

Input and Structure Choices of Neural Networks on the Fuel Flow Rate Prediction in the Transient Operation Condition

Jiamei Deng
Kingston University

Richard Stobart
Loughborough Univ

Andrzej Ordys
Kingston University

ABSTRACT

Measurement accuracy and repeatability for fuel rate is the key to successfully improve fuel economy of diesel engines as fuel economy could only be achieved by precisely controlling air/fuel ratio and monitor real-time fuel consumption. The volumetric and gravimetric measurement principles are well-known methods to measure the fuel consumption of internal combustion engines. However, the fuel flow rate measured by these methods is not suitable for either real-time control or real-time measurement purposes. The problem concerning discontinuous data of fuel flow rate measured by using an AVL 733s fuel meter was solved for the steady state scenario by using neural networks. It is easier to choose inputs of the neural networks for the steady state scenario because the inputs could be chosen as the particular inputs which excited the system in the application. But for transient scenario, such as the NRTC cycle, it will be difficult to choose inputs as there are no excited inputs that could be chosen. This paper attempts to solve the problem of the input choice for the transient state.

This paper illustrates input choices to catch fast and slow dynamics using both phenomenon models and principal component analysis (PCA) and demonstrates a comprehensive and detailed black-box modelling technique for engine applications in the transient condition. A non road transient cycle (NRTC) is used for system identification. Different neural network structures are compared among feed-forward NN, feed-forward NN with delays and non-linear autoregressive model with exogenous inputs (NLARX). The NLARX is the best structure for fuel flow rate prediction in the transient operation of the engine. It is shown that the input choice is reasonable and could be proved by both theory and simulation results. This paper shows that using chosen the inputs, NLARX could predict fuel flow rate accurately with R-square above 0.99.

KEYWORDS: Neural networks, fuel meters, real time application, discontinuous measurement, input choice

CITATION: Deng, J., Stobart, R. and Ordys, A., "Input and Structure Choices of Neural Networks on the Fuel Flow Rate Prediction in the Transient Operation Condition," *SAE Int. J. Engines* 5(4):2012, doi:10.4271/2011-01-2458.

INTRODUCTION

Improving fuel economy under emissions regulations is one of the challenges ahead for diesel engines. Available Technologies for fuel consumption improvement and emissions reduction are heavily dependent on combustion modification technologies, which consist of both injection technologies and real-time advanced control systems. The specified technologies are the aids or complements to the actions of engines. In this technology, fuel flow rate is used in for control purpose, which implies the measurement of fuel

flow rate plays an important role in implementing the emissions reduction techniques and fuel economy improvement. Therefore, establishing a fuel flow rate model with good prediction performance is essential.

The volumetric and gravimetric measurement principles are well-known methods in measuring the fuel consumption of internal combustion engines. In the volumetric approach, the volume of fuel consumed by the engine is recorded. The fuel consumed is the product of the consumed volume and the temperature-dependent fuel density. The drainage time of a

defined volume is measured and converted to fuel consumption. Unfortunately, the measured variable is not suitable in control due to its fundamental characteristic of discontinuity.

Fuel gravimetric systems, such as the AVL 733S Dynamic Fuel Meter, consist of a measuring vessel filled with fuel suspended on a balance system. The fuel consumption rate is determined by calculating the time related weight loss of the vessel, and is measured continuously. As the fuel meter is refilled, the fuel flow rate cannot be measured and thus the fuel rate measurement becomes discontinuous.

The alternative fuel measurement system designs which are popular in engine fuel measurement applications are based on the Coriolis principle, where the measurement technique is replaced by a controlled servo [1, 2]. These designs are generally more expensive but they have the advantages of both continuous fuel measurement and fast response. However, economic fuel metering systems are more popular solutions for most of applications [3, 4].

Using virtual sensors based on data estimation instead of actual sensors represents a potential solution for this problem which is both economic and reliable. Virtual sensors have been utilized in different areas such as engines, emission monitoring, and vehicle dynamics. A key advantage of Neural networks (NNs) is its ability to be used as an arbitrary function approximation mechanism without knowing the complicated underlying process, which is a highly advantageous and economic way to get the correct fuel flow rate measurement.

Neural networks and fuzzy logic have been successfully used in many engine based applications including emissions, boost pressure and EGR prediction [5, 6, 7, 8, 9, 10, 11]. Despite the success of NNs, the most difficult thing is the choice of inputs. How to choose effective inputs based on easily obtained parameters is of great importance in establishing successful models. Exhaustive inputs search is the most difficult task in establishing models in engine application. Schilling [12, 13] chose these inputs which are based on a sensitivity analysis.

In the paper of [14], the problem concerning discontinuous data of fuel flow rate measured by using an AVL 733s fuel meter was solved for steady state scenario by using neural networks. It is easier to choose inputs of the neural networks for the steady state scenario as the inputs could be chosen as the particular inputs which excited the system in the application. But for transient state, such as NRTC cycle, the choice of inputs will be very difficult as there are no excited inputs that could be chosen.

The selection of the inputs for models has been a complicated problem for the automotive engineering industry for a long time. The paper of [15] has a detailed discussion to choose different possible inputs, but those inputs are highly dependent on the available sensors. So far there are no uniform sensors that could be used in all kinds of applications. This paper attempts to develop a uniform,

simple and practical method to solve the input choice problem in transient scenario.

In establishing a fuel flow rate model in the transient scenario, some general and unique problems are required to be considered which are similar to the difficulties in the paper of [14]:

1. The selection of the inputs is difficult. The basic fuel flow equations could not be used directly since many variables are either unmeasurable or unreliable;
2. The transient information is difficult to capture for fuel flow rate measurement;
3. Capturing both steady state and transient fuel flow rate reliably in a single model is extremely difficult;
4. Models are too complicated for real time control and measurement purposes.

Considering the difficulties, this paper reports on work undertaken to develop fuel flow rate models with transient (torque and speed change) states. Torque and speed are the main operating variables of the engine. It is natural to include the speed and torque as the inputs instead of searching different inputs exhaustively in transient scenario as stated in [13]. Therefore, Problem 1 can be overcome by choosing the torque and speed as the inputs of the fuel flow rate model for the transient scenario, such as non-road transient cycle (NRTC). Problem 2 is solved by using derivatives of torque and speed to capture transient states of fuel flow rate. Problem 3 is settled by using neural networks. In this paper, non-linear autoregressive model with exogenous inputs (NLARX) models is used to predict fuel flow rate. Problem 4 will be overcome to design a transient experiment to validate the model robustness. Problem 4 is overcome by using neural network identifiable properties to find a fuel flow rate model with good performance, then reducing the structure size by minimizing the delay parameters, which has been described in the paper [14]:

This paper reports on work which has successfully chose the inputs for the transient state and identified a NLARX model that captures the dynamics of fuel flow rate with an accuracy of R-square of 0.99 for both training and validation. Torque, speed and their derivatives are used as inputs to catch both fast and slow dynamics based on practical theory analysis and principal component analysis (PCA). The training and validation data are collected using NRTC. The model could be used for measurement and control development in practice. This paper demonstrates a comprehensive and detailed modeling technique that could be used in engine applications. The method is highly flexible and could be extended to solve similar problems in other engine applications.

In summary, the modeling process comprises the following five steps:

Step 1. Study the mechanisms of the fuel flow rate meter. The purpose of this step is to understand the properties and

dynamics of fuel flow rate and choose appropriate inputs and modeling techniques.

Step 2. Use PCA to decide the number of inputs.

Step 3. Develop a data collection experiment to catch all of the dynamics of the fuel flow rate.

Step 4. Construct the 'best' fuel flow rate estimation model structure by going through a process of selecting different neural network structures and then evaluating each performance.

Step 5. Identify the unknown parameters in the 'best' model structure and validate the model and reduce the model size by minimizing the delay parameters.

The above procedures will be described in detail in the following.

The continuous parts of the discontinuous fuel flow rate measurement, as measured by the AVL 733S fuel mass balance, are used to train and validate a neural network model. The discontinuous parts of the fuel flow rate are then predicted by using neural networks. It will be shown that the developed neural network model predicts the fuel flow rate with very good accuracy for discontinuous parts of the fuel flow rate measurement.

Section 2 describes the measurement principles of the fuel flow meter used in this work and details of the fuel supply and metering system. Section 3 states the experimental facility for data acquisition and the real-time measurement system integration. Section 4 reviews PCA and the different neural networks architectures that can be considered as candidates for the fuel flow rate model. Section 5 discusses the input choices. Section 6 outlines the detailed procedure for data acquisition and preparation. Section 7 shows the results of fuel flow rate models in the transient state scenario. Section 8 provides conclusions on this work.

FUNDAMENTALS OF THE FUEL FLOW RATE METER AND EXPERIMENTAL SYSTEM

Basic Principle

Similar to the paper of [14], the principle to the AVL 733s will be described in this section. AVL 733s design utilizes the gravimetric measuring principle in which fuel is supplied to the engine from a measuring vessel, the weight of which is continuously measured (AVL) [16].

A schematic in Fig. 1 outlines the main components of the AVL fuel balance. The measuring vessel is suspended on a bending spring. A mechanism which is designed to be both frictionless and hysteresis-less, translates mass into displacement. The displacement is then measured using a contactless capacitive displacement pick-up which sends the resulting voltage signal to a microprocessor. The system is fully automated in respect to both measurement and calibration procedures.

The measurement vessel has both an engine fuel return line and a venting line thus emulating a vehicle fuel tank. This ensures that there is a continuous separation of air and vapour in the measurement system - the engine feed and return lines.

Advantages of the Gravimetric Fuel Measurement

The gravimetric principle used by the meter has the advantage of direct measurement of fuel consumption mass rate and is thus not subject to the additional measurement needs (fuel temperature) of a similar volumetric system and thus reduces the likelihood of experimental error, i.e. the fuel meter is insensitive to fuel temperature changes which can be significant when the measuring circuit includes a return line from the engine.

A further key advantage is the inherent capability to provide a continuous fuel consumption measurement up to the total mass of the measurement vessel when full. For the larger 1800g vessel in the 733S range, this translates into an ability to continually measure fuel consumption for most passenger car sized engines on many of the internationally known test procedures. Total consumption can thus be determined directly without integration and is thus very reliable.

Limitation of Gravimetric Fuel Measurement

The key limitation with a gravimetric fuel measurement system is the time required to refill the measuring vessel and a pre-determined dampening time. Measures can be taken to reduce the refilling time to a minimum but it remains. The frequency of refills is a function of the physical size of the measurement vessel, the configuration of the fuel supply to the meter, the engine fuel consumption rate and the configuration of the fuel meter - whether it has been set to fill to a volume less than the measurement vessel volume (for increased measurement resolution).

Output Signals, Measurement and Conversion

For the work reported in this paper, the fuel measurement system was configured to use the analogue (0-10V) output of the 733S fuel meter which was sampled at greater than 10 Hz (the measurement rate of the fuel meter) by a National Instruments PXI-6255 card mounted in a NI PXI-1042Q chassis. The acquired voltage was then converted into instantaneous fuel rate using a calibration equation, this being a function of the configured measurement parameters of the fuel meter.

EXPERIMENT FACILITY

The engine that is used in this study is a Cat® C6.6 ACERT™ medium-duty off-highway engine. This is a 6

cylinder engine with a Cat common rail fuel system. The engine calibration used in this work produces up to 159kW at rated speed (2200 rpm) with peak torque of 920 Nm at 1400 rpm. The engine has been modified with a high pressure loop EGR system and a variable geometry turbine for experimental purposes.

The engine is fully instrumented to measure air, fuel and cooling system pressures, temperatures and flow rates. The fuel flow rate is measured by AVL 733S Fuel Mass Balance. Emissions data is gathered principally using AVL 415 smoke and 439 opacity meters (offering steady state and transient measurement respectively) and a Horiba 9100 exhaust gas analyzer measuring nitrous oxide, carbon dioxide, carbon monoxide, unburnt hydrocarbons and oxygen. The temperature and pressure at the exhaust port are measured for each of the six cylinders. This is supported by exhaust manifold, post turbocharger, pre/post EGR cooler/control valve temperature and pressure measurements. [Fig. 2](#) shows the engine facility.

NEURAL NETWORKS AND PRINCIPAL COMPONENT ANALYSIS FOR FUEL FLOW RATE MODELS

The field of virtual sensing has become more and more popular, especially in the field of combustion engine control. This is a direct response to ever increasing system complexity and requirement to precisely control many more sub-systems and hence system response to input signals. Many signals such as exhaust NOx and soot concentration (selective catalytic reduction and diesel particulate filter system control) are difficult and expensive to measure reliably on a high-volume production engines for the service life of the engine. In such situations, virtual sensor techniques have been shown to provide very convincing alternatives [17]. Virtual sensors are characterized through black-box approaches such as neural networks. In this work three types of neural networks are used to predict fuel flow rate.

Feed-Forward Structure

The best known feed-forward network is the multilayer perception. These networks have a number of node element layers and have no connection between the node elements in a particular layer [18]. One layer of nodes forms the input layer while a second forms the output layer, with a number of intermediate or hidden layers existing between them. However, in many cases, only one, two or even no hidden layers are employed. A general multilayer feed-forward neural network is composed of three layers: an input layer, an output layer and an intermediate or hidden layer. Neurons of the input layer only act as buffers for sending the input signal x_i to the neurons of the hidden layer. Each neuron of the hidden layer weights its input signal with the strength of the respective connections w_{ji} , and sums it up, and then computes its output y_j as a function f of the sum:

$$y_j = f\left(\sum w_{ji}x_i\right) \quad (1)$$

here, f is activation function, which can be a simple threshold function or a sigmoid, hyperbolic tangent or radial basis function. The output value in the output layer is computed similarly.

Feed-Forward with Time Delay structure

The first architecture is a simple feed-forward network varying in size and depth defined by layer and neuron numbers and the definition of transfer functions which can have a considerable effect on output performance. The interest in this structure lies in the feed-forward structure. In order to catch some of the dynamics of the systems behavior and its impact on the network performance, a feed-forward network with input time delays is tested. The consideration of previous inputs enables the network to map possible dynamics in the systems behavior. This structure is considered for its slight advantage over the previous due to its possible incorporation of previous dynamics of the system.

NLARX Structure

The third model structure is an NLARX model. The output of the model is feed back into the input layer. In addition, the input allows consideration of previous inputs in order to incorporate dynamics within the systems behavior.

The chosen network structure or architecture is crucial for the output performance. Depending on the systems characteristic if linear or non-linear, static or dynamic; the network needs to be designed accordingly. Here, the prediction of fuel flow rate is known as highly dynamic and non-linear behavior, this means a recurrent network structure could be chosen for sufficient predictive results. The NLARX structure can take into account the dynamics of the system by feeding previous network outputs back into the input layer. It also enables the user to define how many previous output and input time steps are required for representing the systems dynamics best. In this paper, a NLARX model is applied. It represents a recurrent neural network, which fits the purpose of non-linearity of the problem.

The inputs of an NLARX model are represented by $u(n)$, which is a vector of speed, torque, the first derivatives of torque and speed, and the second derivative of torque and the outputs are described by $y(n)$, which is the fuel flow rate. A typical structure of an NLARX model is illustrated in [Fig. 3](#). The formulation of this NLARX model can be described as:

$$y(n) = F(y(n-1), \dots, y(n-n_y), u(n-1), \dots, u(n-n_u), \theta) \quad (2)$$

where n_y is the number of past output terms used to predict the current output, n_u is the number of input terms used to predict the current output.

Each output of an NLARX model is a function of regressors which are transformations of past inputs and past outputs. Usually this function has a linear block and a

nonlinear block. The model output is the sum of the outputs of the two blocks. NLARX model training can be expressed as a non-linear unconstrained optimization problem:

$$\min_{\theta} F_M(\theta, Z_M) = \frac{1}{2M} \sum_{k=1}^M \|y(k) - \hat{y}(k|\theta)\|^2 \quad (3)$$

where $Z_M = [y(k), u(k)]_{k=1, \dots, M}$ is a training data set, $y(k)$ represents the measured output which is the measured PM in the training set, $\hat{y}(k|\theta)$ is the NLARX output which is predicted PM, $\|\cdot\|^2$ is 2-norm operation, θ is a parameter vector, where $\theta = [\theta_1, \dots, \theta_i, \dots, \theta_p]$ and p is the number of parameters.

The optimization problem minimizes the averaged distance between the predicted fuel flow rate outputs and the target fuel flow rate of training samples. Predicted fuel flow rate of an NLARX model is a function of regressors which are transformations of past inputs (torque, speed and their derivatives) and past outputs (fuel flow rate). Usually this function has a linear block and a nonlinear block. The predicted fuel flow rate of the model is the sum of the outputs of the two blocks. Typical regressors are simply delayed input or output variables. More advanced regressors are in the form of arbitrary user-defined functions of delayed input and output variables. The training process is described as follows. Given a neural network described by Equation (2), there is an error metric, that we refer to as performance index of Equation (3), that is to be minimized. This index is a representation of the approximation of the network to some given training patterns. The task will be to modify the network parameters θ to reduce the index $F_M(\theta, Z_M)$ over the complete trajectory to achieve the minimal value. In this paper the neural networks are trained using gradient descent algorithms while the initial value of θ is perturbed several times in order to avoid the local minimal solution. The gradient descent methods will calculate the vector $\nabla_{\theta} F_M$

whose elements are $\frac{\delta F_i}{\delta \theta_i} (i=1, \dots, i, \dots, p)$. The training algorithm will find the parameters of the network for which the performance index has reached a desirable value. Given a vectorising trajectory for the network output and training patterns, the performance index is the Euclidean norm of the error matrix of the whole training batch for the output PM. The training process and minimising the model structure will keep the model size reasonable small, so that model could be used for real time control and measurement is described in details in [14].

Principal Component Analysis (PCA)

PCA is used to reduce the input numbers of the NLARX model. PCA generates a new set of variables, which are principal components. Each principal component is a linear combination of the original variables. All the principal components are orthogonal to each other, meaning that no

redundant information exists. The set of principal components forms an orthogonal basis for the space of the data. While there are many ways to construct an orthogonal basis for several columns of data, PCA assumes that the directions along which the variations are the largest include the most information about the model. The most common approach to construct PCA is to use the standardized linear projection, which construct the maximum variance in the projected space [19, 20]. Suppose there is a q -dimensional data set D , the n principal axes $P_1, P_2, P_3, \dots, P_n$, with $n < q$, are orthogonal axes onto which the retained variance is maximum in the projected space. Generally, $P_1, P_2, P_3, \dots, P_n$ can be given by the n leading eigenvectors λ_i of the sample covariance matrix $C = 1/s \sum_{i=1}^s (d_i - m)^T (d_i - m)$, where $d_i \in D$, ($i=1, 2, \dots, S$), m is the sample mean, and s is the number of samples, so that: $CP_i = \lambda_i P_i$.

The n principal components of D are de-correlated in the projected space. Most of the information in the observation vectors is contained in the subspace spanned by the first m principal axes, where $m < q$. Therefore, each original data vector can be represented by its principal component vector with dimensionality m [19].

The first principal component is a single axis in space. When each variable is projected on that axis, a new variable is formed, whose variance is the maximum among all possible choices of the first axis. The second principal component is another axis in space, which is perpendicular to the first axis. The variable on this axis is projected on this second axis and the second new variable is formed whose variance is the maximum among all possible choices of this second axis. This process continues until all variables are used. The full set of principal components is as large as the original set of variables. However, it is natural for the sum of the variances of the first few principal components to exceed 80% of the total variance of the original data. After studying plots of these few new variables carefully, a deeper understanding of the driving forces that generated the original data could be understood. Therefore, dimensionality could be reduced.

INPUT CHOICE

The choice of the input is the primary and difficult task in the modeling process. In order to find the model of rail pressure, Maione et al. used the engine speed, the fuel tank pressure and duty cycle of the regulator valve opening to identify a neural network model in the common rail injection system of compressed natural gas [21].

If the modeling process of [22] is followed, the fuel flow rate of the fuel meter will be obtained by summing the fuel flow rate of all the injectors. The injection rate $\frac{dm}{dt}$ will have the following equation if we assume the temperature T is constant: