On the Use of Longitudinal Data Techniques for Modeling the Behavior of a Complex System

Xaro Benavent [1], Francisco Vegara[1], Juan Domingo[1], and Guillermo Ayala[2]

[1] Instituto de Robótica, Univ. de Valencia,
Polígono de la Coma, s/n, Aptdo. 22085,
46071-Paterna (Spain),
{Xaro.Benavent,Francisco.Vegara,Juan.Domingo}@uv.es,
[2] Dpto. de Estadística e Investigación Operativa, Univ. de Valencia
Dr. Moliner, 50.
46100-Burjasot (Spain),
Guillermo.Ayala@uv.es

Abstract. This work presents the use of longitudinal data analysis techniques to fit the accelerations of a real car in terms of some previous throttle pedal measurements and of the current time. Different repetitions of the same driving maneuvers have been observed in a real car, which constitute the data used to learn the model. The natural statistical framework to analyze these data is to consider it as a particular case of longitudinal data.

Different fits are given and tested as a first step in order to explain the relationship between variables describing the control of the car by the driver and the final variables describing the movement of the vehicle.

Results show that the approach can be valid in those cases in which a temporal implicit dependency can be assumed and in which several realizations of the experiment in similar conditions are available; in such cases an analytical model of the system can be obtained which has the ability to generalize, i.e. to show a robust behavior when faced to input data not used in the model construction phase.

1 Introduction

Many natural phenomena or artificial devices can be modeled as input/output systems, i.e. as entities that take signals from their environment by means of direct acquisition from sensors or by explicit data introduction and that generate outputs that change the state of the system itself and/or of its environment. The problem of modeling such systems with a digital computer consists on obtaining a computing device connectable to sensors if appropriate, or at least an algorithm, that accounts for the (a priori unknown) relationship between inputs and outputs.

In general terms, two main approaches can be used to model an unknown system: the symbolic and the non-symbolic way. The first one use knowledge of the internal behavior of the system that is expressed in algorithmic or mathematical terms; in the case of dynamic systems, differential equations are the

P.M.A. Sloot et al. (Eds.): ICCS 2002, LNCS 2329, pp. 458–467, 2002.

most common formulation. The second way of getting a model is the recollection of data taken along the time at regular time intervals or at predetermined instants; this approach can be divided again into parametric and non-parametric models, depending on whether the functional form of the relationship between inputs and outputs is explicitly stated or not. Linear models are an example of parametric models, since the functional form of the relationship between inputs and outputs is explicitly stated by the experimenter [4].

The approach we propose consists on the estimation of a parametric model for the system, which takes elements from both of the aforementioned possibilities: it needs data collected from the real system but it generates an algebraic expression relating inputs and outputs. The use of parametric models is a standard way to work in statistics: the experimenter proposes a reasonable model for the dependency between inputs and outputs, taking into account the statistical variability, which involves the input variables and some unknown parameters and then optimal values (under a given criterion) for these parameters are estimated. Models with increasing complexity or with special assumptions can be proposed if the data come from a physical phenomenon that generates values along the time (time series) and also if the data represent several realizations of the same stochastic process (i.e. several experiments). When both conditions are met, the appropriate theoretical framework to deal with that case is the analysis of longitudinal data, as described in [3]. This work presents an application of such formulation to the restricted modeling of a real car, considering the usual controls of the car (throttle and brake pedals, handwheel, etc.) as inputs and the accelerations measured by appropriate sensing devices on board of the car as outputs. The final purpose of this model is its usage embedded into the software to control a driving simulator; the simulator contains a cockpit resembling that of a real car, including a handwheel, throttle and brake pedals and gearbox lever. All these controls are appropriately sensorized and they are intended to be the inputs of the model. The simulator has been provided, too with solid state accelerometers that allow the measurement of the acceleration that the driver feels. The outputs of the model should be such accelerations, and the platform will have to be moved so as to produce at least scaled versions of them.

This work is organized as follows: section 2 describes the system and the problem specifying the inputs and outputs; section 3 makes explicit the assumed model and the formulation used and shows how the model has been applied to our problem; section 4 details the experiments and results, and finally section 5 states the conclusions and the proposed future work.

2 Description of the system

Driving simulators are becoming popular as a way to evaluate different security and mechanical issues of real cars, and also to test the psychophysical behavior of typical drivers and to evaluate their reactions in different driving situations [6]. Movement is usually reproduced by means of mechanical platforms connected to electrical or electro-neumatic actuators. To make the simulation platform behave in a way similar to a real car, a model of the car itself is needed; this means to know its behavior in terms of the generated movement and acceleration when the driver interacts with the controls (throttle, brake pedal, handwheel, etc.) in the usual way. Up to now, the prevalent approach has been to model the behavior of the vehicle by means of Newton's laws of mechanics, taking into account the known forces and torques ([5]). Differential equations can be programmed so that their output is the state of movement of the vehicle at each moment, which we will have to reproduce by moving the platform appropriately. The set of differential equations use to be quite complex, several simplifications are normally assumed and the knowledge of several parameters is needed, amongst others suspension stiffness, friction coefficient between the vehicle and the floor and others not easily measurable.

On the contrary, our solution starts by experimenting several typical driving maneuvers in a real car that has been sensorized. This means that several sensors and recording devices have been connected to the main controls to get a record of the input values, and also a triaxial accelerometer was installed on board of the car, so that the accelerations along the x, y and z axis of a given coordinate system are measured and recorded. They are the outputs and the model will have to calculate them from the current and former values of the inputs. In this case is clear that the measured acceleration will depend not only on the current readings of the controls, but also on the past history of the car, which has determined its present state, and also on the time elapsed from the beginning of the maneuver. Due to this, the outputs can be considered as time series. It is common in the context of time series the use of ARMA (Autoregressive Moving Average) or ARIMA (Autoregressive Integrated Moving Average) models [9], but this is not appropriate in our case, since the model we intend to determine has to be obtained from several experiments performed in different conditions (throttle pressed up to different levels, etc.) in order to be able to generalize, so it is said, to account for all the different situations that were used to learn, and also for similar ones, since the behavior of the driver when the model is working on the simulator will be similar, but not identical in each occasion. All these reasons have motivated the choice of models based on the formulation of longitudinal data, since it has theoretical properties appropriate to accomplish these requirements.

3 Linear models for longitudinal data

Longitudinal data consist on several continuous or categorical responses taken from one or more experimental units (different repetitions of the same experiment performed independently). The analysis of longitudinal data is closely related with that of time series but it presents two main differences. First, the different time series are considered as a sample of a population. Second, the interest in the correlation structure of longitudinal data is usually minor, but covariance must be adjusted in the process of data analysis to ensure valid inferences on the structure of the mean of the response.

The main subject of the paper is to approximate the observed acceleration of a car from the previous throttle pedal lectures and the current time. All data are observed at the time instants $(t_1, \ldots, t_n)$. Let y_j and x_j be the acceleration and p explanatory variables (some previous throttle pedal lectures, the time and different functions of them) at the time t_j. The simplest model and the classical option is to assume that

$$y_j = \sum_{k=1}^{p} x_{jk}\boldsymbol{\beta}_k + \epsilon_j \tag{1}$$

where $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)$ are unknown parameters and ϵ_j, the experimental error, a normal (or Gaussian) random variable with zero mean and variance σ^2. It is denoted usually as $\epsilon_j \sim N(0, \sigma^2)$ (from now on, $\sim$ means *distributed as*). Furthermore, the different ϵ_j's are assumed independent and identically distributed with variance σ^2. Based on this assumptions, linear model theory permit us to estimate the unknown parameters and to predict the accelerations. However, this approach can not be applied to our case since the different experimental errors are not independent. Figure 1 shows the autocorrelation function of the errors plotted against the time lag when the time and three previous throttle pedal lectures are used to fit the model. Real data of a straight-line acceleration maneuver was used for this example.

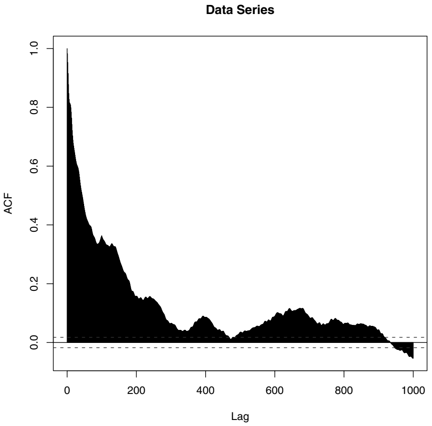

Fig. 1. Autocorrelation function of the residuals observed for a straight-line acceleration maneuver with linear fit (see text for details)

This serial correlation must be taken into account to estimate the parameters $\boldsymbol{\beta}$. We must be aware that have longitudinal data. Different individuals (the different proofs of the same maneuver) are measured repeatedly through time. The natural experimental unit with longitudinal data is the vector $\mathbf{y} = (y_1, \ldots, y_n)$. It is assumed that the same operation is repeated m times, and $\mathbf{y}_i = (y_{i1}, \ldots, y_{in_i})$

are the accelerations observed in the i-th proof at the times $\mathbf{t}_i = (t_{i1}, \ldots, t_{in_i})$ where n_i (which is the total number of times) can be possibly different between different proofs. It is assumed independence between different repetitions but no within a given maneuver observed. This basic hypothesis has to be assumed by the simpler linear model given in equation 1.

It will be assumed that the different $\mathbf{y}_i$'s can be considered as independent realizations of a random vector $\mathbf{Y}_i$ with a multivariate normal (or Gaussian) distribution i.e.

$$\mathbf{Y}_i = X_i\boldsymbol{\beta}' + \boldsymbol{\epsilon}_i \tag{2}$$

being $\boldsymbol{\epsilon} = (\epsilon_{i1}, \ldots, \epsilon_{in_i})$ and

$$\epsilon_{ij} = U_i + W_i(t_{ij}) + Z_{ij}, \tag{3}$$

where U_i is a normal random variable (independent for different i's) with zero mean and variance ν^2; Z_{ij} is another normal random variable with zero mean and variance τ^2 (independent for different i's and j's) and finally W_i is a stationary Gaussian process such that $W_i(t_{ij}) \sim N(0, \sigma^2)$ and whose autocorrelation function is ρ. Two different autocorrelation functions will be used in this work: $\rho(u) = \exp\{-\phi \mid u \mid\}$ and the Gaussian $\rho(u) = \exp\{-\phi u^2\}$. For different i's, it will be assumed that the different realizations of the Gaussian process W_i are independent. Under this model, it can be easily verified that

$$var(Y_i) = V_i = \sigma^2 H_i + \tau^2 I + \nu^2 J, \tag{4}$$

being $H_i(j,k) = \rho(\mid t_{ij} - t_{ik} \mid)$, I the identity matrix and J a matrix of ones. The notation used in the paper has been taken from [3] where the reader can find more details.

The proposed model and the corresponding analysis based on it uses all data jointly. Let $\mathbf{y} = (\mathbf{y}_1, \ldots, \mathbf{y}_m)$ be all the observed accelerations and $\mathbf{t} = (\mathbf{t}_1, \ldots, \mathbf{t}_m)$, the whole set of times. The length of $\mathbf{y}$ and $\mathbf{t}$ would be $N = \sum_{i=1}^{N} n_i$. It is assumed that $\mathbf{y}$ is a realization of a random vector $\mathbf{Y}$ with a multivariate normal distribution given by

$$Y \sim N(X\boldsymbol{\beta}, V(t, \theta)), \tag{5}$$

where $\theta = (\sigma^2, \phi, \tau^2, \nu^2)$ and X is a $N \times p$ matrix where the different X_i's have been stacked. V is a block-diagonal matrix whose non-zero blocks are the V_i's previously considered i.e.

$$\mathbf{Y} = X\boldsymbol{\beta} + \boldsymbol{\epsilon}. \tag{6}$$

The log-likelihood of the observed data $\mathbf{y}$ is then

$$\mathbf{L}(\boldsymbol{\beta}, \theta) = -\frac{1}{2}\{nm \log(\sigma^2) + \sum_{i=1}^{m} \log(\mid V_i \mid)\} + \frac{1}{\sigma^2}(\mathbf{y} - X\boldsymbol{\beta}')' V^{-1} (\mathbf{y} - X\boldsymbol{\beta}'). \tag{7}$$

The parameters $(\boldsymbol{\beta}, \theta)$ will be estimated by using the maximum likelihood estimators (MLE) i.e. the values that give the maximum of the likelihood given by the former equation. The software package *Oswald*, a library of S-PLUS, has been used.[1]

The global behavior of a car can be seen as a juxtaposition of different behaviors depending on the maneuver: acceleration, braking, steering, etc. In this work we will deal exclusively with the straight-line acceleration maneuver, but the behavior of the model could be extended to other driving maneuver by identifying them in the same way as done with the aforementioned maneuver. This was the approach adopted in [1] to build a complete car simulator.

The inputs to the system are the signals that have been obtained from the most important controls of the car: throttle position, brake pedal force, angle of the handwheel and a time vector that will be generated from the beginning of the maneuver. The outputs of the system are the accelerations in the three axes X, Y and Z depending on the maneuver we want to learn. One of the inputs to the system, the time vector, does not refer to any of the controls of the car. This input has been fed into the system because it is important to know if we have pushed the brake pedal or the throttle pedal up to any position (for example, a 50% of its final position) at the beginning of the development of the maneuver or later; indeed, the acceleration of the car will be different in each case.

After doing several tests to determine how many and which previous instants contained relevant information for the dynamic car's modeling, it was decided to use 50 and 100 previous samples from the current instants, for each of the input signals. Given our sampling time, which is $T = 0.01s$, this means half and one second before present.

Since identification was done for each of the driving maneuvers separately, not necessarily all available input had to be used for each of them. For example, straight line acceleration maneuver does not need the brake pedal and handwheel angle, since they are both null in this case. This type of heuristic knowledge may help in the reduction of the dimensionality of the input space.

In order to model the straight-line acceleration maneuver, we need data which are sufficiently representative of the general behavior of the maneuver. In this case we had available data from different runs done for various conditions: throttle pedal pushed at 10%, 30% and 75% of its total allowed run, which are obviously different. Following the notation used by [7] where the concept of *experimental unit* represents the repetition of each of the tests, each experimental unit contains one of the three available data banks for the acceleration maneuver, each of which refers to each of the three different final positions of the throttle pedal.

4 Results

Several experiments have been performed by using the models of equations 2 and 3 with different sets of explanatory variables and assuming different autocorrelation structure. The results are shown in table 1. This table displays the

[1] *Oswald* is a copyright ©1997 David M. Smith and Lancaster University

MLE of the parameters $\hat{\nu}^2$,$\hat{\tau}^2$ and $\hat{\phi}^2$, and the maximum log-likelihood reached. The usual S-PLUS notation for the formulae is used (see [8]).

Table 1. Results with the different models fitted showing the value of the maximum log-likelihood (column headed L) besides the MLE: $\hat{\nu}^2$, $\hat{\sigma}^2$, $\hat{\tau}^2$ and $\hat{\phi}^2$

Num.	Model	$\hat{\nu}^2$	$\hat{\sigma}^2$	$\hat{\tau}^2$	$\hat{\phi}^2$	**L**
1	$y \sim x_1 + x_2 + x_3$ $\rho(u) = exp(-\phi \cdot \lvert u\rvert^2)$	$9.1e^{-9}$	0.223	0.0033	0.0385	-770.9
2	$y \sim x_1 + x_2 + x_3$ $\rho(u) = exp(-\phi \cdot \lvert u\rvert)$	0.517	0.659	$3.79e^{-10}$	0.009	-729.1
3	$y \sim x_1 * x_2 * x_3$ $\rho(u) = exp(-\phi \cdot \lvert u\rvert)$	0.822	0.938	$34.59e^{-11}$	0.0058	-700.8
4	$y \sim poly(x_1, 2) + poly(x_2, 2)$ $+poly(x_3, 2)$ $\rho(u) = exp(-\phi \cdot \lvert u\rvert)$	0.584	0.711	$9.76e^{-11}$	0.0078	-709.1
5	$y \sim poly(x_1, 2) + poly(x_2, 2)$ $+poly(x_3, 2)$ $\rho(u) = exp(-\phi \cdot \lvert u\rvert^2)$	$3.11e^{-8}$	0.2197	0.0032	0.0386	-766.8
6	$y \sim poly(x_1, 2) + poly(x_2, 2)$ $+poly(x_3, 2) + poly(time, 2)$ $\rho(u) = exp(-\phi \cdot \lvert u\rvert)$	0.133	0.263	$3.49e^{-10}$	0.021	-701.9
7	$y \sim poly(x_1, 2) + poly(x_2, 2)$ $+poly(x_3, 2) + tex(t1) + tex(t2)$ $+tex(t3)$ $\rho(u) = exp(-\phi \cdot \lvert u\rvert^2)$	0.089	0.0845	0.0031	0.0501	-703.7
8	$y \sim poly(x_1, 2) + poly(x_2, 2)$ $+poly(x_3, 2) + tex(t1) + tex(t2)$ $+tex(t3)$ $\rho(u) = exp(-\phi \cdot \lvert u\rvert)$	0.042	0.186	$3.94e^{-10}$	0.0298	-697.7

For instance, models 1 and 2 (first and second rows of the table) are denoted by $y \sim x_1 + x_2 + x_3$, which means that the longitudinal data model uses as explanatory variables three throttle pedal lectures in the current and two former instants. The second row indicates the auto-correlation function used. The third model denoted by $y \sim x_1 * x_2 * x_3$ uses as explanatory variables the original variables and all the cross-products of the variables x_1, x_2 and x_3 i.e. x_1x_2, x_1x_3 and so on. The expression $poly(x_1, 2)$ means that the original, x_1 and x_1^2 have been used as explanatory variables. $tex(t1)$ denotes a vector in which time increases linearly; $tex(t2)$ is a vector in which time increases quadratically in the first time interval, up to a certain time t_0 and $tex(t3)$ is a vector in which time increases linearly only from t_0 up to the end of the maneuver. In all the cases t_0 was chosen as 50 sampling periods.

Let us look at the model 4 with more detail. This model uses polynomials of order up to 2 for each of the explanatory variables. The likelihood observed is -709.1 i.e. it is the fifth better fit from this point of view. Note that the parameters estimated are based on three experimental units. The plots in figure 2 show the obtained adjusted data that the model gives for each experimental unit overlaid to the real data. It is clear that the results are not good, since acceleration goes out of the desired range for the real outputs (there are even negative accelerations), but the generalization is acceptable. Nevertheless, the result would be unusable in a simulator since negative accelerations would be opposite to the type of sensation that is expected by the driver.

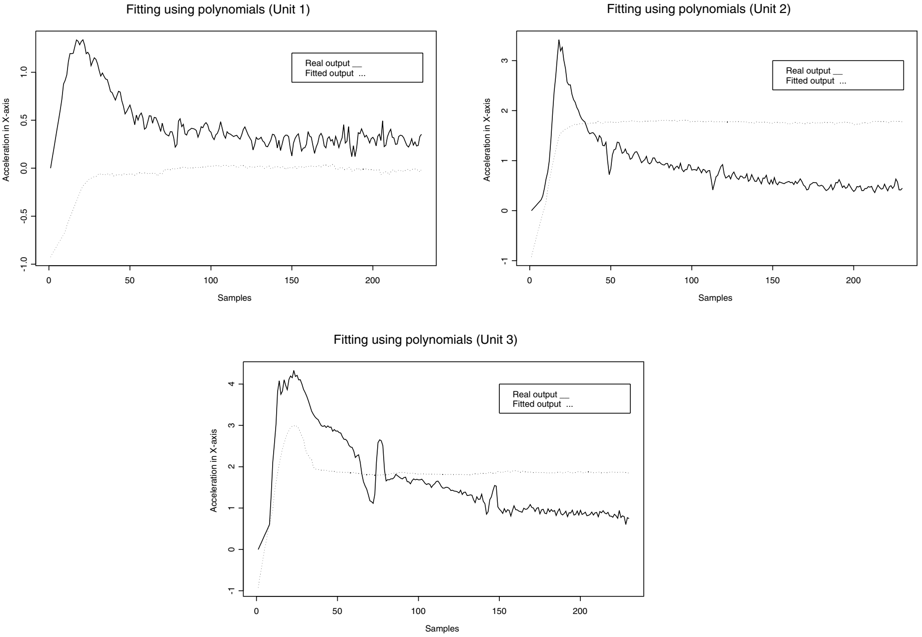

Fig. 2. Comparison between real and fitted values for experimental units 1, 2 and 3 respectively. A model with polynomials of order 2 (experiment 4) has been used.

Model 6 is the first one in which time is used explicitly. Experiments 7 and 8 use the tex function with the two user-defined vectors of time, t_1 and t_2, explained above. These vectors have been created in this way so that a quadratic fitting can be done for the first part of the plot and a linear fitting for the rest of the maneuver. If we observe the behavior of the acceleration in this type of maneuver (output to be fitted) it is clear that the creation of these two vectors is a sensible option. Table 1 shows the obtained results for this type of models; it can be seen that they provide better fits that every previous model giving a maximum log-likelihood of -697.7. Plots in figure 3 show the comparison between the real signal and fitted data. Results are now clearly better; they could be used in our practical case, though not with a completely real sensation. The model could be further refined to get better results by changing the variables in the correlation function and testing other combinations in the formula for the linear model.

Obviously, more explanatory variables and more complex relations provide a better fit. However, they remains two open questions: the variables selection (how many previous pedal lectures and in which times to observe) and a deeper study of the functional relation between the explanatory variables and the observed acceleration.

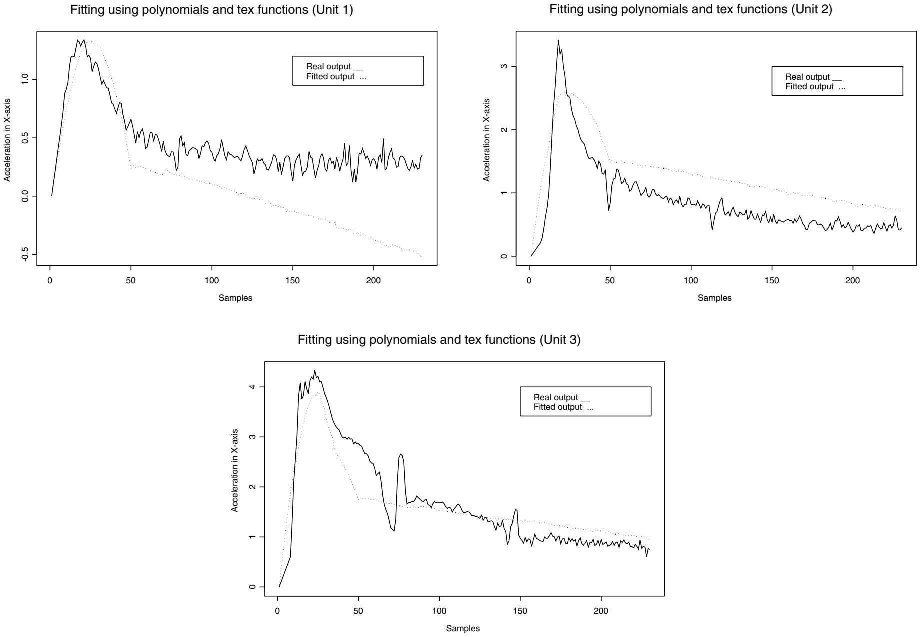

Fig. 3. Comparison between real and fitted values using the formula 8 of table 1 for experimental units 1, 2 and 3 respectively.

5 Conclusions

An important objective of this work was to determine to what extent longitudinal data models can be considered as a valid alternative for the modeling of the dynamic behavior of a real system, or at least of a restricted part of it. The performed experiments shows that, in our case, longitudinal data analysis can be a feasible approach only for sufficiently sophisticated models. Explicit usage of time as an input introduces valuable information, but this can only be done if the initial instant of time (the instant in which the data start to arrive) is known. Moreover, controlling the specific type of dependency of the output with respect to time, and making it different for different time intervals of the experiment increases the goodness of the fit, as can be seen in the second experiment, in which this has been done. On the contrary, the choice of the autocorrelation function of the errors (exponential or Gaussian) does not appear to have a decisive influence, at least for this case.

With respect to the adequacy of the model for its intended purposes, the fit obtained in the second experiment can be considered as sufficient for its use in our driving simulator, given the limited ability of people to perceive absolute values of accelerations or differences of them. Other dynamic systems may require a better fit, but this could probably be achieved by using models involving more complex dependencies between inputs and output. On the other hand, the generalization capabilities of the model are also appropriate for this case, and this is the main reason to choose longitudinal data with preference to simpler linear models for problems involving the control of dynamic systems whose behavior is similar, but strictly different for each trial.

An important practical aspect is the detection of the point in time in which the dependency of the output with time changes, for instance from linear to quadratic, and which roughly corresponds in our case to a change in the automatic gearbox. This could be done by using a different model involving information such as engine speed and remains as a future work.

References

1. X. Benavent. *Modelizacion y control de sistemas no lineales utilizando redes neuronales y modelos lineales. Aplicacion al control de una plataforma de simulacion.* PhD thesis, Universitat de Valencia, 2001.
2. J. P. Chrstos and P. A. Grygier. Experimental testing of a 1994 Ford Taurus for NADSdyna validation. Technical Report SAE Paper 970563, Society of Automotive Engineers, Inc., 1997.
3. P.J. Diggle, K.-Y. Liang, and S.L. Zeger. *Analysis of Longitudinal Data.* Oxford Science Publications, 1994.
4. N.R. Draper and H. Smith. *Applied Regression analysis.* Princeton University Press, 1998.
5. T. D. Gillespie. *Fundamentals of Vehicle Dynamics.* Society of Automotive Engineers, Inc., 1992.
6. S. Nordmark. Driving simulators trends and experiences. In *Driving Simulation Conference. Real Time System 94*, Paris. France, January 1994.
7. D.M. Smith and P.J. Diggle. Oswald: Object-oriented software for the analysis of longitudinal data in S. Technical Report MA94/95, Lancaster University Department of Mathematics and Statistics, 1994.
8. W.N. Venables and B.D. Ripley. *Modern Applied Statistics with S-PLUS.* Springer, third edition edition, 1999.
9. P.J. Brockwell and R.A. Davis. *Time Series: Theory and Methods.* Springer Series in Statistics, Springer, 2nd Edition, 1991.