Model Mining of Multivariate Heterogeneous Time-Varying Data using Heterogeneous Neurons: Its Potential in the Study of Geological Hazards

Valdés, Julio

Publisher's version / Version de l'éditeur:

Proceedings of the 2003 Annual Conference of the International Association for Mathematical Geology (IAMG 2003), 2003

National Research Council Canada
Conseil national de recherches Canada

Institute for Information Technology
Institut de technologie de l'information

NRC · CNRC

*Model Mining of Multivariate Heterogeneous Time-Varying Data using Heterogeneous Neurons: Its Potential in the Study of Geological Hazards **

Valdés, J.
September 2003

* published in Proceedings of the 2003 Annual Conference of the International Association for Mathematical Geology. (IAMG 2003), Portsmouth, UK. September 7-12, 2003. NRC 46515.

Canada

MODEL MINING OF MULTIVARIATE HETEROGENEOUS TIME-VARYING DATA USING HETEROGENEOUS NEURONS: ITS POTENTIAL IN THE STUDY OF GEOLOGICAL HAZARDS

Julio J. Valdés
National Research Council of Canada
Institute for Information Technology
1200 Montreal Rd, Ottawa, K1A 0R6, Canada
julio.valdes@nrc-cnrc.gc.ca

Introduction

Processing heterogeneous information is necessary in many domains, but it is specially important in the study of complex systems like geologic phenomena responsible for natural hazards. The developments in sensor, communication and computer technologies allows the monitoring and storage of different geophysical, geochemical and geological variables on a continuous base. In addition, field observation and laboratory analysis incorporate more information, also changing with time. Some of the problems associated with these large masses of information are: (*a)* the presence of *heterogeneous* variables (i.e. ratio, interval, nominal, ordinal, and other more complex like images, spectra, etc.), *(b)* the *uncertainty* associated with the observed variables in terms of observation or measurement errors, vagueness, subjectivity, etc, and *(c)* the *incompleteness* due to monitoring gaps, irregular sampling frequencies and missing data.
In complex or poorly known processes, knowledge discovery oriented to unveil the underlying structure of the process is crucial, specially for revealing patterns and time dependencies, detecting abnormal behavior, deriving prediction criteria, and constructing forecasting procedures.
This paper discusses a hybrid approach based on a combination of *soft-computing* techniques for model discovery and model-change detection in multivariate time processes with different kind of variables, missing data and uncertainty. The models are represented by hybrid neural networks using *heterogeneous neurons,* which accept as input mixed, fuzzy and missing data. The method finds sets of non-linear models having bounded prediction accuracy over a target signal, and characterizes the overall time dependencies between the heterogeneous time series as probability distributions over their sets of time lags. The main steps are: a search in the space of dependency models, and a study of the probability distributions and their changes in a selected subset having prediction errors within a preset bound.
An example is presented using a public domain set of meteorological data, consisting on 20 precipitation and temperature series from 10 stations.

Heterogeneous Domains and Multivariate Time Series

Processing heterogeneous information is a continuously growing need specially in geosciences where the complex natural systems being studied are characterized by variables of many different kinds (geological, geophysical, geochemical, etc). These variables are represented by magnitudes corresponding to different measurement scales (nominal, ordinal, interval and ratio), and also by more complex types of information as satellite images, instrument records, written reports, diagrams, etc. Some of them are obtained by measuring or recording instruments, whereas others are purely judgmental and subjective, determined by human experts. As consequence, they have different kinds of uncertainty associated with them (imprecision, vagueness, etc).
With the developments on sensor, sampling and laboratory technologies the databases grow at an enormous rate. Also, when observations are conducted for a given period of time (monitoring), not only they contain the information provided by variables which are different in nature, precision and objectivity, but also there is a time dependency component. Moreover, the complexity of the physical observations, electronic and/or mechanical malfunctionings, human errors and other factors, make these databases *incomplete*, usually containing large quantities of *missing values*.
Crucial to the investigation in geosciences is the knowledge discovery process and the data mining of this information, where tasks like classification and prediction are of most importance. However, most data analysis methods work on single-type data, or allows only very few types simultaneously. Many of them have difficulties in handling missing values, and can not account for time dependencies.
A formal approach for describing heterogeneous information was given in (Valdés and Garcia 1997) for constructing neuron models and in (Valdés 2002b) for general observational problems.

For describing heterogeneous observational data, different *information sources* are associated with the attributes, relations and functions. These sources are associated with the nature of what is observed (e.g. point measurements, signals, documents, images, etc). They are described by mathematical sets of the appropriate kind called *source sets* and denoted by Ψ_i, constructed according to the nature of the information source to represent (e.g. point measurements of continuous variables by subsets of the reals in the appropriate ranges, structural information by directed graphs, etc). They should also account for incomplete information, as for example, in the following way: ? is a special symbol denoting the missing information with two basic properties: *(i)* if $? \in S$ (S being an arbitrary set) and f is any unary function defined on S, $f(?) = ?$, and *(ii)* ? is an incomparable element w.r.t any ordering relation in any set to which it belongs.

A heterogeneous domain is defined as a cartesian product of a collection of source sets:

$\hat{H} = \Psi_1 \times \cdots \times \Psi_n$, where $n > 0$ is the number of information sources to consider.

As an example, consider the case of an heterogeneous domain where objects are characterized by attributes given by continuous crisp quantities, discrete features, fuzzy features, time-series, images, and graphs. Individually, they can be represented as Cartesian products of subsets of real numbers($\hat{R}$), nominal ($\hat{M}$) or ordinal sets($\hat{O}$), fuzzy sets($\hat{F}$), set of images ($\hat{I}$) **,** set of time series ($\hat{S}$) and sets of graphs ($\hat{G}$)**,** respectively, all properly extended for accepting missing values. Thus, the heterogeneous, time dependent domain is $\hat{H}^n(t) = \hat{N}^{n_N}(t) \text{x } \hat{O}^{n_O}(t) \text{x } \hat{R}^{n_R}(t) \text{x } \hat{F}^{n_F}(t) \text{x } \hat{I}^{n_I}(t) \text{ x } \hat{G}^{n_G}(t) \text{ x } \hat{S}^{n_S}(t)$, where n_N is the number of nominal sets, n_O of ordinal sets, n_R of real-valued sets , n_F of fuzzy sets , n_I of image-valued sets, n_S of time-series sets, and n_G of graph-valued sets, respectively (Fig-1).

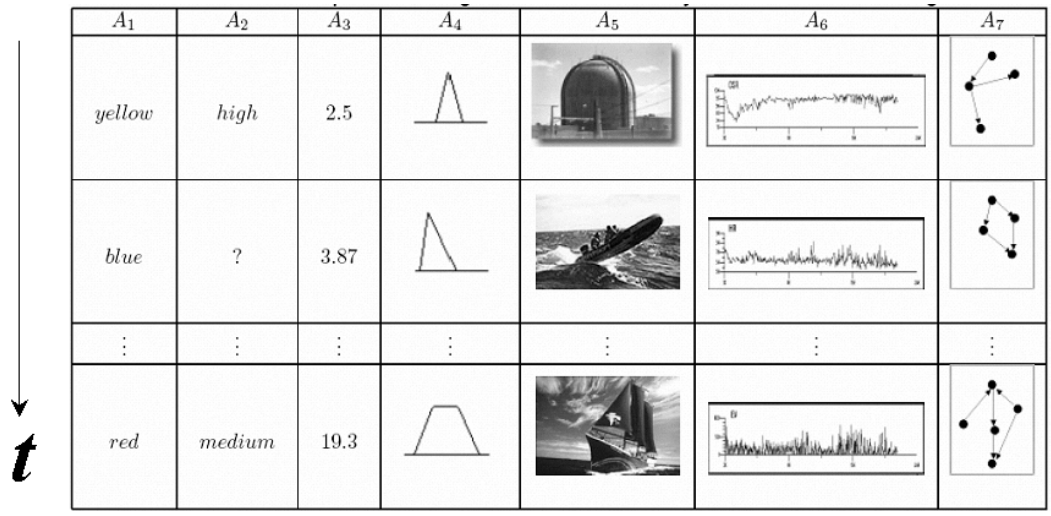


Figure 1. A heterogeneous, time-dependent multivariate process. Each row is an object described by nominal, ordinal, ratio, fuzzy, image, time-series and graph attributes, possibly with missing values (?), for a given time. The sampling interval is assumed to be the same for all attributes.

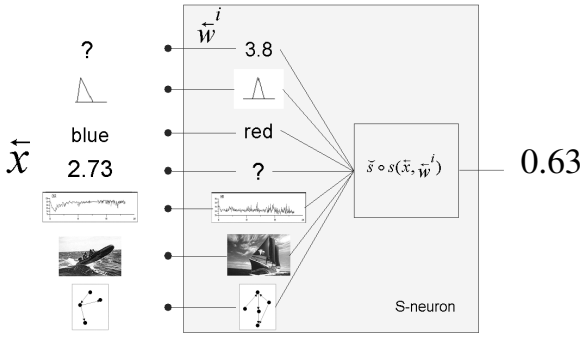
Model Mining with Heterogeneous Neurons and Hybrid Neural Networks

The classical approaches in time series consider mostly univariate, homogeneous (real-valued), time series, without missing values (Box and Jenkins 1994)(Masters 1995)(Pole and others 1994). The purpose of model mining in heterogeneous, multivariate, time varying processes is to discover *dependency models* . A model expresses the relationship between values of a previously selected time series (the target), and a subset of the past values of the entire set of series. Different classes of functional models could be considered, in particular, a generalized non-linear auto-regressive (AR) model like the one given by Equation-1. Note that for the same set of heterogeneous time series, a different model can be obtained for each of the series from the set, with different composition. Without loss of generality, the rest of the discussion will focus on a mining process targeting a single time series.

$$S_{\text{target}}(t) = \mathrm{F}\begin{pmatrix} S_1(t-\tau_{1,1}), S_1(t-\tau_{1,2}), \ldots, S_1(t-\tau_{1,p_1}), \\ S_2(t-\tau_{2,1}), S_2(t-\tau_{2,2}), \ldots, S_2(t-\tau_{2,p_2}), \\ \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots, \\ S_n(t-\tau_{n,1}), S_n(t-\tau_{n,2}), \ldots, S_n(t-\tau_{n,p_n}) \end{pmatrix} \qquad (1)$$

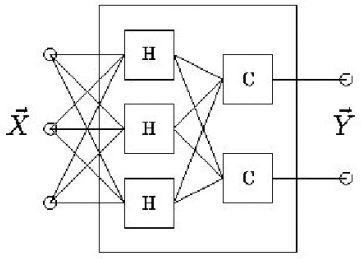
where n is the number of signals, $S_1, S_2, \cdots, S_n$ is the set of signals, F is the unknown function, $p_1, \cdots, p_n$ is the number of lags considered for signal S_i, and $\tau_{i,j}$ is the $j-th$ time lag corresponding to signal i.

A hybrid soft-computing algorithm for solving this kind of problems using genetic algorithms and heterogeneous neural networks (Valdés and García 1997, Valdés and others 2000, Belanche 2000), has been given elsewhere (Valdés 2002). This approach requires the simultaneous determination of: *(i)* the number of required lags for each series, *(ii)* the particular lags within each series carrying the dependency information, and *(iii)* the prediction function. A requirement on function **F** is to minimize a suitable prediction error, usually the root mean squared error (RMSE). This is approached with a soft computing procedure based on: *(a)* exploration of a subset of the model space with a genetic algorithm, and *(b)* use of a similarity-based neuro-fuzzy system representation for the unknown prediction function. Statistical or other classical approaches either have difficulties on handling these kinds of situations or can not handle them at all (Box and Jenkins 1994) (Pole and others 1994). Clearly, the size of the model space to search is immense and it grows exponentially (considering only 10 time series and the first 20 time lags, it contains about 10^{60} models). The prediction function **F** is represented by a hybrid neural network with a hidden layer composed by heterogeneous neurons (h-neurons). A heterogeneous neuron is a general mapping $h: \hat{H} \times \hat{H} \to Y$, where $\hat{H}$ is a heterogeneous domain, and Y is an arbitrary set. If $x, w \in \hat{H}$, and $y \in Y$, then $y = h(x, w)$ A particular class of h-neurons is obtained when Y the real interval [0,1] and h is given by a composition of a similarity function $s(x, w)$ (Chandon and Pinson 1981), and an isotone automorphism $g: [0,1] \to [0,1]$ (in general it is a non-linear function). In this case the h-neuron is given by $h(x, w) = g \circ s(x, y)$ and called a similarity-based neuron (Fig-2B). This neuron model is flexible (heterogeneous data with missing values are its natural input, without the need of data type transformation or imputation of missing values), and it is robust. Also it has the general function approximation property (Belanche 2000).



(A)

Figure 2(A). A similarity-based heterogeneous neuron. Both the input and the neuron weights are objects from a heterogeneous domain. The output is a similarity value.



(B)

Figure 2(B). A hybrid neural network composed by a hidden layer of heterogeneous neurons (H), and an output layer of classical neurons (C). The input is a heterogeneous object and the output a real-valued vector.

Moreover, the s-neuron can be coupled with classical neurons (the scalar product is the aggregation function, and the sigmoid or hyperbolic tangent, the activation), forming hybrid neural networks (Fig-2B).

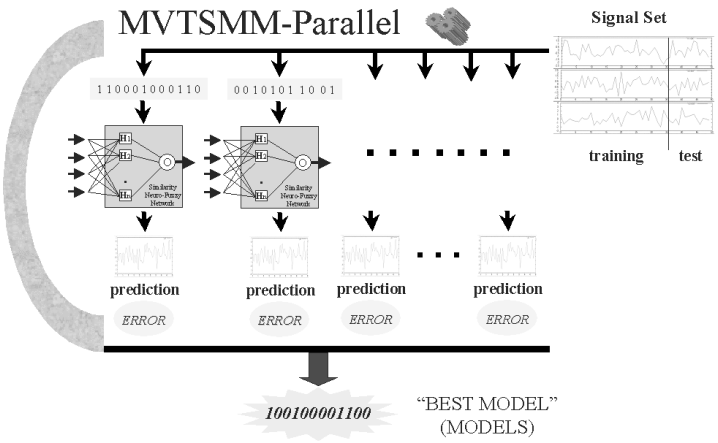


Figure 3. Multivariate Time Series Model Miner System Architecture (MVTSMM). The models with their RMSError under a preset threshold are collected at the end of the mining process.

The MVTSMM system is an implementation of this approach to model mining in heterogeneous time series as a parallel computing algorithm (Fig-3). The arc is the parallel genetic algorithm evolving populations of similarity-based hybrid neural networks. The binary strings encode dependency patterns for the target signal, and for each, a hybrid neural network is constructed and trained with a fast algorithm. The network represents the prediction function **F**, and is applied to a separate time-series test set. At the end of the evolutionary process the best models are collected.

A Model Mining Example with Meteorological Data

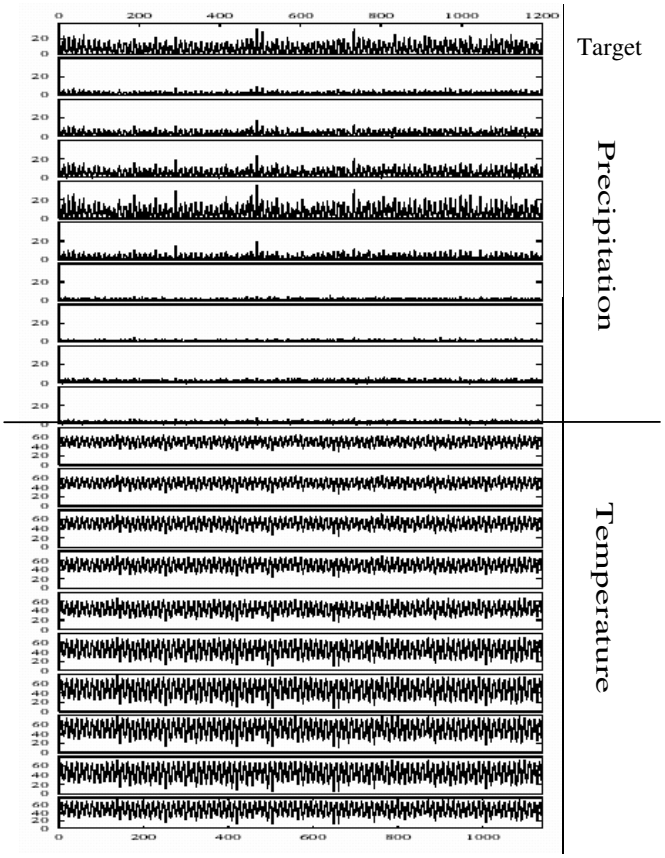


Figure 4. Meteorological data from Washington State. Upper 10 are Precipitation series and lower 10, Temperature series. The first precipitation series was used as target.

A multivariate time series data set consisting of *20* series with *1140* observations of average monthly precipitation and temperatures from different sites in the Washington State (USA) was chosen. They were recorded during the period 1895-1989 (Masters 1995), and compiled by the National Oceanic and Atmospheric Administration (USA). Originally, this data had no missing values (Fig-4). The precipitation signal for the West Olympic Coastal drainage region (the top series) was chosen as the target for prediction. Contrary to the standard practice in time series analysis, no preprocessing was applied to the time series, in order to test the approximation capacity and robustness of the algorithm in the possibly worst conditions. A total of 1001 models were found with a RMSE threshold given by the first quartile (*3.818*) and they are those giving good predictions on the test set.

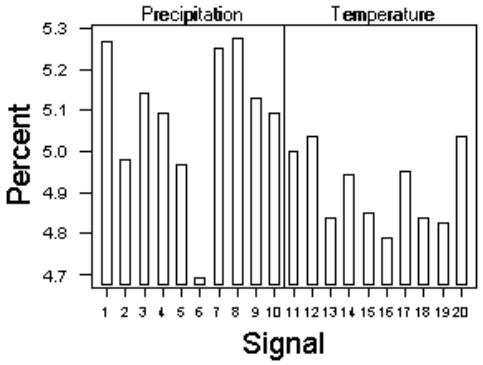


Figure 5. Histogram of signal occurrence in the set of models with RMSError <= 3.8178 (the first quartile).

A rough picture of the contribution of the different signals to this set of interesting models is given by the relative frequency with which time lags from the corresponding signal occur as terms in Equation-1 (Fig5). It is interesting to see that the distribution is clearly multimodal, with some signals tending to contribute more frequently than others, in correspondence with the location of the different hydrological basins w.r.t the North Cascades Mountain Range. Moreover, although the class of precipitation signals (the first *10* in Figure-4), contributes the most (as could be expected, since the target is a precipitation signal), there is an important contribution from the temperature signals as could be expected. Sensitivity analysis should be performed as well, in order to evaluate the relative RMSE impact of the different lags contributing by the different signals.

Anticipating State Changes in Time-Varying Processes: A Simulation Example

One of the most important problems in predicting the behaviour of complex systems is to anticipate changes of state. In order to assess the potential of a detailed analysis of time series information with the MVTSMM algorithm, a single time series of length 900 was constructed as a convex combination mixture of two linear AR processes in the following way:

$X(t) = c_1 X(t-1) + c_2 X(t-2) + \varepsilon(t)$, $Y(t) = c_1 Y(t-2) + c_2 Y(t-3) + \varepsilon(t)$, where $c_1 = 0.4$, $c_2 = 0.5$, and $\varepsilon(t) = N(0,0.001)$ is a gaussian noise with mean=0 and variance=0.001. Clearly, the two processes have the same statistical properties, but differ only in the time lags: (1,2) in the first, and (2,3) in the second, which is the least possible difference between the two (just one single consecutive time lag). The mixed process (Fig-6) is given by $Z(t) = \alpha(t)X(t) + (1-\alpha(t))Y(t)$.

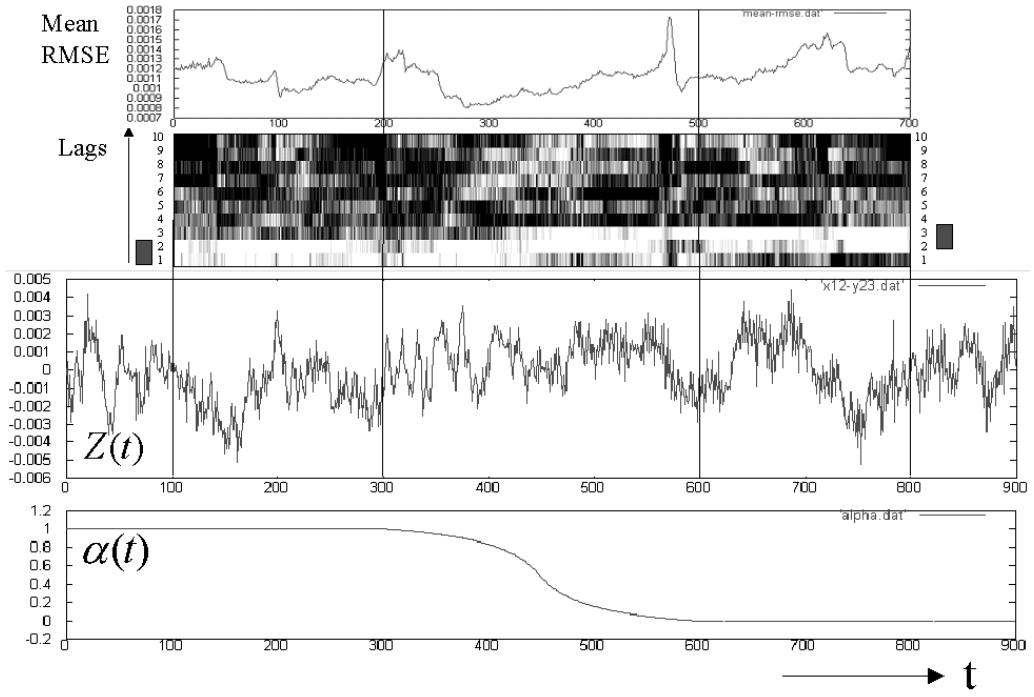


Figure 6. *Z(t)* is a linear AR model computed as a convex combination ($\alpha(t)$)of two linear AR processes with identical coefficients and generating gaussian noise. The upper image shows the probability of each time lag to be included in a model with low prediction error, as a function of time. For each time, the mean RMSE of all models found, as a function of time, is shown at the top.

The convex combination coefficient was not constant, but a function of time, such that $\alpha(t) = 1$ for $t \leq 300$, $\alpha(t) = 0$ for $t \geq 600$, and it has a smooth transition between the two values for $t \in [300,600]$. Then, the MVTSMM algorithm was applied on sliding windows of size 200 along the entire series, exploring models up a maximum depth of 10 time lags. Within each window, the first 100 values were used as test, and the remaining 100 as test. The 10-best discovered models were retained, and the mean RMSE of the 10-best models was computed, as well as an empirical probability distribution for the lags composing the model. In Figure 6 the distributions are displayed as a grey-level image where for each time slice at time *t*, a black-white grey scale spans the [0,1] probability values for the given lag (along the vertical). For *t* under 300, the image contains only two bright horizontal strips, corresponding to lags 1 and 2, as expected, whereas for *t* over 600, the bright strips are those of lags 2 and 3. also as expected. In the transition zone (*t* within 300 and 600), the strip of lag 1 fades whereas that of lag 3 gradually gets brighter, and that of lag 2 remains unchanged. The midpoint (450) where the process starts to be more *Y* than *X* can be clearly identified, with 50 time intervals of anticipation. However, the RMSE and the *Z* process itself show no appreciable change.

The simulation was purposely designed to make the detection difficult, but nevertheless, these results are very preliminary. They show, however, that if the dependency models *are considered as random variables*, an in-depth model mining with techniques like the one outlined may unveil the existence of hidden internal changes within the process. These hidden or subtle changes might indicate that the system is actually in a transient regime, possibly evolving towards a dangerous state. This information could be used as an early warning that a major change is about to come, long before the change happens.

References

Belanche, Ll., 2000, Heterogeneous neural networks: Theory and applications. PhD Thesis, Department of Languages and Informatic Systems, Polytechnic University of Catalonia, Barcelona, Spain,July 2000.

Box, G., Jenkins, G.,1976, Time Series Analysis, Forecasting and Control. Prentice Hall, 592 p.

Chandon, J.L., and Pinson, S., 1981, Analyse typologique. Théorie et applications: Masson, Paris, 254 p.

Masters,T. 1995, Neural, Novel & Hybrid Algorithms for Time Series Prediction. John Wiley & Sons, 544 p.

Pole,A., West M., Harrison J., 1994, Applied Bayesian Forecasting and Time Series Analysis. CRC Press, 432 p.

Valdés, J.J, García R., 1997, A model for heterogeneous neurons and its use in configuring neural networks for classification problems. In Proceedings. of the International. Work-Conference on Artificial and Natural Neural Networks, IWANN'97. Biological and Artificial Computation: From Neuroscience to Technology, (Mira, Moreno-Díaz, Cabestany eds.), Lecture Notes in Computer Science 1240, Springer-Verlag, pp 237-246.

Valdés J.J, Belanche, L, Alquézar, R., 2000, Fuzzy Heterogeneous Neurons for Imprecise Classification Problems. International Journal of Intelligent Systems, 15(3).

Valdés, J.J., 2002, Time Series Models Discovery with Similarity-Based Neuro-Fuzzy Networks and Evolutionary Algorithms. Proc. IEEE World Congress on Computational Intelligence WCCI'2002. IEEE, Hawaii, USA, May 12-17, 0-7803-7278-6/02.

Valdés, J.J., Mateescu G, 2002b, Time Series Models Discovery with Similarity-Based Neuro-Fuzzy Networks and Genetic Algorithms: A Parallel Implementation. Third. Int. Conf. on Rough Sets and Current Trends in Computing RSCTC 2002. Malvern, PA, USA, Oct 14-17. Alpigini, Peters, Skowron, Zhong (Eds.) Lecture Notes in Computer Science (Lecture Notes in Artificial Intelligence Series) LNCS 2475, pp. 279-288. Springer-Verlag.

Valdés, J.J., 2002b, Similarity-based heterogeneous neurons in the context of general observational models: Neural Network World, v.12, no. 5, p. 499-508.

Valdés, J.J.,Barton, 2003, Mining Multivariate Time Series Models with Soft-ComputingTechniques: A Coarse-Grained Parallel Computing Approach. 2003 International Conference on Computational Science and its Applications (ICCSA 2003).Technical Session on Coarse Grained Parallel Algorithms For Scientific Applications. Montreal, May 18-24, 2003. Lecture Notes in Computer Science (Kumar, Gavrilova, Tan, L'Ecuyer eds.) LNCS 2668, Springer-Verlag , pp. 259-268..