View metadata, citation and similar papers at core.ac.uk

brought to you by CORE

provided by Chalmers Publication Library

CHALMERS

Chalmers Publication Library

Data-driven emission model structures for diesel engine management system development

This document has been downloaded from Chalmers Publication Library (CPL). It is the author´s version of a work that was accepted for publication in:

International Journal of Engine Research (ISSN: 1468-0874)

Citation for the published paper:
Grahn, M. ; Johansson, K. ; McKelvey, T. (2014) "Data-driven emission model structures for diesel engine management system development". International Journal of Engine Research

http://dx.doi.org/10.1177/1468087413512308

Downloaded from: http://publications.lib.chalmers.se/publication/193202

Notice: Changes introduced as a result of publishing processes such as copy-editing and formatting may not be reflected in this document. For a definitive version of this work, please refer to the published source. Please note that access to the published version might require a subscription.

Chalmers Publication Library (CPL) offers the possibility of retrieving research publications produced at Chalmers University of Technology. It covers all types of publications: articles, dissertations, licentiate theses, masters theses, conference papers, reports etc. Since 2006 it is the official tool for Chalmers official publication statistics. To ensure that Chalmers research results are disseminated as widely as possible, an Open Access Policy has been adopted. The CPL service is administrated and maintained by Chalmers Library.

(article starts on next page)

Data-driven Emission Model Structures for Diesel Engine Management System Development

Markus Grahn[1, 2], Krister Johansson[1] and Tomas McKelvey[2]

***Abstract*—This paper discusses some specific data-driven model structures suitable for prediction of NO_X and soot emissions from a diesel engine. The model structures can be described as local linear regression models where the regression parameters are defined by two-dimensional look-up tables. It is highlighted that this structure can be interpreted as a B-spline function. Using the model structure, models are derived from measured engine data. The smoothness of the derived models is controlled by using an additional regularization term and the globally optimal model parameters can be found by solving a linear least-squares problem. Experimental data from a 5-cylinder Volvo passenger car diesel engine is used to derive NO_X and soot models, using a leave-one-out cross validation strategy to determine the optimal degree of regularization. The model for NO_X emissions predicts the NO_X mass flow with an average relative error of 5.1% and the model for soot emissions predicts the soot mass flow with an average relative error of 29% for the measurement data used in this study. The behavior of the models for different engine management system settings regarding boost pressure, amount of exhaust gas recirculation, and injection timing has been studied. The models react to the different engine management system settings in an expected way, making them suitable for optimization of engine management system settings. Finally, the model performance dependence on the selected model complexity, and on the number of measurement data points used to derive the models has been studied.**

I. Introduction

Efficient design of advanced engineering systems requires a model based approach. A model is normally a simplified description of a complex scenario, physical unit or system which is tailored towards the intended use in the design process. This paper focuses on models and sub-models for use in diesel engine management system (EMS) design and optimization.

The EMS design is primarily driven by a goal to minimize fuel consumption while keeping emissions within the legislative bounds. The engine emission certification cycles, e.g. NEDC [11], include a large portion of transient operation. Hence, the engine simulation model for EMS design must be at a detail level such that transient engine behavior can be modeled including emissions.

From an EMS design perspective an engine model can be decomposed into a number of interconnected sub-systems. The combustion in the cylinder is primarily dependent on the properties of the ingested air charge, temperatures in the cylinder wall and how the fuel is delivered. The fuel delivery into the cylinder is more or less directly controlled by the EMS while the composition of the air charge and cylinder wall temperature are highly dependent on the past engine history including the past EMS control of the air system components, e.g. EGR valve control, EGR cooling, variable geometry turbine (VGT) and intake throttle settings. The air system is “driven” by the exhaust from the combustion, hence the properties of the exhaust gas from the combustion influences the different air system components. The air system components and the cylinder comprise an interconnected loop which has a dynamic behavior which is important to include in order to accurately model the transient behavior of the total engine system. This dynamic dependence also includes turbo and compressor dynamics. The combustion process results in crankshaft torque which is the driving force of the vehicle. By including a vehicle model, and a driver model making gear-shifts and torque requests a full dynamic drive cycle can be simulated numerically.

Levels of emissions of nitrogen oxides (NO_X) and soot in diesel engines are limited by law. Hence, a simulation based EMS design approach need means to also model the produced amounts of these matters. In this contribution we focus on emission models which can predict the emissions of a diesel engine given partial knowledge of the boundary conditions of the combustion. Examples of such conditions are properties of ingested air charge, engine speed, amount of fuel injected, injection strategy and timing etc.

In general, models can be derived from two opposite directions. The classical approach is to use first principles modeling, leading to multi-dimensional computational fluid dynamic (CFD) models combined with detailed models for the combustion chemistry. Such models give insight into fundamental properties of combustion but are less able to quantitatively predict the levels of emissions and the computation times are very long. Models of this type are described in detail in [22] and [24].

Less demanding is to use zero-dimensional or low-dimensional combustion models, which are based on first principle models, but that are substantially reduced in model complexity. Examples of this are models for NO_X emissions based on the extended Zeldovich mechanism described in [9] and [1]. Although less computationally demanding, these models are too simple to give accurate predictive information [13]. Another example of reduced models is a mean value model for soot emissions described in [18].

On the opposite side of the scale is the use of experimental data from a specific engine geometry where the boundary conditions for the combustions are varied as much as possible.

[1] Department 97500 Complete Powertrain Engineering, Volvo Car Corporation, Gothenburg, Sweden

[2] Department of Signals and Systems, Chalmers University of Technology, SE-412 96, Gothenburg, Sweden

Such data, of course, have excellent predictive performance for the cases covered in the experiment if measurement errors can be neglected but provide no information about engine conditions not included in the experiment set. If the behavior of the predicted variable is assumed to slowly vary in comparison with how densely the conditions are varied during the experiments, a smooth function can be used to interpolate between the experimental points. Such models are known as data-driven, or black-box since the predicted outputs are based on simple functions of the measured data. If an interpolating approach is adopted the complexity of the model in terms of number of parameters is the same as the number of measurements. In some cases when a very large amount of measurements are at hand it might be required, for example due do storage requirements, to use less complex models. Then the experimental data can be used to derive models of lower complexity by minimizing the prediction error for the measured data. Such an approach is also beneficial when measurements are corrupted by noise. A lower model complexity will suppress the influence of the measurement noise on the prediction performance (variance error). However, using a lower model complexity increases the approximation error (bias error) and these two types of errors must be balanced out. This is commonly known as the bias and variance trade-off.

Several different types of data-driven emission models are described in the literature. Examples of this are models based on neural networks [2], [4], models based on Gaussian processes [3], global regression models [5], and global-local model approaches where a global model for the emissions is constructed by switching or weighing between different local models depending on the engine speed and injected fuel operating point of the engine [21], [16], [14].

This paper will discuss some specific data-driven model structures suitable for prediction of $\mathrm{NO_X}$ and soot. The model structures can be described as local linear regression models where the regression parameters are defined by two-dimensional look-up tables. It is highlighted that this structure can be interpreted as a B-spline function and we show how the globally optimal model parameters can be found by solving a linear least-squares problem. Experimental data from a 5-cylinder Volvo passenger car diesel engine is used to derive $\mathrm{NO_X}$ and soot models, where a leave-one-out cross validation approach is used to control the smoothness of the functions.

II. Data-driven prediction models

The basic assumption behind a prediction model is the existence of a functional mapping $f(\cdot)$ from the input space (domain), $x \in \mathbb{R}^m$ to the output space $y \in \mathbb{R}$ (the codomain) also denoted as

$$y = f(x) \tag{1}$$

Given samples of data pairs (x^i, y^i), $i = 1, \ldots, N$ it is desirable to infer the functional relation $f(\cdot)$. This inference can practically be achieved by employing a parametrized function

$$\hat{y}(x, \alpha) = \hat{f}(x, \alpha) \tag{2}$$

where $\alpha \in \mathbb{R}^n$ is the vector of parameters. For practical reasons most often a linearly parametrized model structure is employed

$$\hat{y}(x, \alpha) = \sum_{j=1}^{n} \alpha_j B_j(x) \tag{3}$$

where $B_j(x)$ are known functions, $\mathbb{R}^m \to \mathbb{R}$, and α_j denotes the j-th component in the vector α. Often $B_j(x)$ are called basis functions or regression functions. Given the data the parameter vector α can be determined by minimizing the difference between the data output y^i and the model $\hat{y}(x^i, \alpha)$ using a suitable metric. Employing the Euclidean distance as a metric the optimal parameter vector is

$$\hat{\alpha} \triangleq \arg\min_{\alpha} \sum_{i=1}^{N} \|y^i - \sum_{j=1}^{n} \alpha_j B_j(x^i)\|^2. \tag{4}$$

This metric is favorable from a numerical point of view since the criterion to optimize is convex and an analytical solution exists. Minimizing the Euclidean distance is also equivalent to maximizing the likelihood function, if measurement errors are assumed to have a Gaussian distribution [17]. The minimizing argument is the solution(s) to the set of linear equations

$$\sum_{i=1}^{N} (y^i - \sum_{j=1}^{n} \alpha_j B_j(x^i)) B_k(x^i) = 0, \quad k = 1, \ldots, n \tag{5}$$

known as the normal equations. Employing a vector notation for the known function values

$$B(x) \triangleq \left[B_1(x), B_2(x), \ldots, B_n(x)\right]^T \tag{6}$$

the predictor can be expressed as an inner vector product

$$\hat{y}(x, \alpha) = \alpha^T B(x) \tag{7}$$

Introducing the vector

$$\mathbf{y} \triangleq \left[y^1, y^2, \ldots, y^N\right]^T \in \mathbb{R}^N \tag{8}$$

and matrix

$$\mathbf{B} \triangleq \left[B(x^1), B(x^2), \ldots, B(x^N)\right]^T, \tag{9}$$

the minimization problem in (4) can be rewritten as

$$\hat{\alpha} \triangleq \arg\min_{\alpha} \|\mathbf{y} - \mathbf{B}\alpha\|^2 \tag{10}$$

If the matrix $\mathbf{B}$ has full rank and $N \geq n$, the minimizing vector $\hat{\alpha}$ is unique and is given by

$$\hat{\alpha} = (\mathbf{B}^T\mathbf{B})^{-1}\mathbf{B}^T\mathbf{y} \tag{11}$$

If the vector $\mathbf{y}$ is in the range of $\mathbf{B}$ the model will interpolate the data, i.e.

$$y^i = \hat{y}(x^i, \hat{\alpha}), \quad i = 1, \ldots, N. \tag{12}$$

If $n = N$ and $\mathbf{B}$ has full rank, the range of $\mathbf{B}$ is $\mathbb{R}^N$ and the model will be interpolating for any value of $\mathbf{y}$. When $n < N$ we say that the model complexity has a lower dimension than the data and, in general, the model will approximate the data, i.e. $\|\mathbf{y} - \mathbf{B}\hat{\alpha}\|^2 > 0$. The behavior of the prediction model for x-values between data samples x^i is, besides the dependency on the training data samples themselves, also dependent on how the basis functions $B_j(x)$ are chosen.

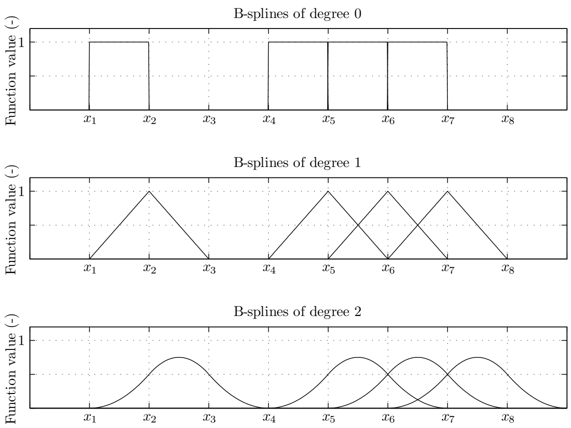

Fig. 1. Illustration of B-splines of degree 0, 1, and 2. The figure shows one isolated B-spline and several overlapping ones.

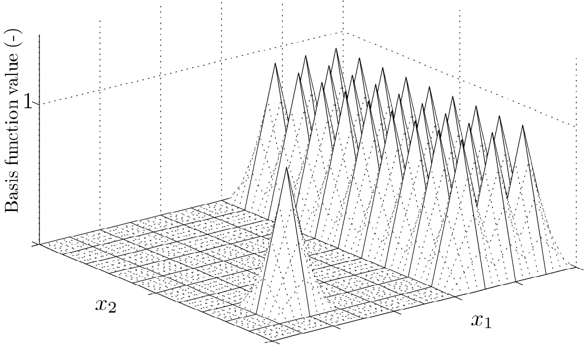

Fig. 2. Illustration of two-dimensional B-splines of degree 1. The figure shows one isolated B-spline and several overlapping ones.

A. B-spline functions

B-spline functions are well suited for interpolating and approximating data-driven models [7], [8]. B-spline is short for basis spline, and a B-spline is constructed by smoothly joining polynomial segments. The points on the indexing axis, where the segments come together, are called knots. The shape of a B-spline depends on its degree q. Scalar B-splines of different degrees are shown in Figure 1, which illustrates the general properties of a B-spline of degree q:

- it consists of $q+1$ piecewisely constructed polynomials, each of degree q.
- the polynomial pieces join at q inner knots.
- at the joining knots, derivatives up to order $q-1$ are continuous.
- the B-spline is positive on a domain spanned by $q+2$ knots, everywhere else it is zero.
- at a given x, $q+1$ B-splines are non-zero.

A function built up by B-splines is called a B-spline function. A scalar B-spline function $\mathbb{R} \rightarrow \mathbb{R}$ with a given knot distribution can be expressed in a similar manner as in (3)

$$\hat{y}\left(x,\alpha\right) = \sum_{j=1}^{n} \alpha_j \beta_j\left(x\right) \tag{13}$$

where $\hat{y}\left(x,\alpha\right)$ is the function value for the input value x, α_j is the B-spline coefficient for B-spline number j, $\beta_j\left(x\right)$ is the value for B-spline number j at the input value x, and n is the number of B-splines used to build up the function. For a B-spline of order 1 and a knot placing which coincides with the data x^i, $n=N$ and $\hat{\alpha}_i = y^i$ and the B-spline will interpolate the data. In this case the interpolation will be locally linear between two data samples. A multivariate function $\mathbb{R}^m \rightarrow \mathbb{R}$ can be generated from scalar B-spline functions in several ways. An *additive* B-spline function can be defined as

$$\hat{y}\left(x,\alpha\right) = \sum_{k=1}^{n}\sum_{j=1}^{m} \alpha_{j,k} \beta_{k,j}\left(x_k\right) \tag{14}$$

If scalar splines for each component of the vector x is multiplied together a *tensor* spline function is obtained. A bivariate tensor spline function, $\mathbb{R}^2 \rightarrow \mathbb{R}$, can be expressed as

$$\hat{y}\left(x,\alpha\right) = \sum_{k=1}^{n_1}\sum_{j=1}^{n_2} \alpha_{k,j} \beta_k\left(x_1\right)\beta_j\left(x_2\right) \tag{15}$$

In Figure 2 the 2-D splines resulting from the tensor product of two scalar splines of order 1 is illustrated. The classical 2-D table lookup technique using linear interpolation between tabulated values is also called bilinear interpolation. For a given value $x = [x_1, x_2]^T$ within the square of the 2+2 axis-points, the local interpolation of y is given by

$$y = \gamma_0 + \gamma_1 x_1 + \gamma_2 x_2 + \gamma_3 x_1 x_2 \tag{16}$$

where γ_i, i=1,...,4 are selected such that $y = y^i$ for the four axis-point values x_i. The 2-D tensor spline function constructed from scalar splines of order 1 yields the same interpolating function as the bilinear interpolation table lookup technique.

B. Regularization

The method of fitting B-spline functions to given data points leads to the minimization problem defined in (4) and (10) . If the matrix $\mathbf{B}$ has full rank, this minimization problem has only one solution. However, if the matrix $\mathbf{B}$ is rank deficient, there are several solutions which minimize the error. In practice, this typically occurs when there are too few measured data points available within some range of the B-spline function, and if the basis functions are not chosen adequately. Also if the matrix $\mathbf{B}$ is ill-conditioned the solution to (10) most likely will lead to poor prediction models. To handle this, the concept of regularization can be used. The idea is to introduce a penalty for the difference between the B-spline coefficients corresponding to adjacent knots [20]. This results in a smoothing effect. A non-negative tuning parameter regularizes the influence of the penalty, with large values leading to heavy smoothing and vice versa. Including a term for regularization the minimization problem becomes

$$\min_{\alpha}\left(\|\mathbf{y} - \mathbf{B}\alpha\|^2 + \lambda\|\mathbf{D}_d\alpha\|^2\right) \tag{17}$$

where the regularization coefficient λ is a tuning parameter, and the matrix $\mathbf{D}_d$ constructs d_{th} order differences of α. The first difference of α, $\mathbf{D}_1\alpha$, is the vector with elements $\alpha_{j+1} - \alpha_j$, $j = 1 \ldots m$. The matrix $\mathbf{D}_1$ is sparse, with $d_{j,j} = -1$ and $d_{j,j+1} = 1$ and all other elements zero. By repeating this computation, we arrive at higher differences. $\mathbf{D}_2\alpha$ is the vector with elements $\{(\alpha_{j+2} - \alpha_{j+1}) - (\alpha_{j+1} - \alpha_j)\}$. It is highly unlikely that the possible null space of $\mathbf{B}$ intersects with the structured null space of $\mathbf{D}_d$ of dimension d, leading to that the construction $\mathbf{B}^T\mathbf{B} + \lambda\mathbf{D}_d^T\mathbf{D}_d$ has full rank and the solution to (17) is

$$\hat{\alpha}_\lambda = \left(\mathbf{B}^T\mathbf{B} + \lambda\mathbf{D}_d^T\mathbf{D}_d\right)^{-1}\mathbf{B}^T\mathbf{y} \tag{18}$$

For two-dimensional B-splines, the regularization matrix $\mathbf{D}_d$ has to be adapted to be used in a two-dimensional structure.

It can be noted that for two-dimensional B-spline functions, smoothing by means of regularization can be controlled individually in the different axis directions of the B-spline function depending on which rows in the regularization matrix that are included. Smoothing can be applied to only one of the axis directions, or to both directions. Also, smoothing can be applied to act in both directions, but with different scaling by using different values for different rows in the regularization matrix.

Regularization introduces a tradeoff between fitting the model to the data and smoothness of the model. Various methods to find an optimal value for the regularization coefficient λ are discussed in for example [10] and [23]. Here, a leave-one-out cross-validation strategy (LOOCV) is employed. This is performed by removing one of the data points, estimating the model parameters using all other data points, and then evaluating the prediction error for the point that was removed. This calculation is repeated for each of the data points, and the root mean square error (RMSE) of the predictions is calculated. The regularization coefficient λ is chosen such that the RMSE of the leave-one-out predictions is minimized.

III. Data-driven emission models

The emission model structure introduced in [13] is a regression model where different emission affecting signals are used as independent variables (inputs) and where all the regression parameters are given by two-dimensional bilinear interpolation maps with the engine speed and the injected fuel amount as inputs. Thus, for a given engine speed and injected amount of fuel, the emissions are predicted by a linear regression model and hence predict how changes in the independent variables affects the emission. The idea behind the model structure is that most of the actuators in a typical diesel engine management system are controlled by a feedforward map based on the engine speed and injected fuel. Therefore, using a regression model where all regression parameters also are dependent on the engine speed and injected fuel, enables the model to handle variations in emission affecting states that depend on speed and fuel without having to include them directly as inputs to the regression model.

The model structure can be described mathematically as

$$\hat{y}\left(x,z\right) = f_0\left(x_1,x_2\right) + \sum_{i=1}^{n} z_i \cdot f_i\left(x_1,x_2\right) \tag{19}$$

where $\hat{y}\left(x,z\right)$ denotes the predicted emission, x_1 and x_2 are the input signals engine speed and injected fuel respectively, z_i are other emission affecting input signals to the model, and finally $f_0\left(x_1,x_2\right)$ and $f_i\left(x_1,x_2\right)$ are model parameters represented by two-dimensional bilinear interpolation maps or equivalently, two-dimensional tensor spline functions of degree 1.

Given a fixed knot-spacing in the x space for the two-dimensional spline functions, it is clear that the predicted emission $\hat{y}\left(x,z\right)$ is an affine function of the parameter vectors α for each of the spline functions. Hence, the model structure can be written in the standard form (3). Fitting the model to data including a regularization term is then given by (17). The model structure has been implemented for modeling of NO_X and soot emissions from a passenger car diesel engine.

A. NO_X modeling

The formation of NO_X is strongly dependent on the availability of oxygen, and the temperature [15]. To include these physical properties for the combustion process, the input signals chosen for the model for NO_X emissions, in addition to the engine speed and the injected fuel, are the injection timing, the pressure in the intake manifold and the ratio of oxygen in the intake manifold. Furthermore, as described in for example [5] and [25], the NO_X emissions have been found to correlate better with exponentials of different input signals. A possible explanation for this could be that the chemical reactions responsible for NO_X formation have reaction rates and equilibrium with exponential behavior. Furthermore, the model structure is used to model a dimensionless measure of the NO_X emissions. A dimensionless measure is created by dividing the NO_X mass flow with the engine speed and injected fuel amount according to

$$N\tilde{O}_X = \frac{60 \cdot NO_X}{2.5 \cdot x_1 \cdot x_2} \tag{20}$$

where $N\tilde{O}_X$ denotes the dimensionless measure of the NO_X emissions (-), NO_X the NO_X mass flow (mg/s), x_1 the engine speed (rpm), and x_2 the injected fuel amount per cycle (mg/cycle). The value 60 in the equation translates the engine speed from revolutions per minute to revolutions per seconds, and the value 2.5 is the number of combustion cycles per engine revolution in a five cylinder engine. The dimensionless measure can be interpreted as mass of NO_X emissions relative to injected fuel mass. Instead of directly modeling $N\tilde{O}_X$ with the regression model the logarithm of the emission level is used. The complete structure for the NO_X model is

$$\log\left(\frac{60 \cdot N\hat{O}_X}{2.5 \cdot x_1 \cdot x_2}\right) = f_{N0}\left(x_1,x_2\right) + \sum_{i=1}^{3} z_{Ni} \cdot f_{Ni}\left(x_1,x_2\right) \tag{21}$$

where z_{N1} denotes the injection timing, z_{N2} the fraction of oxygen in the intake, z_{N3} the pressure in the intake manifold,

TABLE I
INPUT AND OUTPUT SIGNALS FOR THE MODEL FOR NO_X EMISSIONS

	Name	Description	Unit
Output signal:	$\hat{NO}_X$	Estimated NO_X mass flow	g/s
Input signals:	x_1	Engine speed	rpm
	x_2	Injected fuel amount	mg
	z_{N1}	Injection timing	CAD
	z_{N2}	Oxygen fraction in the intake	-
	z_{N3}	Intake manifold pressure	Pa

TABLE II
INPUT AND OUTPUT SIGNALS FOR THE MODEL FOR SOOT EMISSIONS

	Name	Description	Unit
Output signal:	$\hat{soot}$	Estimated soot mass flow	mg/s
Input signals:	x_1	Engine speed	rpm
	x_2	Injected fuel amount	mg
	z_{s1}	Injection timing	CAD
	z_{s2}	Global equivalence ratio	-
	z_{s3}	Partial pressure of oxygen in the intake	Pa

and f_{N0}, f_{N1}, f_{N2} and f_{N3} are two-dimensional bilinear interpolation maps with the engine speed and the injected fuel as inputs. The output and input signals to the model for NO_X emissions are summarized in Table I. When optimizing the model parameters in (17) the i-th component of the vector $\mathbf{y}$ is given by $\log\left(\frac{60\cdot NO_X^i}{2.5\cdot x_1^i\cdot x_2^i}\right)$ where the superscript i on the variables denotes the i-th value in the data set. This means that the model is fitted in the $\log$ domain.

B. Soot modeling

Soot formation and soot oxidation are the two important mechanisms influencing the engine-out level of soot emissions. The formation of soot is mainly dependent on the equivalence ratio. Large amount of soot is formed when combustion takes place at high equivalence ratios within the cylinder. The oxidation of soot is mainly dependent on the temperature and the availability of oxygen late in the combustion phase [15]. To be able to represent the main mechanisms for soot formation and oxidation, the signals chosen as inputs for the model, besides engine speed and injected fuel amount, were the global equivalence ratio, the injection timing, and the partial pressure of oxygen in the intake manifold. Furthermore, also the soot emissions have been found to correlate better with exponentials of different input signals [5], [25]. Similarly, as for the NO_X emission model, a dimensionless measure of the soot emissions is created by dividing the soot mass flow with the engine speed and the injected fuel amount. The complete structure for the soot model is given by

$$\log\left(\frac{60\cdot \hat{soot}}{2.5\cdot x_1\cdot x_2}\right) = f_{s0}\left(x_1, x_2\right) + \sum_{i=1}^{3} z_{si}\cdot f_{si}\left(x_1, x_2\right) \tag{22}$$

where $\hat{soot}$ denotes the estimated soot mass flow, x_1 the engine speed, x_2 the injected fuel amount, z_{s1} the injection timing, z_{s2} the global equivalence ratio, z_{s3} the partial pressure of oxygen in the intake, and f_{s0}, f_{s1}, f_{s2} and f_{s3} are two-dimensional bilinear interpolation maps with the engine speed and the injected fuel as inputs. The output and input signals to the model for soot emissions are summarized in Table II.

C. Engine measurement data

To derive and validate the models, measurement data from a 5-cylinder Volvo diesel engine were used. The engine is equipped with a common-rail injection system, a turbocharger with variable geometry, charge air cooling, an exhaust gas recirculation (EGR) system with cooling, and has a displacement volume of 2.4 liters.

Measurements were performed on the engine in the complete speed and load operating area, ranging from 750 to 4750 rpm and from 0 to 60 mg of injected fuel per cylinder and cycle. Within the range, a number of speed/fuel operating points were selected, and for each of the selected operating points a set of experiments were carried out according to a D-optimal design of experiment methodology, using the engine speed, the amount of injected fuel, the injection timing, the duty cycle to the EGR valve, and the duty cycle to the variable geometry turbine (VGT) as control variables. For each selected operating point the engine speed was varied within a range of 500 rpm, the injected fuel amount within a range of 10 mg, the injection timing within a range of 12 CAD, and the duty cycle to the turbine and EGR valve within the full working range for each set of experiments. This means that the working range of the combustion system was exploited close to as fully as possible regarding the engine air system and the injection timing, using only steady-state engine operation. In total, 3713 steady-state measurements were performed on the engine using this methodology.

The fuel rail pressure and the injection strategy were set according to the settings in the engine management system, and therefore depended only on the engine speed and amount of injected fuel. The injection strategy varied between using a minimum amount of two injections per cycle to using up to four injections per cycle. Injection masses for the different injections and dwell times between the injections were different for different engine speed/fuel operating points.

The engine was equipped with sensors such that the pressure in the intake manifold, the fresh air mass flow, and the exhaust gas recirculation mass flow could be measured. The engine was also equipped with measurement systems for NO_X and soot emissions. A Horiba chemiluminescence measurement system was used to measure NO_X emissions and an AVL Smoke Meter was used to measure soot emissions. From the EMS, the engine speed, the injection timing, and the injected fuel mass were registered.

IV. RESULTS

A. Model complexity investigation

The complexity of the given model structure in terms of number of parameters is directly proportional to the number of spline knots used in the two-dimensional B-spline functions f_i

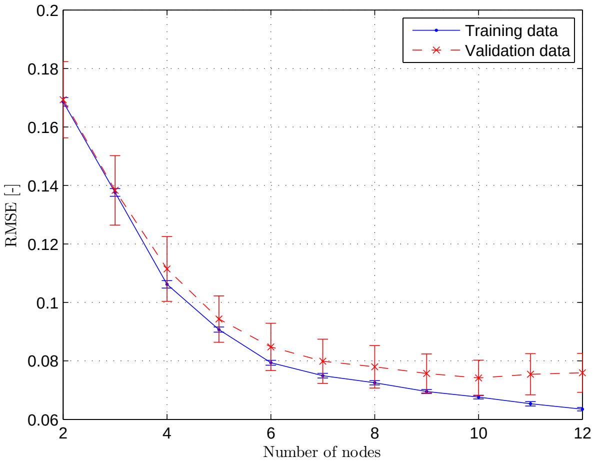

Fig. 3. The performance of the model for NO_X emissions for different number of nodes used in the interpolation maps. The solid blue line shows the RMSE of the logarithm of the dimensionless NO_X emissions for the data points used to optimize the model parameters, and the dashed red line shows the RMSE for the model validation data points. The mean value and standard deviation for 100 models using different data sets for model parameter optimization and model validation are shown for each number of nodes.

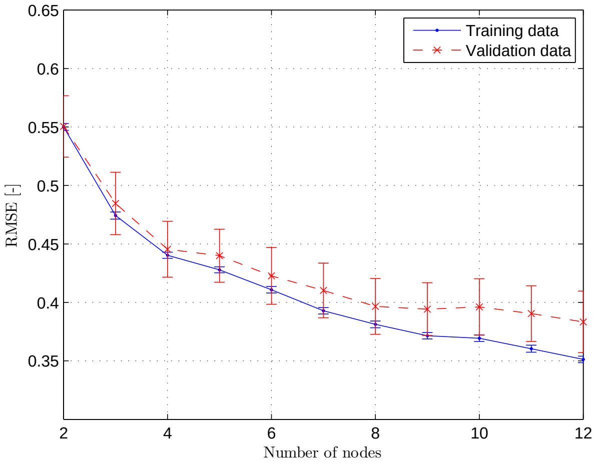

Fig. 4. The performance of the model for soot emissions for different number of nodes used in the interpolation maps. The solid blue line shows the RMSE of the logarithm of the dimensionless soot emissions for the data points used to optimize the model parameters, and the dashed red line shows the RMSE for the model validation data points. The mean value and standard deviation for 100 models using different data sets for model parameter optimization and model validation are shown for each number of nodes.

in (21) and (22). Different number of nodes, i.e. spline knots, were tested for the models, ranging from two nodes to twelve nodes in each of the two dimensions. For each of the different number of nodes, equally distributed points between 750 and 4750 rpm and between 5 and 60 mg injected fuel per cycle were added for the two input dimensions respectively.

From the measurements, 90% of the data points were randomly chosen to be used to optimize the model parameters, and the remaining 10% of the data points were used for validation. The parameters in the models for NO_X (21) and soot (22) emissions were optimized according to (17). A first order difference regularisation was applied to the data fitting, where a leave-one-out cross-validation strategy was used to set the value of the regularisation coefficient λ according to the description in Section II-B. The full procedure was repeated 100 times according to a Monte Carlo simulation methodology using different randomly chosen validation data sets for each tested number of nodes.

The results for the NO_X model and for the soot model, using different numbers of nodes in the interpolation maps are illustrated in Figures 3 and 4. As expected, the prediction performance of the models is poor when very few nodes are used. The performance increases as the number of nodes increase until the number of nodes is around eight for both models. It can be noted that the prediction performance does not significantly decrease again as the number of nodes are increased even further. The reason for this is that the regularization acts as a reduction of the effective model complexity and therefore prevent overfitting. Based on the prediction performance of the models shown in Figures 3 and 4, the number of nodes in the interpolation maps were chosen to be eight both for the NO_X and for the soot model. Using more nodes increase model complexity without significantly increasing the prediction performance of the models.

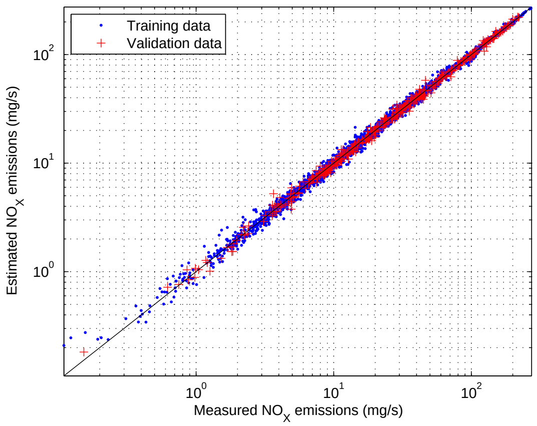

Fig. 5. Measured versus estimated NO_X mass flow when eight nodes are used in all interpolation maps in the model for NO_X emissions. The blue dots show the data points used to optimize the model parameters, and the red crosses the data points used for validation.

B. Emission model results

Using eight nodes, the model for NO_X emissions estimates the NO_X mass flow (g/s) with an average relative error of 5.1% for the validation data. An illustration of the model performance regarding NO_X mass flow is shown in Figure 5. The resulting interpolation maps, $f_{N_i}, i = 0 \ldots 3$, as defined in (21) are illustrated in Figure 6. The model for soot emissions estimates the soot mass flow (mg/s) with an average relative error of 29% for the validation data when eight nodes are used. The model performance regarding soot mass flow is illustrated in Figure 7. The resulting interpolation maps, $f_{s_i}, i = 0 \ldots 3$,

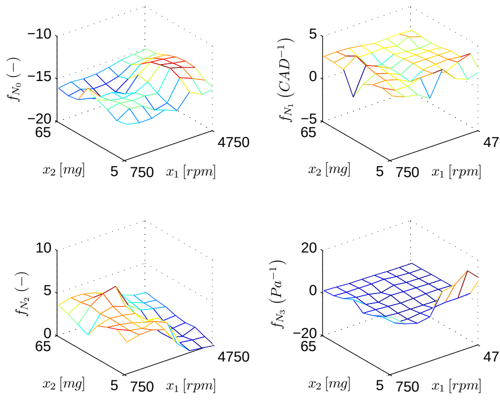

Fig. 6. Illustration of the resulting interpolation maps, $f_{N_i}, i = 0 \ldots 3$, as defined in (21), for the NO_X model when eight nodes are used in all the maps.

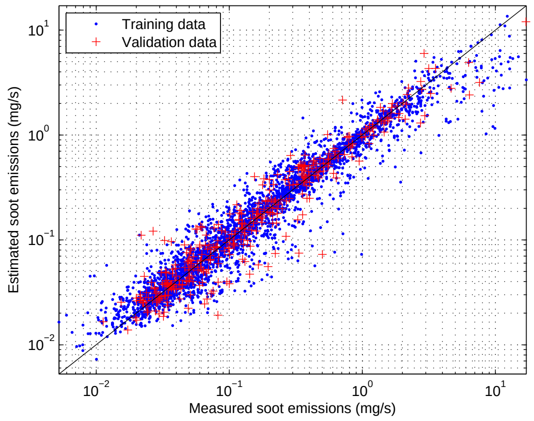

Fig. 7. Measured versus estimated soot mass flow when eight nodes are used in all interpolation maps in the model for soot emissions. The blue dots show the data points used to optimize the model parameters, and the red crosses the data points used for validation.

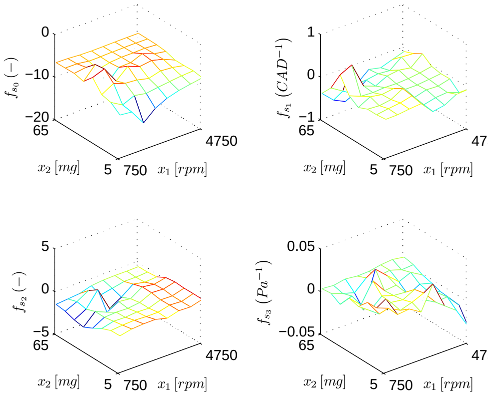

Fig. 8. Illustration of the resulting interpolation maps, $f_{s_i}, i = 0 \ldots 3$, as defined in (22), for the soot model when eight nodes are used in all the maps.

as defined in (22) are illustrated in Figure 8. The interpolation maps shown in Figures 6 and 8 do not have direct physical interpretations. However, as discussed in [13], the emission affecting inputs signals to the models can be chosen to be represented as deviations from nominal values instead of, as here, by their absolute values. By doing this, the resulting interpolation maps would have meaningful interpretations. The interpolation maps f_{N0} and f_{s0} could then be interpreted directly as the logarithm of estimated NO_X and soot emissions during steady-state nominal engine operation, and the interpolation maps $f_{Ni}, i > 0$ and $f_{si}, i > 0$ could be interpreted directly as how deviations from nominal values for the various input signals affect the emissions.

It can be noted that the prediction performance of soot emissions is worse than the prediction performance of NO_X emissions. This is expected, and there are several reasons for this. First, engine-out soot emissions is a result of the difference between formed soot and oxidized soot during the combustion [15]. This means that two different phenomena are relevant for the engine-out soot emissions, making them in general difficult to predict. Also, soot emissions are more difficult to measure accurately than NO_X emissions [19].

To illustrate the behavior of the models for NO_X and soot emissions regarding changes in controllable EMS settings, one engine operating point was chosen. The operating point is defined by; engine speed of 1500 rpm, injected fuel amount of 18 mg per cycle, pressure in the intake manifold of 1.3 bar, ratio of oxygen in the intake manifold of 0.16, and injection timing of 5.5 CAD after top dead center for the main injection. Using this operating point, the pressure in the intake manifold was varied between 1 bar and 2 bar, the oxygen ratio in the intake manifold was varied between 0.15 and 0.21, and the injection timing was varied between 4 CAD before top dead center to 8 CAD after top dead center respectively. The influence on estimated NO_X and soot emissions due to these EMS settings are illustrated in Figures 9, 10, and 11.

For the chosen engine operating point, the NO_X emissions increase and the soot emissions decrease when the injection occurs earlier and vice versa. Also, the NO_X emissions increase and the soot emissions decrease when the boost pressure is increased and vice versa. Finally, the NO_X emissions increase and the soot emissions decrease when the oxygen ratio in the intake manifold is increased (i.e., when the EGR rate is decreased) and vice versa. All this is expected and complies with basic properties of normal diesel combustion as described in e.g. [15].

C. Measurement data availability analysis

An analysis on how the performance of the models for NO_X and soot emissions depend on the number of measured data points used for the model parameter optimization has been performed. First, 10% of the measurements in the complete data set are randomly chosen as validation data points. A given number of the remaining data points are randomly chosen to optimize the model parameters, using the LOOCV method for setting the value of the regularisation coefficient λ, as described in Section II-B. The performance of the resulting

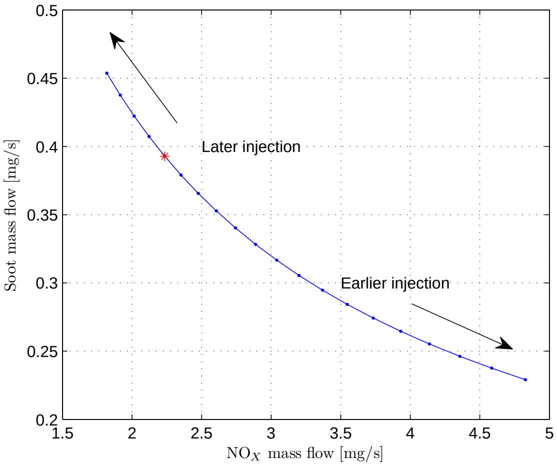

Fig. 9. The red star shows the estimated NO$_X$ and soot mass flows for the engine operating point defined by; engine speed of 1500 rpm, injected fuel amount of 18 mg per cycle, pressure in the intake manifold of 1.3 bar, ratio of oxygen in the intake manifold of 0.16, and injection timing of 5.5 CAD after top dead center for the main injection. The blue line shows the estimated NO$_X$ and soot mass flows when the injection timing is varied from 4 CAD before top dead center to 8 CAD after top dead center.

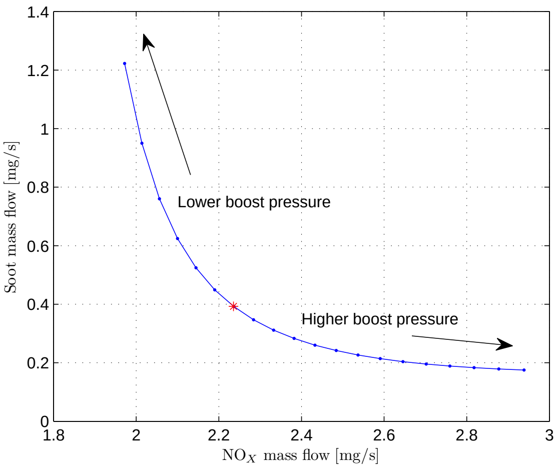

Fig. 10. The red star shows the estimated NO$_X$ and soot mass flows for the engine operating point defined by; engine speed of 1500 rpm, injected fuel amount of 18 mg per cycle, pressure in the intake manifold of 1.3 bar, ratio of oxygen in the intake manifold of 0.16, and injection timing of 5.5 CAD after top dead center for the main injection. The blue line shows the estimated NO$_X$ and soot mass flows when the pressure in the intake manifold is varied between 1 and 2 bars.

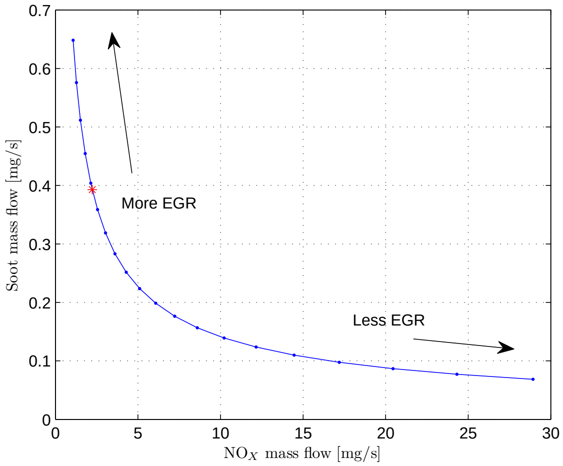

Fig. 11. The red star shows the estimated NO$_X$ and soot mass flows for the engine operating point defined by; engine speed of 1500 rpm, injected fuel amount of 18 mg per cycle, pressure in the intake manifold of 1.3 bar, ratio of oxygen in the intake manifold of 0.16, and injection timing of 5.5 CAD after top dead center for the main injection. The blue line shows the estimated NO$_X$ and soot mass flows when the oxygen ratio in the intake manifold is varied from 0.15 to 0.21.

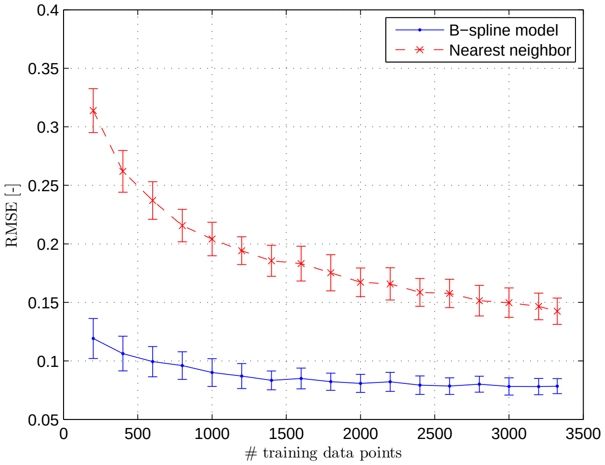

Fig. 12. The solid blue line shows the RMSE for the validation data points using the model for NO$_X$ emissions when different number of data points are used to optimize the model parameters. The dashed red line shows the RMSE for the validation data points when instead using a nearest neighbor model approach. The lines show the mean values and standard deviations of the RMSE for 100 randomly chosen data sets. The figure shows the RMSE of the logarithm of the dimensionless emissions.

model is validated by calculating the RMSE of the validation data points. For a given number of data points for model parameter optimization, this procedure is repeated 100 times, with randomly chosen data sets for parameter optimization and model validation. Finally, the complete procedure is repeated for different number of data points used for the model parameter optimization. Eight nodes were used in all interpolation maps in this study.

As a comparison, the validation points were also estimated using a nearest-neighbor approach for each tested set of parameter optimization and model validation data. Each validation data point was estimated by the value of the point in the parameter optimization data set with smallest Euclidean distance with respect to the model inputs. The inputs were scaled according to the overall working range of the inputs when calculating the distances.

The results of this study for NO$_X$ emission modeling are illustrated in Figure 12, and for soot emission modeling in Figure 13. As seen in the figures, the prediction performances of the B-spline models are uniformly better than the prediction

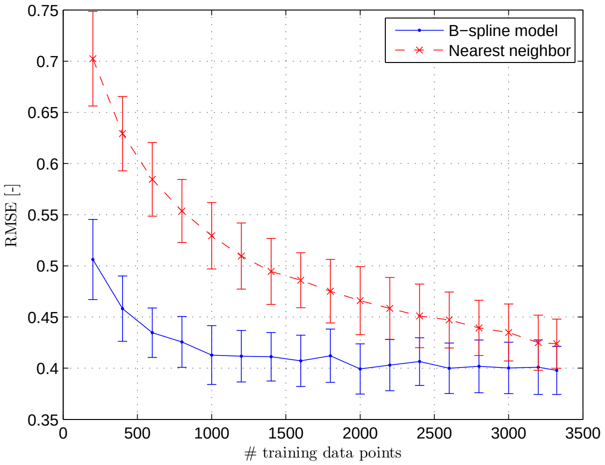

Fig. 13. The solid blue line shows the RMSE for the validation data points using the model for soot emissions when different number of data points are used to optimize the model parameters. The dashed red line shows the RMSE for the validation data points when instead using a nearest neighbor model approach. The lines show the mean values and standard deviations of the RMSE for 100 randomly chosen data sets. The figure shows the RMSE of the logarithm of the dimensionless emissions.

performances when using a nearest neighbor approach. As expected, the prediction results improve when the number of data points used to optimize the model parameters increase. As an example, the number of data points that are needed in order for the prediction RMSE of the models not to increase by more than 10%, compared to when the maximum amount of measurement data is used, is 1400 for the $\mathrm{NO_X}$ model and 600 for the soot model respectively.

V. Discussion

The performance of the models for $\mathrm{NO_X}$ and soot emissions when derived using the B-spline method presented in this paper is better than the performance of the models when they are derived according to the method described in [13]. This is achieved even though fewer nodes are used for the interpolation maps.

Using eight axis-points in each direction for the four B-spline function maps means that each of the emission models have a total of $4 \times 8^2 = 256$ parameters that are fitted using the 3713 data points. The fitted model hence corresponds to a data compression ratio of $3713/256 = 14.5$.

There are several advantages of this proposed method to optimize the model parameters. The models are described mathematically using the B-spline approach, and the model parameters can be calculated directly as an explicit solution to a convex linear least-squares problem. The models are also derived using all measured data points simultaneously, which leads to that measurements do not have to be performed in the structured way as presented in [13], with a particular local design of experiments.

The method of regularization is used to control the smoothness of the interpolation maps. The optimal smoothing parameters in the different interpolation maps, and possibly also in different axis directions of the maps would be interesting to explore further. Also, the influence of the spline knot placing also needs further attention. Various methods for analyzing optimal regularization coefficients and optimal knot placing for B-spline functions are described in e.g. [20].

The presented models for $\mathrm{NO_X}$ and soot emissions react on changes in the EMS settings for the controllable quantities boost pressure, EGR rate, and injection timing. This makes the models useful in the application of optimizing EMS settings with respect to $\mathrm{NO_X}$ and soot emissions. In this paper, the models have been evaluated using only steady-state engine operation. The main difference between steady state engine operation and transient engine operation with respect to emission formation is caused due to the dynamics in the engine air system [12]. Since the models account for deviations in boost pressure and EGR rate, the models could potentially be able to perform accurate prediction results also during transient engine operation. This, however, needs to be verified.

The equivalence between B-spline functions of degree 1 and linear/bilinear interpolation, and the method to use B-spline functions to optimize the parameters in models consisting of interpolation tables and maps could possibly be used in several other applications. Interpolation based models of similar structure are common in a typical engine management system [6], and also in several other systems.

VI. Summary

Data-driven models for $\mathrm{NO_X}$ and soot emissions based on two-dimensional bilinear interpolation maps have been described and studied. The models have been expressed as B-spline functions, and the model parameters have been optimized using measurement data from a 5-cylinder Volvo passenger car diesel engine. The concept of regularization has been used to control the smoothness of the functions, where the degree of regularization has been chosen using a leave-one-out cross validation strategy.

Using eight nodes in all the interpolation maps in the models, the model for $\mathrm{NO_X}$ emissions predicts the $\mathrm{NO_X}$ mass flow with an average relative error of 5.1%, and the model for soot emissions predicts the soot mass flow with an average relative error of 29% for the measurements used in this study. The behavior of the emission models regarding changes in the EMS controllable parameters boost pressure, EGR rate, and injection timing has been studied. The models show an expected behavior that complies with well-known basic properties of diesel engine combustion.

Finally, the prediction performances of the models depending on the model complexity, in terms of number of model parameters, and also depending on the amount of engine measurement data used to optimize the model parameters have been evaluated.

VII. Funding

This work was supported by the Swedish Energy Agency.

References

[1] M. Andersson, B. Johansson, A. Hultqvist, and C. Nöhre. A Real Time NOx Model for Conventional and Partially Premixed Diesel Combustion. *SAE Technical Paper 2006-01-0195*, 2006.

[2] Chris Atkinson, Marc Allain, and Houshun Zhang. Using Model-Based Rapid Transient Calibration to Reduce Fuel Consumption and Emissions in Diesel Engines. *SAE Technical Paper 2008-01-1365*, 2008.

[3] B. Berger, F. Rauscher, and B. Lohmann. Analysing Gaussian Processes for Stationary Black-Box Combustion Engine Modelling. In *Proceedings of the 18th IFAC World Congress*, pages 10633–10640, 2011.

[4] I. Brahma and C.J. Rutland. Optimization of Diesel Engine Operating Parameters Using Neural Networks. *SAE Technical Paper 2003-01-3228*, 2003.

[5] I. Brahma, M.C. Sharp, and T.R. Frazier. Empirical Modeling of Transient Emissions and Transient Response for Transient Optimization. *SAE Technical Paper 2009-01-1508*, 2009.

[6] Urs Christen and Rainer Busch. the Art of Control Engineering: Science Meets Industrial Reality. *2012 IFAC Workshop on Engine and Powertrain Control, Simulation and Modeling*, 2012.

[7] Carl de Boor. *A Practical Guide to Splines*. Springer, 1978.

[8] Paul Dierckx. *Curve and Surface Fitting with Splines*. Oxford University Press, 1995.

[9] R. Egnell. A Simple Approach to Studying the Relation between Fuel Rate, Heat Release Rate and NO Formation in Diesel Engines. *SAE Technical Paper 1999-01-3548*, 1999.

[10] Paul H. C. Eilers and Brian D. Marx. Flexible Smoothing with B-splines and Penalties. *Statistical Science*, 11(2):89–121, 1996.

[11] EU. Directive 98/69/EC of the European Parliament and of the Council of 13 October 1998 relating to measures to be taken against air pollution by emissions from motor vehicles and amending Council Directive 70/220/EEC. *Official Journal of the European Communities*, L 350-41:1–56, December 1998.

[12] William Glewen, David Heuwetter, David Foster, Michael Andrie, and Roger Krieger. Analysis of Deviations from Steady State Performance During Transient Operation of a Light Duty Diesel Engine. *SAE Int. J. Engines*, 5(3), 2012.

[13] Markus Grahn, Krister Johansson, Christian Vartia, and Tomas McKelvey. A Structure and Calibration Method for Data-driven Modeling of NO_X and Soot Emissions from a Diesel Engine. *SAE Technical Paper 2012-01-0355*, 2012.

[14] Christoph Hametner and Stefan Jakubek. Local model network identification for online engine modelling. *Information Sciences*, 220:210 – 225, 2013.

[15] John B. Heywood. *Internal Combustion Engine Fundamentals*. McGraw-Hill, 1988.

[16] Markus Hirsch, Daniel Alberer, and Luigi del Re. Grey-Box Control Oriented Emissions Models. In *Proceedings of the 17th IFAC World Congress*, pages 8514–8519, 2008.

[17] S. Kay. *Fundamentals of Statistical Signal Processing, Volume I: Estimation Theory*. Prentice Hall, 1993.

[18] P. Kirchen and K. Boulouchos. Development and Validation of a Phenomenological Mean Value Soot Model for Common-Rail Diesel Engines. *SAE Technical Paper 2009-01-1277*, 2009.

[19] Z. Gerald Liu and Devin R. Berg. An Analysis of Methods for Measuring Particulate Matter Mass Emissions. *SAE Technical Paper 2008-01-1748*, 2008.

[20] Brian D Marx. P-spline varying coefficient models for complex data. In *Statistical Modelling and Regression Structures*, pages 19–43. Springer, 2010.

[21] M. Mrosek, H. Sequenz, and R. Isermann. Control Oriented NO_X and Soot Models for Diesel Engines. In *Proceedings of the 6th IFAC Symposium Advances in Automotive Control*, pages 234–239, 2011.

[22] Norbert Peters. *Turbulent Combustion*. Cambridge University Press, 2000.

[23] Grace Wahba. *Spline Models for Observational Data*. SIAM, 1990.

[24] Jürgen Warnatz, Robert W. Dibble, and Ulrich Maas. *Combustion: Physical and Chemical Fundamentals, Modeling and Simulation, Experiments, Pollutant Formation*. Cambridge University Press, 4th edition, 2010.

[25] Sebastian Paul Wenzel. *Modellierung der Ruß- und NO_X-Emmissionen des Dieselmotors*. PhD thesis, Otto-von-Guericke-Universität, 2006.