

Empirical Modeling of Transient Emissions and Transient Response for Transient Optimization

Indranil Brahma, Mike C. Sharp and Tim R. Frazier

Cummins Inc

Copyright © 2009 SAE International

ABSTRACT

Empirical models for engine-out oxides of Nitrogen (NO_x) and smoke emissions have been developed for the purpose of minimizing transient emissions while maintaining transient response. Three major issues have been addressed: data acquisition, data processing and modeling method. Real and virtual transient parameters have been identified for acquisition. Accounting for the phase shift between transient engine events and transient emission measurements has been shown to be very important to the quality of model predictions. Several methods have been employed to account for the transient transport delays and sensor lags which constitute the phase shift. Finally several different empirical modeling methods have been used to determine the most suitable modeling method for transient emissions. These modeling methods include several kinds of neural networks, global regression and localized regression. Global regression models have been shown to be the most suitable technique for transient emission predictions. Within global regression models, the best model form, model transforms and model terms have been explored. The chosen models, built on transient data, processed to account for transient lags and sensor delays, have been shown to predict transient emissions and transient response over different transient cycles with reasonable accuracy.

INTRODUCTION

Modern diesel engines systems equipped with complex fuel injection systems, air handling systems, combustion and aftertreatment systems in response to increasingly stringent emission norms have significantly increased the engine calibration effort. As a result, there is a growing realization that model based calibration can yield rich

rewards. Most of the model based calibration effort has traditionally been focused on steady state operation. Only a few researchers like Atkinson et al. [1,40] have ventured to address transient calibration. In the current work, the authors have tried to model transient emissions for the purpose of transient calibration. Due to the nature of the problem and the expected large scale multi-objective optimization, the authors have only looked at empirical modeling methods.

Even though significant advances have been made in steady-state engine calibration processes such as adaptive online data acquisition [2], rule based calibration [3] and simulation based calibration [4,5], function approximators might be the weakest link in the entire process, in part because emission formation in engines is an extremely complex non-linear process, sensitive to a large number of interacting factors. The most commonly used method for steady-state engine calibration is the 'Response Surface Method' or RSMs [6,7] which typically use Design of Experiments (DOE) to obtain data using a structured test plan and then typically use regression models to study the relationship between the independent and response variables. These models then drive the DOEs towards further data acquisition or final engine calibrations. Other examples of practical applications of RSMs for engine system and engine subsystem calibration can be found in the work of Montgomery and Reitz [8,9], Edwards and Pilley [10], Burk et al. [11] and Dimopoulos et al. [12]. Examples of works where distinct engine operating points have been optimized individually or have been considered as a group to meet some overall cycle emission limits while maintaining mechanical and thermal limits at each engine operating point include works by Kampelmuhler et al. [13] who used second order models at each engine operating point to develop an automated calibration process, Schmitz et al. [14] who used a Lagrangian

function for multi-point optimization as part of an calibration process development, Brooks et al. [15] who have used the Matlab Model Based Calibration (MBC) toolbox to develop calibrations to comply with the New European Drive Cycle and Knafl et al. [16] who have used second order models to develop dual-use engine calibrations. Traditional RSM methods have used least-square linear regression models including linear, pure quadratic and full quadratic models, but some recent modeling techniques have demonstrated significant promise and advantages over traditional RSMs. Neural network models used for engine optimization such as [17,18] can be used to represent the entire dataset. A structured test plan or DOEs are not usually required and any amount of non-linearity can be handled. A lot of recent work [27-39,41] has gone into utilizing the power of neural networks for engine modeling. Qiang et al. [19] have used multiple neural networks trained at distinct engine operating points and used fuzzy logic to interpolate at intermediate points for the purpose of engine calibration. Other recently used techniques include Kriging techniques, Radial Basis Functions, Splines and Support Vector machines. Guerrier et al. [20] have examined Linear Growth Models and Radial Basis Function Models for engine calibration. The authors feel that phenomenological models such as the soot model developed by Bayer and Foster [21,22] and Bayesian techniques [23,24] as well as Gaussian Processes [25] show significant promise in engine calibration development. The Design of Experiments for Powertrain Engineering Consortium (DEPE) led by Ricardo Inc. have claimed that their Stochastic Modeling Methods have outperformed polynomials, neural networks and Radial Basis Functions [26].

Except for the works of Atkinson et al. [1,40] who have used neural networks to predict transient engine operation, all the above mentioned works deal primarily with steady state calibration. Previous work by the authors[41] has dealt with modeling for steady state calibration in which the authors compared empirical methods for steady state engine modeling: global regression, localized regression and neural networks (found to work best) and presented a new variant of localized regression. The current work deals with transient modeling only and is different from steady state modeling in a couple of ways. The most important difference is that in case of transient modeling, data acquisition and data processing is not easy and is very important. Secondly the model forms chosen for steady state data are not necessarily suited to transient data. Both these important distinctions are presented in this work. Global regression was found to be better suited to predicting transient data than either neural networks or localized regression, a result that was unexpected and is related to the correlated nature of transient data.

In the following sections, data acquisition, data processing and model forms are treated separately because they are equally important to the success of the transient model.

DATA ACQUISITION

The data for this work was generated by a heavy duty diesel engine and acquired over a series of transient federal cycles as well as specially designed in-house transient cycles. Data was collected at a frequency of 5 Hz and comprised of fuel system parameters such as injected fuel mass, injection pressure, injection timing, air handling parameters such as EGR fraction and fresh air flow and emissions data from a NO_x sensor and opacity meter.

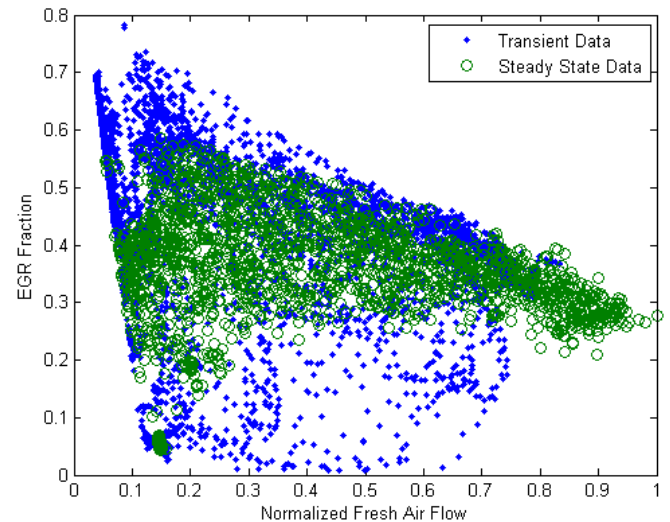


Figure 1. Steady state data in two dimensions overlay-ed on transient data

A variety of transient cycles were used for model building. This variety was preferable but not essential. Based on model predictions built from one kind of transient cycle and used over a different transient cycle, it is expected that the training data for the model might have even comprised of pseudo-transients or a set of transient segments without including data from the cycle over which the transient calibration was being developed. The important requirement observed was that the transient data used for building the model needed to be generated from several different calibrations. In order to understand this, consider the data required for building steady state empirical models. The usual practice is to either conduct a Design of Experiments (DOE) tailored to the model form or in some cases to collect random space filling data which has its own advantages. With transient data, structured DOE's or random space filling is ruled out because the model independent variables are correlated by the particular engine calibration used to generate the data. For an engine running off a calibration, fuel system parameters will be virtually identical and air-handling parameters will tend to be identical (variance resulting from previous history) at the same speed and load. This will cause the data to be highly correlated and an empirical model built off such data will fail when confronted with data correlated differently, i.e. data from a different calibration or the solution space of an optimizer.

The other possibility was to use steady state data to build models. Efforts to use steady state models to predict transient emissions were not very successful because of 2 possible reasons:

a) The transient data spanned areas of multidimensional space that are difficult or impossible to achieve in steady state operation primarily because of air-handling constraints. And it is these extremes of high/low EGR fractions, lean/rich A:F ratios that cause the smoke and NO_x spikes that are so important but the hardest features to predict. We used leverage [41] as a metric to assess and restrict extrapolation. The leverage limit was exceeded for large and important portions of transient data when the model was built from steady state data. For example the engine might momentarily achieve high EGR flows at reasonable fresh air flow rates that cannot be achieved at steady state operation because of turbocharger constraints. Figure 1 demonstrates this by showing steady state and transient data plotted together in the fresh air flow-EGR fraction space.

b.) Transient emissions and some engine independent parameters such as fuel rate were measured differently than steady state operation. An empirical model built from data collected off one measurement system might not work well on data collected off another measurement system.

The second reason can be easily addressed, but the first reason can be overcome with some effort, by designing special air-handling systems including those that are external to the engine. Using steady state data has two major advantages:

a) Transient transport delays and sensor lags need not be addressed

b) The independent variable can be varied randomly or according to a structured DOE, they are not correlated through a set of calibrations as is the case for transient data.

As mentioned before, the authors did not have much success at using steady state models to predict transient emissions, but believe that reason (b) above will result in significant model robustness. The subsequent sections describe the process of creating models from transient data.

DATA PROCESSING

The previous section described why transient data was used for creating empirical transient models for emissions and torque. As mentioned before, data was acquired at a rate of 5 Hz which means that each data point corresponds to the average of 1 to 3 engine cycles, over the FTP cycle. Figure 2 shows the number of engine cycles represented by each data point, over the FTP cycle.

Of course it is desirable to resolve the data as fine as possible so 15 Hz would have been ideal because then there would be no averaging involved. 15 Hz data was not available but attempts to smooth the data and reduce the resolution had adverse effects on the model. Averaging seems to hurt the model, this is intuitive (unless the model is linear in the raw parameter space), because it smudges out the input-output relationship.

Model inputs comprised of engine speed, injected fuel mass, injection pressure, injection timing, EGR fraction and charge flow and model outputs were the emissions data from a NO_x sensor and opacity meter, as well as measured torque. The reason that some data processing is required with transient data is that the model input data at any time does not correspond exactly to model output data at that time, i.e. there is a phase lag between engine events and emission or torque events. Emission and torque phase lags are discussed below:

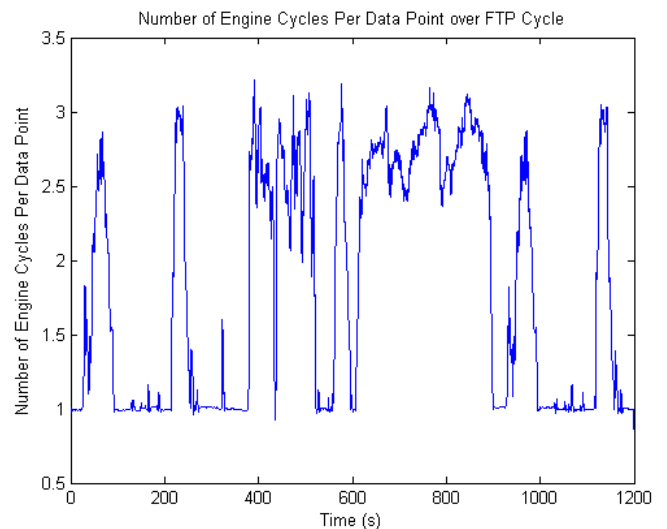


Figure 2. Number of engine cycles represented by each data point, over the FTP cycle.

EMISSION PHASE LAGS

For NO_x emissions captured by a sensor on the exhaust manifold, the phase lag between measured engine parameters and recorded NO_x measurements comprised of the transport delay plus sensor lag. For modeling purposes, a constant phase lag was assumed at each data point. This lag was determined by building multiple regression models with delayed NO_x outputs, each time increasing the delay by 0.2 seconds (5 Hz data). The delay corresponding to the lowest model residuals was chosen as the constant phase lag comprising of the transport delay and sensor lag. Since the phase lag corresponded to approximately 0.8 seconds and the transport lag was calculated to be very small, the calculated phase lag was assumed to be mostly sensor lag. This corresponded well with the sensor specifications. For smoke opacity measurements the phase lag was also similar, though the transport delay was substantial because of the very downstream

location of the opacity meter and opacity measurement was faster than the NO_x sensor lag.

With the constant phase lag determined as above, the relative standard deviation σ/μ , fell by about 35% for NO_x emissions and 10% for particulate emissions. This relative standard deviation was averaged over all points predicted over various different cycles and is described later.

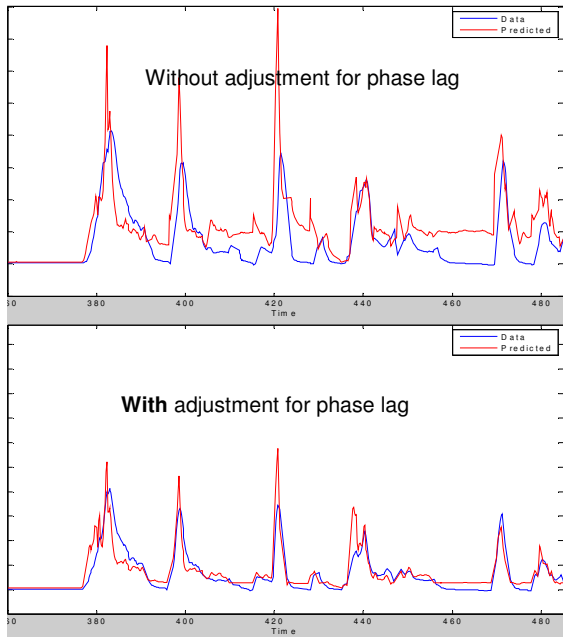


Figure 3. Normalized NO_x predictions with and without adjusting for the constant phase lag.

Figure 3 shows the difference between using a constant phase shift to realign the model independent parameters and model responses as described in the 'Data Processing' section. The constant phase shift approach has the following shortcomings:

- The transport delay should depend on mass flow rate and will not be constant, although the transport delay was small (percent of total) for the NO_x sensor, it was the main component of the opacity lag.
- The NO_x sensor lag might depend on temperature, velocity and composition of the exhaust and might not be constant.
- The inputs to the model are also phase shifted a little bit from each other. The EGR fraction or charge flow measurements in particular might not correspond to the in-cylinder parameters exactly because they were measured outside and some distance away from the cylinder.

None of the above shortcomings have been addressed in the current work because of the following issues. For (a), the transport delay could possibly be calculated but for the NO_x sensor but it seemed negligible and needed to be combined with a fixed (estimate of, not really fixed) sensor lag. while for opacity measurement, the lumped phase lag seemed more convenient because the opacity measurement response time was not known. This will be

addressed in future work. For (b) attempts were made to determine a variable lumped phase lag by determining the delay between NO_x peaks and EGR fraction troughs. Figure 4 shows a section of the FTP cycle at the top and a sub-section zoomed in at the bottom. Each major NO_x peak corresponds to an EGR fraction trough. An algorithm was developed to identify 'major' NO_x peaks (to separate peaks from noise, based on the duration and magnitude of the peak, i.e. if the peak point is above a certain threshold and is greater than a certain number of points on either side, the point is classified as a peak) and find the time lag between the corresponding EGR fraction troughs. This time lag when averaged corresponded well with the calculated constant lag. It can be seen from figure 4 that one 'major' peak has been missed. This is an example of a miss by the algorithm to identify major NO_x peaks because the algorithm has to distinguish between random noise and peaks. If we choose to believe the varying time delay between the peaks and troughs, we could adjust our training data such that independent parameters (inputs) and model responses (outputs) have a variable phase lag at each point. To maintain the one-to-one correspondence between independent parameters and model responses, model responses would need to be interpolated. This method was ultimately not chosen because it was complex, applied only to NO_x emissions (This method is not applicable to smoke emissions because smoke peaks could not be dominated by one single overwhelming parameter like EGR fraction) and attempts to correlate this variable phase lag to exhaust conditions across a range of datasets were unsuccessful.

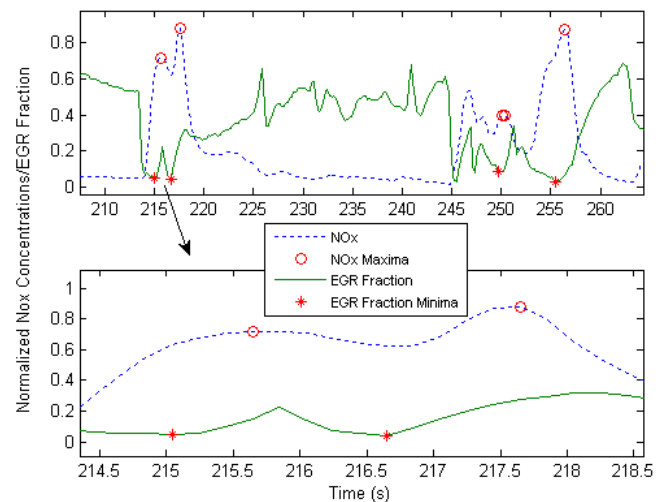


Figure 4. A section of the FTP cycle showing how each NO_x peak corresponds to an EGR fraction trough and the time delay between the two is the phase lag comprising of (primarily) sensor delay and transport delay from EGR fraction measurement location to NO_x sensor.

As for shortcoming (c), it should be possible to synchronize each independent parameter based on measurement locations and flow velocities and the independent variables and model responses would need to be interpolated to generate synchronized training data, but at 5 Hz data rate the interpolation would be feasible.