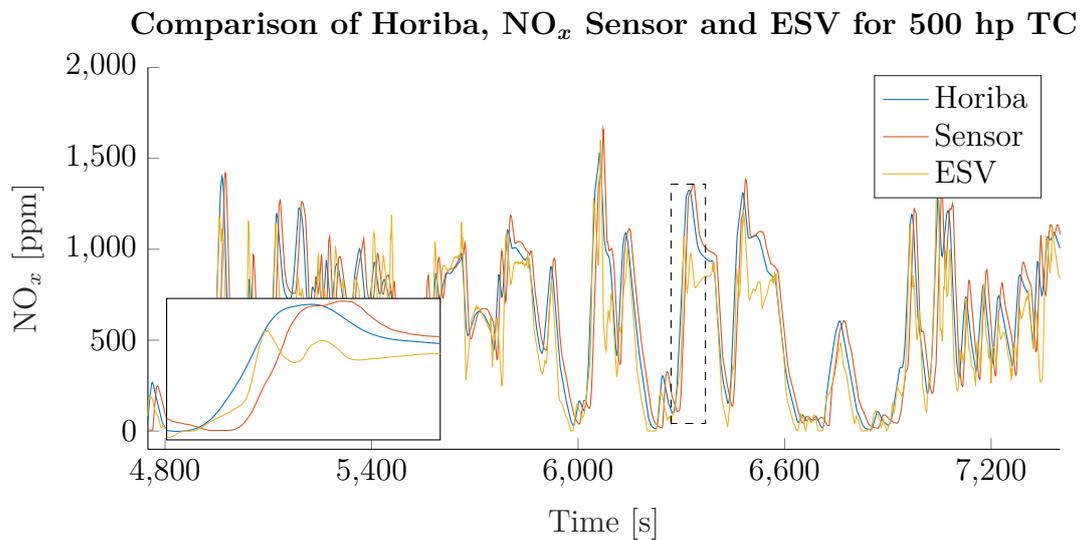
Comparison of Horiba, NO$_x$ Sensor and ESV for 500 hp TC

Figure A.10: Sequence from the validation dataset of the TC engine that shows that ESV follows the Horiba measurements well. It have the same amplitude and keeps up in the transient parts. The contents in the dashed square are shown in the plot in the left corner. Here it can be seen that the transient behaviour is better for the ESV than for the sensor.