DEALING WITH INCOMPLETE DATA RECORDS IN QUALITATIVE MODELING AND SIMULATION OF BIOMEDICAL SYSTEMS

Angela Nebot*
Institut de Cibernètica
Universitat Politècnica de Catalunya
Diagonal 647, 2na. planta
Barcelona 08028, Spain
nebot@ic.upc.es

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

ABSTRACT

In the biomedical domain, it is very common to find that the internal structure of the systems under investigation are totally or partially unknown, making it impossible to use analytical models. Therefore, qualitative modeling and simulation techniques with their inherent tolerance for uncertainty and ambiguity provide an excellent platform for the analysis of biomedical systems that may be difficult to model in a more precise fashion. Moreover, even where quantitative models are available, qualitative models may constitute an important complement to the more classical quantitative models.

One of the major problems in biomedical qualitative modeling is the lack of information. Inductive, pattern–based modeling techniques are extremely data hungry. It is therefore essential for behavioral qualitative methodologies to have available a large amount of rich data to work with. Unfortunately, in biomedical applications, this is hardly ever the case.

The lack of information may have several different causes, all of them related to acquisition difficulties. The problems are further amplified when the data records obtained from medical experiments are incomplete.

In this paper, a technique called *missing data option* is proposed that allows to work with incomplete medical data records. This technique represents an enhancement to a previously introduced qualitative modeling and simulation methodology entitled *fuzzy inductive reasoning*.

INTRODUCTION

Even when sufficient and sufficiently rich data are available, incomplete data sets can still pose difficult problems to the modeler, and this situation is unfortunately all too common in the biomedical domain.

Biomedical data records are notorious for being incomplete. A patient on a heart monitor is routinely taken off the monitor while the nurse is cleaning him or her. A particular instrument may exist only in one copy. Although the instrument is in use by one patient, it is temporarily removed in order to give it to another patient who needs it more urgently. Signal detectors that are taped to the patients' body often fall off during the night. The recording device (e.g. a tape cassette) is full and is not replaced for a while. There are dozens of circumstances that can produce gaps of information for one, several, or all of the parameters. Qualitative methodologies that cannot deal with missing data values are therefore quite useless for biomedical applications.

In this paper, a missing data feature is presented that enables the researcher to work with sets of incomplete data, and extract as much information from them as they contain. The feature makes it possible to convert incomplete quantitative data sets to reduced qualitative data sets in order to derive the best possible qualitative model for prediction of future system behavior. This feature has been developed for use within fuzzy inductive reasoning.

The inductive reasoning methodology had originally been developed by G. Klir (Klir 1985) as a tool for general system analysis, to study the conceptual modes of behavior of systems. One implementation of this methodology is SAPS–II (Cellier and Yandell 1987). Fuzzy measures were introduced into the methodology independently by (Klir and Folger 1988; Klir 1989; Wang and Klir 1992) and by (Li and Cellier 1990). Even more recently, SAPS–II has been advocated as a tool for qualitatively studying the behavior of highly complex non–linear technical systems (Cellier et al. 1992; de Albornoz and Cellier 1993a, 1993b) as well as biomedical systems (Nebot et al. 1993).

FUZZY INDUCTIVE REASONING

The fuzzy inductive reasoning (FIR) methodology is composed of four basic functions: *fuzzy recoding* (fuzzification), *fuzzy optimization* (qualitative modeling), *fuzzy forecasting* (qualitative simulation), and *fuzzy regeneration* (defuzzification).

*Supported by Consejo Interministerial de Ciencia y Tecnología (CICYT), under project TIC93-0447.

Recoding denotes the process of converting a quantitative variable to a qualitative variable. In general, some information is lost in the process of recoding, but fuzzy recoding avoids this problem. Each quantitative value of a variable is recoded into a qualitative triple, whereby the first component is the class value, the second component is the fuzzy membership function, and the third component is the side value. The side value indicates whether the qualitative value is to the left or to the right of the maximum of the fuzzy membership function.

In SAPS–II, qualitative knowledge about a system under study is represented by a large data matrix with one row per sample, and as many columns as there are variables in the system to be modeled. Thus, each row represents one data record, and each column represents one trajectory. This data matrix is called the *raw data matrix*. For example:

$$\begin{array}{lc} time & \begin{array}{ccccc} u_1 & u_2 & u_3 & y_1 & y_2 \end{array} \\ \begin{array}{l} 0.0 \\ \delta t \\ 2 \cdot \delta t \\ 3 \cdot \delta t \\ \vdots \\ (n_{rec}-1) \cdot \delta t \end{array} & \left(\begin{array}{ccccc} \ldots & \ldots & \ldots & \ldots & \ldots \\ \ldots & \ldots & \ldots & \ldots & \ldots \\ \ldots & \ldots & \ldots & \ldots & \ldots \\ \ldots & \ldots & \ldots & \ldots & \ldots \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ \ldots & \ldots & \ldots & \ldots & \ldots \end{array} \right) \end{array} \qquad (1)$$

where u_i are the inputs, y_i the outputs, n_{rec} is the number of data records, and δt is the sampling interval.

From the raw data matrix, SAPS–II is able to find a model that represents the system. In the process of modeling, finite automata relations are to be found between the recoded variables that make the resulting state transition matrices as deterministic as possible. If such a relationship has been found for every output variable, the behavior of the system can be forecast by iterating through the state transition matrices. The more deterministic the state transition matrices are, the better the certainty that the future behavior will be predicted correctly. Such a finite automata relation could take the form:

$$y_1(t) = \tilde{\mathrm{f}}(y_3(t-2\delta t), u_2(t-\delta t), y_1(t-\delta t), u_1(t)) \qquad (2)$$

Equation (**??**) can be represented in a matrix form as follows:

$$\begin{array}{lc} {}_t\backslash^x & \begin{array}{ccccc} u_1 & u_2 & y_1 & y_2 & y_3 \end{array} \\ \begin{array}{l} t-2\delta t \\ t-\delta t \\ t \end{array} & \left(\begin{array}{ccccc} 0 & 0 & 0 & 0 & -1 \\ 0 & -2 & -3 & 0 & 0 \\ -4 & 0 & +1 & 0 & 0 \end{array} \right) \end{array} \qquad (3)$$

The negative elements in this matrix denote inputs of our qualitative functional relationship. The positive value is the output of the qualitative relationship. In inductive reasoning, such a representation is called a *mask*. A mask denotes a dynamic relationship between qualitative variables. The best possible mask is found in a process of optimization using the *fuzzy optimal mask function* of SAPS–II.

Once the optimal mask has been determined, it can be applied to the given raw data matrix, flattening a dynamic relationship out into a static relationship. The mask can be shifted over the episodical behavior, picking out the selected inputs and outputs, and writing them together in one row. The result of this operation is called the *input/output model* of the system. The entries (rows) of the input/output model can then be sorted lexicographically. The result of this sorting operation is a *behavior matrix*. The behavior matrix is a finite state machine. For each combination of input values, it shows which output is most likely to be observed.

Forecasting is now a straightforward procedure. The mask is simply shifted further down beyond the end of the raw data matrix, future inputs are read out from the mask, and the behavior matrix is used to determine the future output, which can then be copied back into the raw data matrix. In fuzzy forecasting, it is essential that, together with the qualitative output, also a fuzzy membership value and a side value are forecast. Thus, fuzzy forecasting predicts an entire qualitative triple from which a quantitative variable can be regenerated whenever needed.

In fuzzy forecasting, the membership and side functions of the new input are compared with those of all previous recordings of the same qualitative input contained in the behavior matrix. The five inputs with the most similar membership and side functions are identified, and the new output value is computed as a weighted average of the outputs of the five nearest neighbors (Mugica and Cellier 1993).

The regeneration process is the inverse process of recoding, converting a qualitative triple to a quantitative value.

MISSING DATA OPTION

How has this methodology been enhanced to allow to work with incomplete data records? First of all, a global variable has been introduced that allows the user to define how missing data points are marked in the recorded data. Medical data bases usually denote missing data by a physically impossible value, such as -999. SAPS–II allows the user to declare that e.g. -999 is supposed to mark missing data, and in this way, the data base can be accessed directly by SAPS. This feature thus permits the user to work directly with the data recorded at the hospital without first preprocessing them.

The fuzzy recoding function computes a qualitative triple for each quantitative data entry in the measurement data matrix, as it was done in the previous version. When a missing data value is encountered, the class value of the corresponding qualitative triple is not computed, keeping the missing data marker as the class value. The membership function is set to one, and the side value is set to zero. In this manner,

the place of the missing data elements within the *raw data matrix* (the recoded version of the measurement data matrix) is known.

In order to identify the model that represents the system, the fuzzy optimal mask function is used. It performs an optimization that finds the best model (mask) by means of a mechanism of exhaustive search through all possible models (masks). 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 using an uncertainty measure and a complexity measure.

In order to evaluate the quality of a mask, it is necessary to have available the input/output matrix. As was mentioned earlier, the input/output matrix is obtained by shifting the mask over the episodical behavior, picking out the selected inputs and outputs, and writing them together in one row. Therefore, it is possible that the input/output matrix contains missing elements, that might get used during the quality evaluation. Therefore, it is essential to eliminate from the input/output matrix all data records that are contaminated by missing values. The optimal mask function has been modified to eliminate all contaminated records from the input/output matrix.

When the optimal model that decribes the system is found, it is used to forecast the future outputs of the system. The new forecasting function goes through the input/output matrix deleting all rows that are contaminated by missing values. This corresponds to the elimination of illegal rules from a rule base. Once the input/output matrix is free of contaminated data records, the behavior matrix is computed from it, and the class, membership and side values of the current output can be forecast.

It is important to distinguish between two types of past data: (i) the *history data* that is being used to recognize similar behavioral patterns in the past, and (ii) the *immediate past data* that is used by the recursion of the finite state machine. Whereas missing values in the history data base are comparatively harmless since contaminated records can simply be eliminated, missing data values among the immediate past data are much more critical. If the current forecast requires an immediate past value that is missing, the routine needs to backtrack to first come up with a prediction of that value.

The modifications needed to upgrade the *regenerate function* were trivial. Contaminated qualitative triples are simply converted to missing quantitative values.

LIMITATIONS OF PREDICTABILITY

How many missing data values can be tolerated before the forecasting power of the fuzzy inductive reasoner deteriorates? This question is difficult to answer in a precise quantitative fashion. The missing data feature is implemented in such a way that missing data do not affect the forecasting power of the model per se. It is the lack of training data that affects the forecasting power. Thus, as long as the training data set is sufficiently rich, missing data will not affect the forecasting very much, i.e., missing data can be compensated for by redundancy in data records. A significant degradation of the forecasting power will only be experienced when the number of missing data records is so large that the richness of the training data set no longer suffices to compensate for the loss, or if the missing data are in some way systematic. For example, if the missing data always occur when one of the variables is at its peak value, then obviously, the training data set no longer contains any information as to how the model should behave when that variable is at its peak value, and will be unable to forecast appropriately in that situation. After all, forecasting in fuzzy inductive reasoning is only a smart way of remembering (associating) similar past behavior with the current situation. Two examples are used in the paper to clarify these statements.

Biomedical Application

The biomedical system presented in this example is a subsystem of the human central nervous system, more specifically, it concerns the *venous tone controller*. The most important characteristics of this system can be captured by a SISO model where the input signal is the carotid pressure and the output signal represents the control of the venous tone, i.e., a signal that influences the compliance of the vene that finally dictates the blood pressure in the vene itself.

Figure 1 shows the input and output signals that are used to identify the qualitative model. In this data set, there are no missing values.

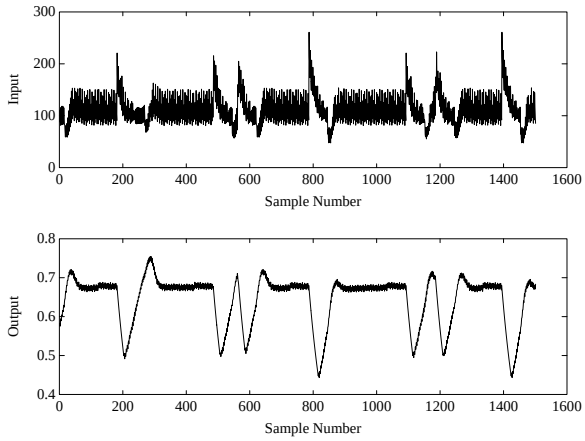

Figure 1: Input/Output Data for Model Identification

The time distance between two neighboring sampling points is 0.24 seconds.

The best qualitative model (the optimal mask) for representing the venous tone controller is the following:

$$
\begin{array}{c|cc}
t \backslash x & u & y \\
t-2\delta t & -1 & 0 \\
t-\delta t & 0 & -2 \\
t & -3 & 1
\end{array}
\qquad (4)
$$

This masks indicates that the control output at the current time depends on its own past one time step back, and on the carotid pressure at the current time as well as two time steps back.

The forecasting results obtained from this model are presented in figure 2. The solid line represents the measurement data, whereas the dashed line denotes the forecast. As can be seen, they are very close. The fast time constant is captured with high accuracy, and also the slow time constant is captured quite well.

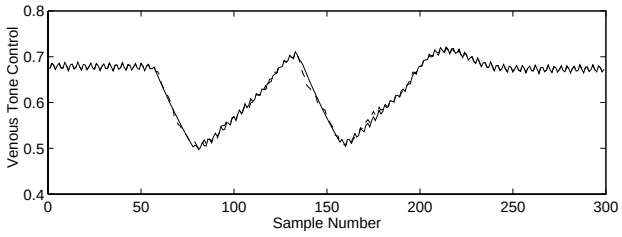

Figure 2: Forecast of the Biomedical System without Missing Data

It is evident that the identified qualitative model represents the system very well. Note that the data used in the model identification process did not contain any missing values.

A series of tests have been carried out modifying the number and position of missing data entries. A few of them are presented in this paper.

Adjacent Missing Data

In this test, different amounts of missing values have been inserted in the identification data set adjacent to each other.

In a first test, 10% of the identification data are declared missing, namely those from sampling point 501 to 650. As can be seen from figure 1, this portion of the data is the one that is most similar to the curve to be forecast. Therefore, although the amount of missing data is not very high, its loss is significant for the prediction, as can be seen from figure 3.

The prediction still works, but its quality is definitely reduced.

In a second test, the identification data set contains a gap of 40% missing values from sampling point 800 to 1400. Here, a different mask is obtained as the best model to represent the system:

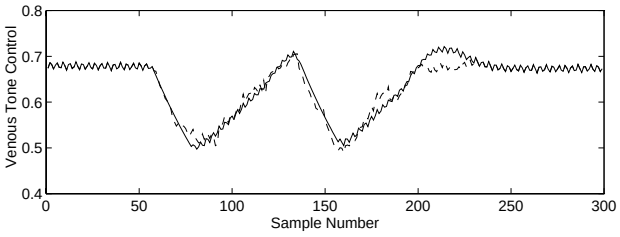

Figure 3: Forecast of the Biomedical System with 10% of Adjacent Data Missing

$$
\begin{array}{c|cc}
t \backslash x & u & y \\
t-2\delta t & -1 & 0 \\
t-\delta t & -2 & -3 \\
t & -4 & 1
\end{array}
\qquad (5)
$$

The set of missing data chosen in this test is not similar to the curve to be predicted (cf. figure 1), and consequently, it is not essential for the model to capture them. Therefore, the prediction is quite good and definitely better than in the previous case in spite of the much larger chunk of missing data, as shown in figure 4.

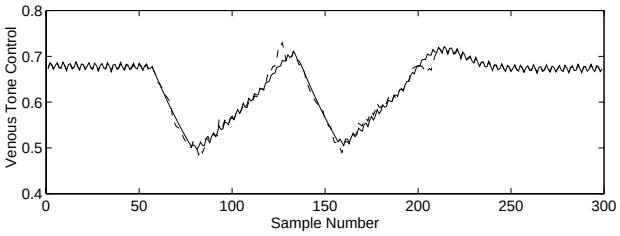

Figure 4: Forecast of the Biomedical System with 40% of Adjacent Data Missing

Scattered Missing Data

In this test, 40% data values are missing like in the previous one. However, they are scattered throughout the data file. Groups of 50 missing elements are distributed arbitrarily along the identification data toggling between input stream and output stream: from sample 51 to 100, 91–140, 232–281, 262–311, 423–472, 451–500, 623–672, 731–780, 821–840, 961–1010, 1061–1110, 1310–1354, and 1410–1459. In this test, the original model obtained with the full data set is still valid. The results are shown in figure 5.

As can be seen from figure 5, the forecasting power is not significantly reduced by the introduction of the missing data values. This is due to the fact that the relevant information contained in the identification data set is not lost.

Linear Model Application

The linear system presented in this example is a position servo–mechanism. The input of the system is a square wave signal, and the output is the angular position of the servo measured in radians. The input and output signals used to

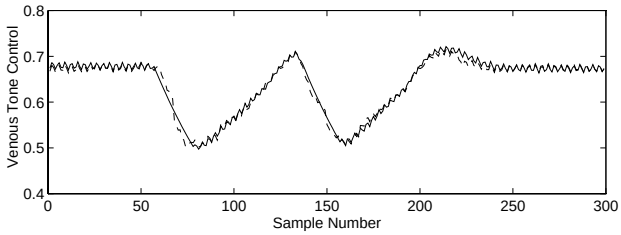

Figure 5: Forecast of the Biomedical System with 40% of Scattered Missing Data

identify the model are shown in figure 6. The data don't contain any missing value.

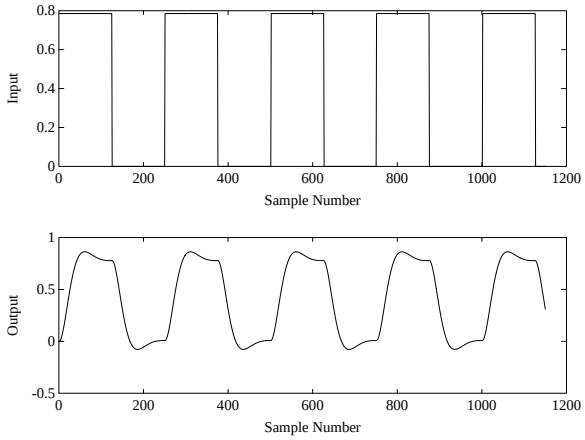

Figure 6: Input/Output Data for Model Identification

The best model found representing the position servo–mechanism is the following:

$$\begin{array}{lcc} {}_t\backslash^x & u & y \\ t-2\delta t & -1 & -2 \\ t-\delta t & -3 & -4 \\ t & -5 & 1 \end{array} \qquad (6)$$

The results of the prediction using this model are presented in figure 7.

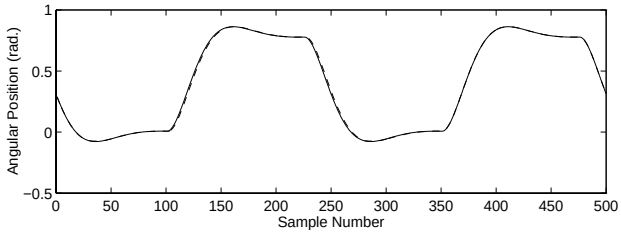

Figure 7: Forecast of the Linear System without Missing Data

As can be seen, the results obtained are excellent, which is not further surprising thanks to the regularity of the output pattern.

Adjacent Missing Data

In a first test, different amounts of missing data values are inserted in the identification data set adjacently. As can be seen in figure 6, the redundancy on the original data is quite large, therefore the forecast results obtained with 50% missing data included in the raw data are as good as the results shown in figure 7.

When the amount of missing data reaches 60% (from sample 1 to 690), the output forecast is affected by the reduction of the available data, as can be seen in the first plot of figure 8. If the amount of missing data is further increased to 75% (from sample 1 to 855), the results get worse (second plot of figure 8).

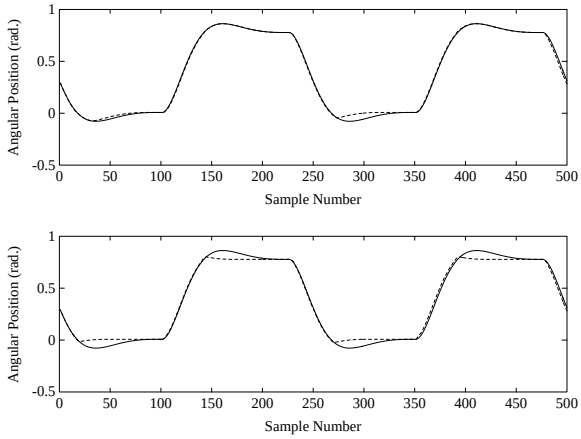

Figure 8: Forecast of the Linear System with 60% and 75% of Adjacent Data Missing

Scattered Missing Data

If the missing data are always located at the maximum of the output variable, the training data set no longer contains any information indicating how the model should behave when the position of the servo–mechanism is at its maximum. In the next test, the missing data gaps were inserted at samples 51 to 90, 301–340, 551–590, 801–840, and 1051–1090 of the output variable, corresponding to a data loss of 17%. The results of the forecast are shown in figure 9.

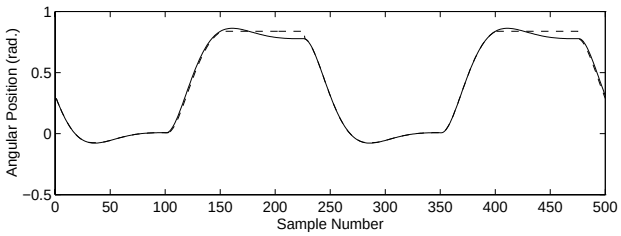

Figure 9: Forecast of the Linear System with Scattered Missing Data

Evidently, a severe degradation of the forecasting power at the peak takes place due to the missing behavioral infor-

mation in the identification data set.

CONCLUSIONS

This paper contains a detailed explanation of how the missing data feature is implemented for use in fuzzy inductive reasoning. It is described how the crucial functions of the FIR methodology: recode, optimal mask, forecast, and regenerate, have been enhanced to allow to work reliably with incomplete data streams.

A discussion of the limitations to predictability with respect to the amount of missing data encountered in the input/output data of the system under investigation is also presented in the paper.

Two different applications have been used to show those limits: a (generic) linear state–space model, and observations of input/output behavior stemming from a biomedical system.

The tests done with both systems show that the limitations to predictability are difficult to quantify in a precise fashion. The degradation of the forecasting power depends on the richness and the redundancy of the data records in the data history. It is the lack of training data (previous behavioral experience) that affects the forecasting power, and not the presence or absence of data gaps.

An important application of the missing data feature is described in (Nebot and Cellier 1994).

ACKNOWLEDGMENTS

The authors are thankful to Dr. Rafael Huber of the Institut de Cibernètica of the Universitat Politècnica de Catalunya for his support on this research.

REFERENCES

Cellier, F.E., A. Nebot, F. Mugica, and A. de Albornoz. 1992. "Combined Qualitative/Quantitative Simulation Models of Continuous–Time Processes Using Fuzzy Inductive Reasoning Techniques." In *Proc. SICICA'92, IFAC Symposium on Intelligent Components and Instruments for Control Applications* (Málaga, Spain, May 20–22), 589–593.

Cellier, F.E., and D.W. Yandell. 1987. "SAPS–II: A New Implementation of the Systems Approach Problem Solver," *Internat. Journal of General Systems*, no. 13 (4): 307–322.

de Albornoz, A., and F.E. Cellier. 1993a. "Qualitative Simulation Applied to Reason Inductively About the Behavior of a Quantitatively Simulated Aircraft Model." In *Proc. QUARDET'93, IMACS Internat. Workshop on Qualitative Reasoning and Decision Technologies* (Barcelona, Spain, June 16–18), 711–721.

de Albornoz, A., and F.E. Cellier. 1993b. "Variable Selection and Sensor Fusion in Automatic Hierarchical Fault Monitoring of Large Scale Systems." In *Proc. QUARDET'93, IMACS Internat. Workshop on Qualitative Reasoning and Decision Technologies* (Barcelona, Spain, June 16–18), 722–734.

Klir, G.J. 1985. *Architecture of Systems Problem Solving*, Plenum Press, New York.

Klir, G.J. 1989. "Inductive Systems Modelling: An Overview," in: *Modelling and Simulation Methodology: Knowledge Systems' Paradigms* (M.S. Elzas, T.I. Ören, and B.P. Zeigler, Eds.), Elsevier, Amsterdam, The Netherlands.

Klir, G.J., and T.A. Folger. 1988. *Fuzzy Sets, Uncertainty, and Information*, Prentice–Hall, Englewood Cliffs, N.J.

Mugica, F., and F.E. Cellier. 1993. "A New Fuzzy Inferencing Method for Inductive Reasoning." In: *Proc. International Symposium on Artificial Intelligence* (Monterrey, México, September 20–24), 372–379.

Li, D., and F.E. Cellier. 1990. "Fuzzy Measures in Inductive Reasoning." In *Proc. 1990 Winter Simulation Conference* (New Orleans, LA), 527–538.

Nebot, A., F.E. Cellier, and D.A. Linkens. 1993. "Controlling an Anaesthetic Agent by Means of Fuzzy Inductive Reasoning." In *Proc. QUARDET'93, IMACS Internat. Workshop on Qualitative Reasoning and Decision Technologies* (Barcelona, Spain, June 16–18), 345–356.

Nebot, A., and F.E. Cellier. 1994. "Preconditioning of Measurement Data for the Elimination of Patient–Specific Behavior in Qualitative Modeling of Medical Systems." In *Proc. CISS'94, First Joint Conference of International Simulation Societies* (Zurich, Switzerland, August 22–25).

Wang, Z., and G.J. Klir. 1992. *Fuzzy Measure Theory*, Plenum Press, New York.