TIME SERIES PREDICTION USING FUZZY INDUCTIVE REASONING A CASE STUDY

Josefina López
Institut de Cibernètica
Univ. Pol. de Catalunya
Diagonal 647, 2a. planta
Barcelona 08028, Spain
Lopez@IC.UPC.Es

Gabriela Cembrano
Institut de Cibernètica
Univ. Pol. de Catalunya
Diagonal 647, 2a. planta
Barcelona 08028, Spain
Cembrano@IC.UPC.Es

François E. Cellier
Elect. & Comp. Engr. Dept.
University of Arizona
Tucson, AZ 85721
U.S.A.
Cellier@ECE.Arizona.Edu

Abstract

This paper presents the application of fuzzy inductive reasoning (FIR) to time–series analysis and forecasting. This methodology had previously been applied to modeling and control of dynamic input–output systems [de Albornoz 96], [Mugica 95], [Nebot 94]. The research effort discussed in this paper presents a first attempt at assessing the suitability of this qualitative modeling methodology for forecasting time–series, i.e., for predicting the future development of signals on the basis of their own past, without identifying the systems that produce these signals. The performance of FIR in time–series forecasting is compared with Box–Jenkins methods and neural networks in a case study.

Keywords: Qualitative Simulation, Forecasting, Fuzzy Inductive Reasoning, Time Series, Water Demand Prediction.

INTRODUCTION

Given a stable physical plant. Its eigendynamics die naturally out due to the stability properties of the plant. Consequently, the outputs of the plant are predominantly driven by its inputs, and not by its eigendynamics. Mathematical models of such systems can either be obtained deductively from first principles, or inductively by observing the inputs and outputs of the plant over a given period of time. In both cases, the mathematical model obtained is expected to be characterised by the same stability properties as the physical plant it represents. Hence it is possible to make predictions of future outputs over an essentially unlimited time horizon as long as the system characteristics remain the same and as long as the future input streams are precisely known. Errors made in the prediction will not accumulate, but get dissipated away as a consequence of the stability properties of the model.

Time series analysis is fundamentally different from the mathematical modeling and simulation of systems in several respects. Although a time series can be interpreted as an output of a system, it is, by definition, an output of an *unknown* system. Neither the system characteristics nor the input functions are known, and consequently, time–series analysis must content itself with estimating future output values by means of extrapolation from their own past. The more benign (i.e., stable) the unknown system that produces the time series, the less likely this will work. In the extreme case, one could imagine a system with a transfer function of 1.0 driven by white noise. Clearly, any attempt at performing a time–series analysis on the output of this system must be futile, whereas a systems analysis would work perfectly, since the input functions are expected to be known in that case.

The scientific community has been interested in the analysis of time series for a long time. Time series are important when studying the behavior of a system that is not completely understood, such as the time–varying intensity of a star, or the stock market. Clearly, one should not expect similarly accurate results when forecasting a time series as when simulating a known system with known inputs. It is also important to recognize that the time horizon of a meaningful prediction will, in this case, usually be limited, and in fact, may be rather short. Furthermore, a successful prediction of a time series depends on the characteristics of the time series. It is to be expected that a stationary or quasi–stationary process can be predicted better and over a longer time horizon than a non–stationary process.

In this paper, a new technique for the analysis of time series shall be presented, a technique that hitherto has only been used for qualitative modeling and simulation of systems. The performance of this new methodology, called *Fuzzy Inductive Reasoning (FIR)* shall be compared with that of a Box–Jenkins approach as well as with a neural network solution. The case study that is presented in this paper is limited to a single time series, describing the water demand of a section of the city of Barcelona. This time series has been selected, because earlier attempts at predicting this time series have previously been reported in the literature.

Several methodologies have been used in the past to analyze and forecast time series. The methodologies that were reported in the literature range from linear modeling [Chatfield 89], [Priestley 81], [Box 76] to more complex non–linear techniques [Tong 90], [Volterra 59]. Yet more recently, neural networks and state–space reconstruction have successfully been applied to the analysis of time series [Casdagli 92], [Weigend 90]. Learning methods for systems modeling, such as neural networks and FIR, derive implicit models of the systems in question, by extracting regularities in the input/output behavior from a set of training examples. These same methods can also be applied to the analysis of time series. In this case, they exploit regularities within the observed output patterns.

FUZZY INDUCTIVE REASONING

Fuzzy inductive reasoning is a modeling and simulation methodology that generates a qualitative input/output model of a system by finding the best possible fuzzy finite state machine between discretized (fuzzified) input and output states of the system. The methodology is composed of the following main functions:

- *Fuzzification.* The process of converting quantitative variables into qualitative triples is called *recoding*. The first component of the triple is the class value, the second is the fuzzy membership function value, and the third is the side value [Cellier 96]. The process of recoding is applied to each observed variable (trajectory) separately. The recoded qualitative episodical behavior is stored in three matrices, one containing the class values, the second storing the membership function values, and the third keeping the side values. Each column of these matrices represents one of the observed variables, and each row represents on recorded state. The trajectory behavior can thus be separated into a set of input trajectories, u_i, concatenated from the right with a set of output trajectories, y_i, as shown in the following example (three inputs and one output):

$$\begin{array}{l} time \\ 0.0 \\ \delta t \\ 2 \cdot \delta t \\ 3 \cdot \delta t \\ \vdots \\ (n_{rec}-1) \cdot \delta t \end{array} \begin{array}{c} \begin{array}{cccc} u_1 & u_2 & u_3 & y_1 \end{array} \\ \begin{pmatrix} \dots & \dots & \dots & \dots \\ \dots & \dots & \dots & \dots \\ \dots & \dots & \dots & \dots \\ \dots & \dots & \dots & \dots \\ \vdots & \vdots & \vdots & \vdots \\ \dots & \dots & \dots & \dots \end{pmatrix} \end{array} \quad (1)$$

- *Qualitative Modeling.* Once the quantitative trajectory behavior has been recoded into a qualitative episodical behavior, the process of modeling consists of finding finite automata relations between the recoded variables that make the resulting state transition matrices as deterministic as possible. Such a relation is called a mask. An example of a mask might be:

$$\begin{array}{c} {}_t\backslash^x \\ t-2\delta t \\ t-\delta t \\ t \end{array} \begin{array}{c} \begin{array}{cccc} u_1 & u_2 & u_3 & y_1 \end{array} \\ \begin{pmatrix} 0 & 0 & 0 & -1 \\ 0 & -2 & -3 & 0 \\ -4 & 0 & 0 & +1 \end{pmatrix} \end{array} \quad (2)$$

The negative elements in this matrix denote inputs of the qualitative functional relationship, so–called m–inputs. The above example has four m–inputs. The positive value represents the m–output. A mask denotes a dynamic relationship between qualitative variables. A mask has the same number of columns as the episodical behavior to which it is applied, and it has a certain number of rows. The number of rows of the mask matrix is called the *depth* of the mask. The mask can be used to flatten a dynamic relationship out into a static relationship. A mask candidate matrix is an ensemble of all possible masks, from which the best one is chosen by a mechanism of exhaustive search. The mask candidate matrix contains -1 elements where the mask has a potential m–input, it contains a $+1$ element where the mask has its m–output, and it contains 0 elements to denote forbidden connections. Thus, the mask candidate matrix for the previous four–variable example will be:

$$
\begin{array}{c|cccc}
t\backslash^{x} & u_1 & u_2 & u_3 & y_1 \\
t-2\delta t & -1 & -1 & -1 & -1 \\
t-\delta t & -1 & -1 & -1 & -1 \\
t & -1 & -1 & -1 & +1
\end{array} \qquad (3)
$$

Each of the possible masks is compared to the others with respect to its potential merit. The optimality of the mask is evaluated with respect to the maximization of its forecasting power. The Shannon entropy measure is used to determine the uncertainty associated with the forecasting of the desired output state, for given feasible input states [Cellier 96].

- *Qualitative Simulation.* Once the optimal mask has been determined, it can be applied to the given raw data matrix resulting in an input/output matrix. Since the input/output matrix contains functional relationships within single rows, the rows of the input/output matrix can be sorted in alphanumerical order. The result of this operation is called the *input/output behavior matrix* of the system. The input/output behavior matrix is a finite state machine. For each combination of input values, it shows which output is most likely to be observed [Cellier 96].

- *Defuzzification.* This is the inverse function of the recoding process. In fuzzy inductive reasoning, it is called *regeneration* [Cellier 96].

For a more detailed information on this methodology, the reader is referred to [Nebot 94], [Mugica 95], [Cellier 96].

TIME SERIES FORECASTING

In the previous section, it was shown how FIR can be applied to identifying a model of a system. Each FIR model is characterized by a set of mask inputs, the so–called m–inputs that best determine the output to be predicted, i.e., the so–called m–output. Whereas the m–output is always an output of the system, the m–inputs can be chosen among the input variables and older values of the outputs. Hence the same methodology can also be used for modeling time series. In this case, there aren't any recorded input variables, and all m–inputs must be chosen among earlier values of the recorded outputs.

In the simplest (yet most difficult to predict) case, only a single variable, the time series, has been observed, the future values of which are to be predicted on the basis of their own past. In this case, the mask candidate matrix will have n rows and one column. The mask candidate vector contains -1 elements where the mask has potential inputs, a $+1$ where the mask has its output (always at the bottom of the vector), and it contains 0 elements to denote forbidden connections.

In order to decide the depth of the mask, the autocorrelation function is used. The mask should be made as deep as is necessary to capture the significant autocorrelation coefficients.

WATER DEMAND PREDICTION

As a case study, FIR techniques have been applied to the modeling and forecasting of the series of daily water demand in an important section of Barcelona. The data contain daily water demand of roughly two years. The reason for the choice is the availability of two other models developed previously, which will be useful for comparison [Quevedo 88], [Griñó 92]. A plot of the series is shown in Figure 1. The process is approximately stationary and its variance is roughly constant, as can be seen in the figure. It is a quasi–stationary process with weekly and seasonal quasi–periodic characteristics. A seven–day cycle can be observed on examination of the series. During the weekends, less water is being used than during week days. In addition, it can be noticed that the water demand drops during the month of August as a consequence of the summer holiday period. Shorter drop periods can also be observed during the Easter week and other public holidays.

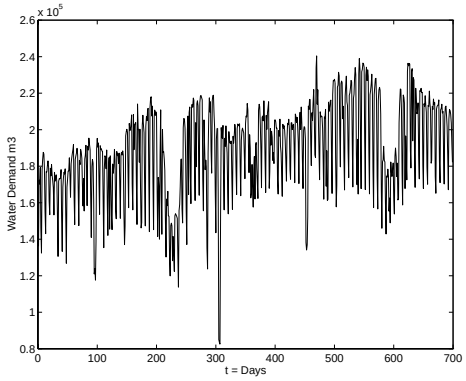

Figure 1: Daily water demand time series January 1985 - November 1986

Figure 2 depicts the autocorrelation function. It clearly shows the weekly cycle. It is quite evident

that the water demand on any given day is strongly correlated with the water demand seven days earlier.

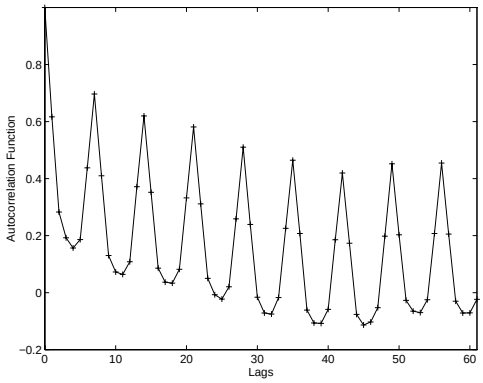

Figure 2: Autocorrelation function of daily demand series

SAPS–II, our implementation of the FIR methodology, was used as a qualitative simulator. SAPS–II is a toolbox of Matlab. The data were divided into two different sets: 570 days (from January 1985 to 24 July 1986) for the training, and 128 days (from 25 July to 29 November 1986) for the test.

The steps for the qualitative simulation were the following:

- *Fuzzification.* The quantitative data were recoded into three classes each. The series does not require any data preprocessing before the modeling stage.

- *Qualitative Modeling.* The autocorrelation function (Figure 2) shows the significance of the periods 7, 14, 21. A mask candidate vector of depth 15 was proposed to SAPS–II:

$$\begin{array}{ll} & u_1 \\ t-14\delta t & \begin{pmatrix} -1 \\ -1 \\ \cdots \\ -1 \\ \cdots \\ -1 \\ -1 \\ -1 \\ +1 \end{pmatrix} \\ t-13\delta t & \\ \cdots & \\ t-7\delta t & \\ \cdots & \\ t-2\delta t & \\ t-\delta t & \\ t & \end{array} \qquad (4)$$

The following optimal mask was found by SAPS–II:

$$\begin{array}{ll} & u_1 \\ t-14\delta t & \begin{pmatrix} -1 \\ 0 \\ 0 \\ -2 \\ 0 \\ 0 \\ -3 \\ +1 \end{pmatrix} \\ t-13\delta t & \\ \cdots & \\ t-7\delta t & \\ \cdots & \\ t-2\delta t & \\ t-\delta t & \\ t & \end{array} \qquad (5)$$

The optimal mask shows the 7–day weekly dependency. The mask found by SAPS–II is perfectly reasonable. Clearly, the last day's demand should be taken into consideration, as well as the demand of one and two weeks back. This is precisely what SAPS–II recommended. This is consistent with the structure used by the ARIMA model of this time series [Quevedo 88].

- *Qualitative Simulation.* Figure 3 shows the forecasting obtained from this qualitative model. The dashed line represents the predicted values, and the solid line shows the real values. The look–ahead period was one day, i.e., FIR was asked to predict todays water demand on the basis of the measurements taken yesterday, one week back, and two weeks back. None of the predictions ever relied on previously predicted data. This is exactly what also both of the previous models did. The experimental design was driven by the requirements of the city government who were not interested in longer–range predictions. FIR has been able to predict the behavior of the time series quite well. It learned the 7–day cycle, and it was able to predict also the slower frequency characteristics reasonably well. The larger errors during the month of August are caused by data deprivation. There is clearly a lack of relevant data in the experience data base, i.e., the data that were used for training the model. Other methodologies used preprocessing or intervention analysis to take care of situations, where known external events cause effects on the values of the series. No such techniques have yet been used in the FIR model, although they may prove useful in the future. The results are, nevertheless, satisfactory in that the errors are largely redressed in a few days, i.e. the model adapts to the new situation quickly. This is true both for sudden drops and for sudden increases in the water consumption.

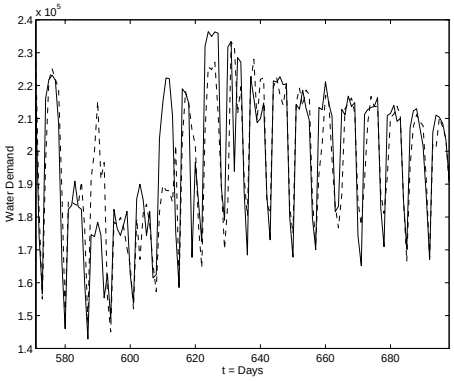

Figure 3: Water Demand FIR prediction from 25 July to 29 November 1986

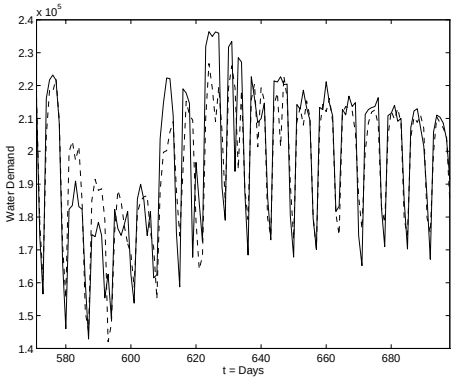

Figure 4: Water Demand NN prediction from 25 July to 29 November 1986

COMPARISON OF FIR WITH OTHER TECHNIQUES

The same water demand time series had previously been studied using two different methodologies, Neural Networks [Griñó 92] and Box–Jenkins [Quevedo 88].

For the application of the Box–Jenkins method [Quevedo 88], the water demand series was deseasonalized by differencing at lag 7, i.e., by using:

$$dy(t) = y(t) - y(t-7) \tag{6}$$

as input to the ARIMA model.

SAPS–II, on the other hand found a qualitative model in a totally automated fashion. The model has three inputs and one output. The optimal mask suggests the most significant lags. The lags are the same as the significant coefficients in the fitted ARIMA(7,7,7) model [Quevedo 88].

[Griñó 92] suggested a neural network with 15 inputs nodes, 20 hidden nodes, and 1 output node. Figure 4 shows the prediction of this model.

The mean percentage error obtained with the ARIMA model, the neural network model, and the FIR model were 4.5%, 4.2%, and 4.6%, respectively. It is important to remark that, in the ARIMA model, intervention analysis was performed, and the ARIMA model was augmented by adding deterministic components to account for seasonal variations. This required a much more tedious analysis of the series. In the neural network model, the number of layers and the number of neurons per layer had to be chosen very carefully, and the training of the neural network was a rather slow process.

The authors believe that the results obtained with SAPS–II, with no data preprocessing at all and with no manual intervention of the designer, show an excellent potential of this methodology for the application of time–series predictions.

CONCLUSIONS

A first attempt at assessing the potential of FIR for time–series forecasting has been presented. The results obtained for thegiven case study are very encouraging in that the forecasting redresses errors quickly, and the prediction error obtained is comparable to those of ARIMA modeling and neural networks. An asset of SAPS–II, as compared with the other two methodologies, is that it generates the model quickly and in a completely automated fashion.

At this stage of the research, two main conclusions can be drawn that would generalize to the applicability of FIR to predict other time series of quasi–stationary processes:

1. SAPS will efficiently and effectively detect the significant lags in the series that lead to an optimal mask.
2. The FIR model of a time series is mostly an autoregressive model, so that the errors are assumed to be uncorrelated. Therefore it will work well on series that correspond to deterministic, or autoregressive stochastic processes.

Other series corresponding to largely stochastic

processes where the errors are correlated are modeled correctly inasmuch as correlation of the errors can be captured with a finite number of autoregressive terms. It is important to remark that this is very often the case in time–series analyses. However, a further step in the research is underway in order to endow SAPS with capabilities to treat stochastic series where the errors are significantly correlated.

References

[Box 76] Box, G.E.P., and F.M. Jenkins. (1976), *Time Series Analysis: Forecasting and Control*, 2nd ed. Oakland, CA: Holden-Day.

[Casdagli 92] Casdagli, M., and S. Eubank. (1992), "Nonlinear Modeling and Forecasting," Santa Fe Institute Studies in the Sciences of Complexity, Proc. Vol. XII. Redwood City: Addison-Wesley.

[Cellier 96] Cellier, F.E., A. Nebot, F. Mugica, and A. de Albornoz. (1996), "Combined Qualitative /Quantitative Simulation Models of Continuous–Time Processes Using Fuzzy Inductive Reasoning Techniques," in *Intl. J. Gen. Syst.*, **24**(1–2):95-116.

[Chatfield 89] Chatfield, C. (1989), *The Analysis of Time Series*, 4th ed. London: Chapman and Hall.

[de Albornoz 96] de Albornoz, A. (1996), "Inductive Reasoning and Reconstruction Analysis: Two Complementary Tools for Qualitative Fault Monitoring and Decision Support in Large–Scale Systems," Ph.D. Thesis, Universitat Politècnica de Catalunya.

[Griñó 92] Griñó, R. (1992), "Neural Networks for Univariate Time Series Forecasting and Their Application to Water Demand Prediction," *Neural Network World*, pp. 437–450.

[Mugica 95] Mugica, F. (1995), "Diseño Sistemático de Controladores Difusos Usando Razonamiento Inductivo," Ph.D. Thesis, Universitat Politècnica de Catalunya.

[Nebot 94] Nebot, A. (1994), "Qualitative Modeling and Simulation of Biomedical Systems using Fuzzy Inductive Reasoning," Ph.D. Thesis, Universitat Politècnica de Catalunya.

[Priestley 81] Priestley, M.B. (1988). *Nonlinear and Non–Stationary Time Series*. New York: Academic Press.

[Quevedo 88] Quevedo, J., G. Cembrano, A. Valls, and J. Serra. (1988) "Time Series Modeling of Water Demand — A Study on Short–Term and Long–Term Predictions," *Water Demand Prediction and Allocation*. p. 268. Sec. V. ed. Byron Coulbeck and Chun-Hou Orr. *Computer Application in Water Supply*.

[Tong 90] Tong, H. (1990), *Nonlinear Time Series Analysis: A Dynamical System Approach*, Oxford: Oxford University Press.

[Volterra 59] Volterra, V. (1959), *Theory of Functionals and of Integral and Integro–Differential Equations*, New York: Dover.

[Weigend 90] Weigend, A.S., B.A. Huberman, and D.E. Rumelhart. (1990) "Predicting the Future: A Connectionist Approach,", *Intl. J. Neur. Sys.*, **1**:193-209.