Proceedings of the IASTED International Conference
Intelligent Systems and Control 2000
August 14-16, 2000 – Honolulu, Hawaii, USA

On Fuzzy Model Identification and the Gas Furnace Data

Wael Farag
Case New Holland Technology Center
7 South 600 County Line Rd.
Burr Ridge, IL 60521-6975, USA
wael.farag@cnh.com

Ahmed Tawfik
University of Windsor
School of Computer Science
Windsor, ON N9B 3P4, Canada
ahmed.tawfik@computer.org

Keywords: Fuzzy modeling, linear modeling, identification, benchmarks, nonlinear modeling.

ABSTRACT

Benchmarks play an important role in the validation of fuzzy and neural modeling approaches. This paper provides new analytical insights into a commonly used benchmark: the gas furnace data set, and some of the results reported for this benchmark in fuzzy systems literature. The linearity of the data set is examined and linear models, which use the same inputs as fuzzy models are presented. Two highly nonlinear systems are proposed as useful benchmarks in nonlinear modeling and control.

I. INTRODUCTION

The gas furnace data set is a commonly used benchmark. The data set consists of 296 input-output measurements sampled at a fixed interval of 9 seconds. The measured input u(k) represents the flow rate of the methane gas in a gas furnace and the output measurement y(k) represents the concentration of carbon dioxide in the gas mixture flowing out of the furnace under a steady air supply. This data set has been used extensively as a benchmark example for process identification. Box and Jenkins [2] used a time-series based approach to develop a model based on the gas furnace data set. This model has been referred to, repeatedly, as the *linear model*. This may be rather misleading because there are several possible linear models that can model this process depending on the choice of inputs and the fitting criterion. The data set has since been used by many researchers including Tong [13], Pedrycz [9], Xu and Lu [15], Sugeno and Tanaka [10], Sugeno and Yasukawa, [11], and Wang [14]. Table I lists the inputs used and the mean squared error (MSE) reported in these studies.

Most studies have used the inputs y(k-1) and u(k-4) which have the highest correlation with the output y(k). The smallest MSE reported appears in the study by Sugeno and Tanaka (1991). However, this MSE seems to be inaccurate.

Study	Inputs	Rules	MSE
Tong (1978)	y(k-1), u(k-4)	19	0.469
Pedrycz (1987)	y(k-1), u(k-4)	81	0.320
Xu (1987)	y(k-1), u(k-4)	25	0.328
Sugeno (1991) FM	y(k-1), y(k-2), y(k-3), u(k), u(k-1), u(k-2)	2	0.068
Sugeno (1991) FM’	y(k-1), u(k-4)	2	0.359
Sugeno (1993)	y(k-1), u(k-3)	6	0.190
Wang (1996)	y(k-1), u(k-4)	5	0.158

Table 1. Reported results.

Implementing the two rules, which are also published in the study, the calculated mean squared error from the rules is 0.247 that leads us to believe that the reported MSE had a misprint. For further validation, we compared the plots appearing in the study to those resulting from our implementation (see Figure 1). After the publication of these results, studies that used this data set including Sugeno and Yasukawa [11] and Wang [14] attempted to explain the impressive performance of the FM algorithm for this particular data set. (Note that the MSE result reported in the same work [10]] for FM’ should read 0.525 (instead of 0.359)).

317-083

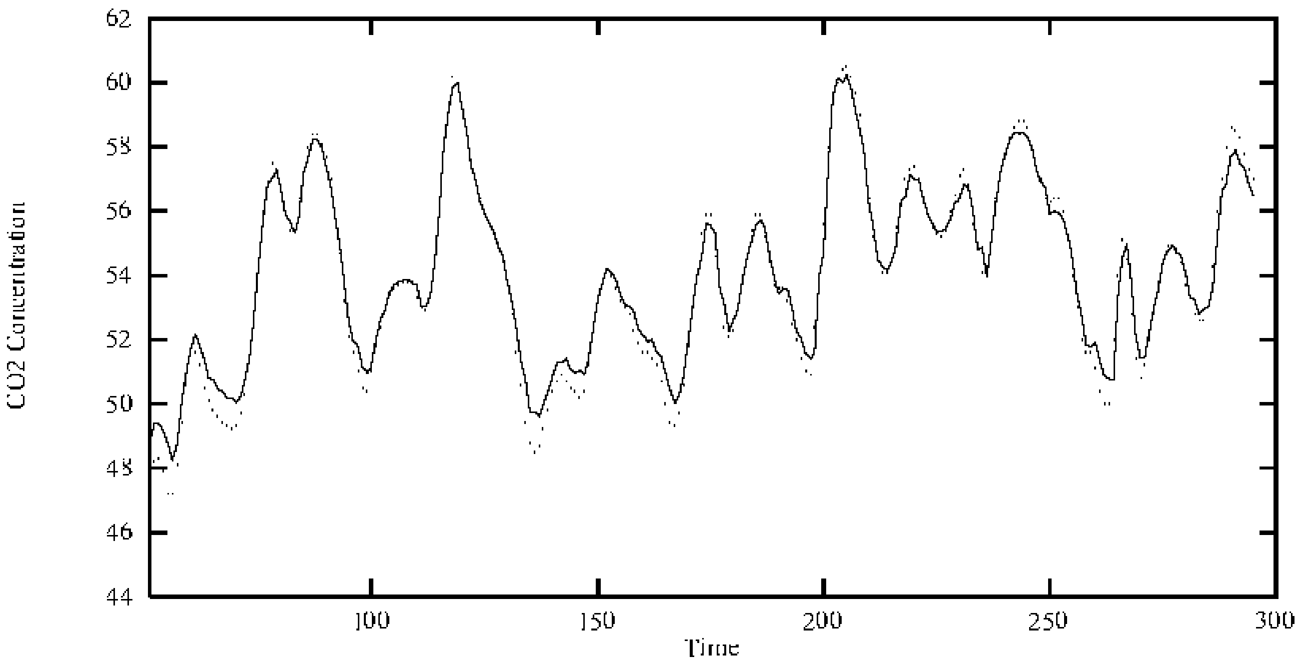


Figure 1. The gas furnace data

II- FUZZY PROCESS IDENTIFICATION

The knowledge representation in fuzzy modeling can be viewed as having two classes. The first (class A) as suggested by *Takagi and Sugeno* [12] can represent a general class of static or dynamic nonlinear systems. It is based on "fuzzy partition" of input space and it can be viewed as the expansion of piecewise linear partition represented as

$$R_i \; : If \; x_1 \, is \, A_1^i \; and \, x_2 \, is \, A_2^i \, and \; \cdots \; and \, x_m \; is \, A_m^i$$

$$then \; y^i = a_0^i + a_1^i x_1 + \cdots + a_m x_m$$

where R_i ($i=1, 2, ..., c$) denotes the i^{th} fuzzy rule, and x_j ($j= 1,2, ..., m$) is the input and y^i is the output of the fuzzy rule R_i. $A_1^i, A_2^i, ..., A_m^i$ $(i=1,2,...,c)$ are fuzzy membership functions which can be bell-shaped, trapezoidal, or triangular, … etc., and usually they are not associated with linguistic terms. Sugeno and Tanaka's [10] approach approximates a nonlinear system with a combination of several linear systems by decomposing the whole input space into several partial fuzzy spaces and representing each output space with a linear equation as in (1). A drawback of this type of knowledge representation is that it does not allow the output variables to be described in linguistic terms.

The second class of knowledge representation (class B) in fuzzy models is developed by *Mamdani* [7] and used by *Lin and Lee* [6] and *Sugeno and Yasukawa* [11]. The knowledge is presented in these models using rules of the form:

$$R_i \; : If \; x_1 \, is \, A_1^i \; and \, x_2 \, is \, A_2^i \, and \; \cdots \; and \, x_m \; is \, A_m^i$$

$$then \; y^i \; is \, B^i$$

where $A_1^i, A_2^i, ..., A_m^i, B_i$ $(i=1,2,...,c)$ are fuzzy membership functions which are bell-shaped, trapezoidal, or triangular, etc., and usually associated with linguistic terms. This approach has some advantages over the first approach. The linguistic terms in the consequent parts make this model more intuitive, understandable and give more insight into the model structure. Also, this modeling approach is easier to implement than the first approach [5].

Table 2 classifies the studies involving the gas furnace data as a Class A model or Class B model based on the above definition.

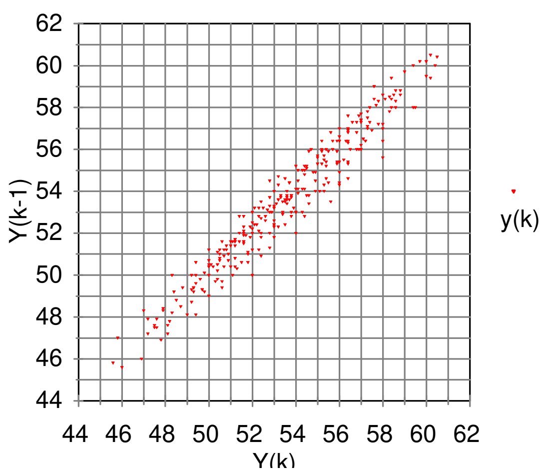
Study	Class	Study	Class
Tong (1978)	B	Sugeno (1991) FM'	A
Pedrycz (1987)	B	Sugeno (1993)	B
Xu (1987)	B	Wang (1996)	A
Sugeno (1991) FM	A	Farag (1998)	B

Table 2. Classification of fuzzy system identification algorithms.

III. THE GAS FURNACE DATA

The ability to model non-linearity is one of the main reasons for the success of control systems using fuzzy control systems. Therefore, a highly nonlinear benchmark would be appropriate for testing different fuzzy control/modeling approaches. However, the gas furnace data set is known to represent an approximately linear input-output behavior. The following quote from [2] indicates that the data set is linear and that the data were taken at a time when the system behaves linearly:

> "In an investigation on adaptive optimization, a gas furnace was employed in which air and methane combined to form a mixture of gases containing CO_2 (carbon dioxide). The air feed was kept constant, but the methane feed rate could be varied in any desired manner and the resulting CO_2 concentration in the off gases measured. The continuous data [of Figure 1] were collected to provide information about the dynamics of the system over a region of interest where it was known that an approximately linear steady state applied. "



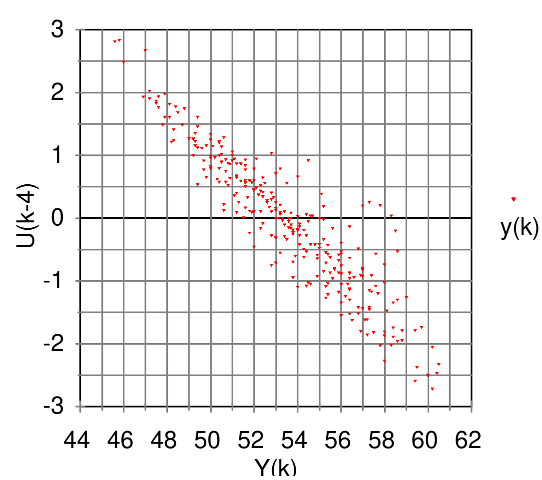


Figure 2. Linearity of the predictors

Linear trends in the relationships between the commonly used predictors: y(k-1) and u(k-4), and the output variable y(k) are apparent from the plots in Figure 2.

Given that the trend underlying the data is linear, a linear model should perform well. A linear model based on the linear model [2] resulted in a MSE of 0.202.

$$y(k) = 0.55\, y(k-1) - 0.52\, u(k-3) - 0.4\, u(k-4) - 0.51\, u(k-5) + N(k)$$

$$N(k) - 1.53\, N(k-1) + 0.63\, N(k-2) = a(k)$$

where N(k) is the noise in sample k and a(k) is white noise.

However, using the two commonly used input: y(k-1) and u(k-4), we get the following linear model:

$$y(k) = 0.67163\, y(k-1) - 1.07163\, u(k-4) + 17.5693.$$

This model results in a mean squared error of 0.179. Another linear model relating y(k) to y(k-1), y(k-2), y(k-3), u(k), u(k-1), u(k-2), u(k-3), and u(k-4):

$$y(k) = 1.625988\, y(k-1) - 0.90389\, y(k-2) + 0.168927\, y(k-3) - 0.07794\, u(k) + 0.225577\, u(k-1) - 0.22026\, u(k-2) - 0.55682\, u(k-3) + 0.2788755\, u(k-4) + 5.8154$$

has a MSE of 0.06051. For all three linear models, the residuals (difference between model predictions and observed outputs) form a completely random sequence with respect to the inputs using correlation analysis. The linear models are therefore valid models [1].

Inputs	MSE
y(k-1) , u(k-3), u(k-4), u(k-5)	0.202
y(k-1),u(k-4)	0.179
y(k-1), y(k-2),y(k-3),u(k), u(k-1),u(k-2) u(k-3), u(k-4)	0.061

Table 3. Linear models with different inputs

Table 3 summarizes the MSE results obtained for linear models and the corresponding inputs. Note that a linear model is restricted enough that the complete dataset is used for both model generation and model testing. To avoid following the noise in the dataset, nonlinear models usually use half the data for model generation and the other half for testing.

IV- FUZZY MODELS OF THE GAS FURNACE

Class A and Class B fuzzy identification methods have both used the gas furnace data for testing. While fuzzy models may not be the most appropriate models for this data set, models from both classes have achieved MSE results comparable to linear models. Realizing the linearity of the data, ideally a Class A fuzzy model would generate a single rule with an empty condition, and a consequence that correspond to the linear model. A Class A model with a large number of rules may indicate a tendency to follow the data too closely instead of generalizing the real underlying trends. Such model may produce a small MSE despite this problem.

For a class B model, the number of rules does not necessarily have the same significance. Ideally, the class B model would deduce the linearity of the data but this may depend on the choice of membership functions. However, the main advantage of a Class B model lies in the linguistic variable characterization of the system which can be easily verified against a linear model. For example, consider the following rule table from Farag [4]:

	u(k-4)						
y(k-1)	NL	NM	NS	ZE	PS	PM	PL
NL				***NS***	***NS***	***NM***	***NL***
NM		***ZE***	***NVS***	***NVS***	***NVS***	***NS***	***NS***
NS		***ZE***	***ZE***	***NVS***	***NVS***	***NVS***	
ZE		***PVS***	***ZE***	***ZE***	***NVS***	***NVS***	
PS	***PS***	***PVS***	***PVS***	***ZE***	***ZE***	***ZE***	
PM	***PS***	***PS***	***PS***	***PVS***	***PVS***	***PVS***	
PL	***PL***	***PM***	***PM***	***PM***	***PM***		

Table 4. Linguistic variable model for the gas furnace

The linguistic model in Table 4 divides the input and output spaces into 7 each: NL, NM, NS, ZE, PS, PM, PL corresponding to negative large, negative medium, negative small, zero, positive small, positive medium, and positive large. Note that negative here does not necessarily signify negative in numeric terms, this is just relative to the zero linguistic variable. Comparing the table above with the linear model equation (5) above, we notice that y(k) decreases as u(k-4) increases for a constant y(k-1) - suggesting a negative u(k-4) coefficient. However, y(k) increases as y(k-1) increases which indicate that the regression coefficient is positive.

V. DISCUSSION

A benchmark suite for the evaluation of nonlinear system identification methods is needed. The gas furnace data set is widely used as a benchmark, which seems to require careful consideration as to what can be concluded from testing nonlinear modeling techniques on linear data. It may be more appropriate to use highly non-linear data sets in addition to or instead of the gas furnace data set. Non-linear data can be easily generated from equations.
For example, Chen and Billings [3] use the equation

$$y(k) = (0.8 - 0.5\,e^{-y^2(k-1)})\,y(k-1) - (0.3 + 0.9\,e^{-y^2(k-1)})\,y(k-2) + 0.1\sin(\pi y(k-1)) + n(k)$$

with the initial values set to y(-1) =0.1, y(-2) =0.1, and the noise term n(k) = Gaussian(0,0.01).
Another example from Narendra et al. [8] uses a second order difference equation in the form:

$$y(k) = \frac{y(k-1)\,y(k-2)[\,y(k-1) + 2.5]}{1 + y^2(k-1) + y^2(k-2)} + u(k)$$

VI. CONCLUSIONS

Benchmarks are widely used for empirical validation and verification in intelligent system identification. Inaccurate reporting of benchmark results may cause confusion concerning the suitability of particular system identification techniques to given modeling tasks. Results for the gas furnace data had suffered from this problem due to two results [10]. The revised results are presented here.

In general, linear modeling techniques perform well on linear data. It is therefore surprising to see nonlinear models outperform linear models for a linear dataset such as the gas furnace dataset. This apparent anomaly is analyzed more closely here. By choosing the model inputs to match those used in the more successful nonlinear techniques, a linear model that outperforms such models is presented. The choice of input is therefore of great importance.

The mean squared error minimization should never be the only performance measure for a fuzzy process identification benchmark. The ability to capture the underlying linear trend and the ability to generalize the actual trend in the data instead of following closely the noise are also important performance issues that may be deduced from testing a nonlinear technique on linear data.

REFERENCES

[1] S. Billings and Q. Zhu, Model validation tests for multivariate nonlinear models including neural networks, *Int. J. Control*, 62(4), 749-766, 1995.

[2] G. E. P. Box and G.M. Jenkins, *Time Series Analysis: Forecasting and Control*, 2nd ed. San Francisco: Holden-Day, 1970.

[3] S. Chen and S. Billings, Neural Networks for Nonlinear Dynamic System Modeling and Identification, *Int. J. Control*, 56, 1992, 319-346.

[4] W. Farag, V. Quintana, and G. Lambert-Torres, A Genetic-based Neuro-fuzzy Approach for Modeling and Control of Dynamical Systems, *IEEE Trans. on Neural Nets*, 9 (5), 1998, 756-767.

[5] E. Kim, M. Park, and M. Park, A New Approach to Fuzzy Modeling, *IEEE Trans. on Fuzzy Systems*, 5(3), 1997, 328-337.

[6] C.T. Lin and C. S. G. Lee, Neural-Network-Based Fuzzy Logic Control and Decision System, *IEEE Trans. on Computers,* 40(12), 1991, 1320-36.

[7] E. Mamdani, Advances in the linguistic synthesis of fuzzy controllers, *Int. J. Man-Machine Studies*, 8, 1976, 669-678.

[8] K.S. Narendra, and K. Parthasarathy, Identification and Control of Dynamical Systems Using Neural Networks, *IEEE Trans. Neural Net*.,1(1), 1990.

[9] W. Pedrycz, An identification algorithm in fuzzy relational systems, *Fuzzy Sets Syst.*. 13, 1984,153-167.

[10] M. Sugeno and K. Tanaka, Successive Identification of a Fuzzy Model and its Application to Prediction of a Complex System, *Fuzzy Sets Syst*., 42, 1991, 315-334.

[11] M. Sugeno, and T. Yasukawa, A Fuzzy-Logic-Based Approach to Qualitative Modeling, *IEEE Trans. on Fuzzy Systems*, 1(1), 1993, 7-31.

[12] T. Takagi and M. Sugeno, Fuzzy Identification of Systems and Its Applications to Modeling and Control, *IEEE Transactions on Systems, Man, and Cybernetics*, 15(1), 1985.

[13] R. M. Tong, Synthesis of Fuzzy Models For Industrial Processes: Some Recent Results, *Int. J. General Systems*, 4, 1978, 143-162.

[14] L. Wang and R. Langari, Complex Systems Modeling via Fuzzy Logic, *IEEE Trans. on Syst., Man, and Cybernetics*, 26(1), 1996, 100-106.

[15] C.W. Xu and Y.Z. Lu, Fuzzy model Identification and self-learning for dynamic systems, *IEEE Trans. Syst. Man and Cyber*., 17, 1987, 683-689.