Proceedings of the 2009 IEEE International Conference on Systems, Man, and Cybernetics
San Antonio, TX, USA - October 2009

An Evolutive Interval Type-2 TSK Fuzzy Logic System for Volatile Time Series Identification.

Juan Carlos Figueroa García.
Laboratory for Automation, Microelectronics and Computational Intelligence (LAMIC).
Universidad Distrital Francisco José de Caldas, Bogotá - Colombia.
e-mail: jcfigueroag@udistrital.edu.co

***Abstract*—This paper presents a study case of the US Dollar - Colombian Peso Exchange Rate by using a First-order Interval Type-2 TSK Fuzzy Logic System. This study case is specially interesting because it presents a volatile behavior which is a complex problem for classical analysis. Their results are checked by statistical tests as Bayesian, Akaike, Hannan-Quin criteria, Goldfeld-Quant, Ljung-Box, ARCH, Runs and Turning Points which provide appropriate criterions to test the solution.**

Some methodological aspects about a hybrid implementation between Evolutive Optimization and First order Interval Type-2 TSK FLS are presented. Additionally, the selected type-reduction algorithm is the *IASCO* algorithm proposedby Melgarejo in [1] since it presents better properties than other algorithms.

I. Introduction and Motivation.

During the last 50 years, economic and financial Time Series have become as a complex problem for identification and prediction because those series have volatile behavior. This volatility implies important problems for statistical models which do not provide good practical results.

During the last 50 years, economic and financial Time Series have become as a complex problem for identification and prediction because those series show a volatile behavior. This implies important problems for statistical models which do not provide good practical results.

This work is focused on methodological aspects to design a First-order Interval Type-2 Takagi-Sugeno Fuzzy Logic System *(IT2FSK)* for volatile Time Series identification, as a continuation of the work of Figueroa and Soriano in [2]. Some important aspects about the design of Type-2 fuzzy input sets, TSK outputs, rule base and inference process are described with an implementation on a study case. To do so, we use the *IASCO* type reduction algorithm proposed by Miguel Melgarejo in [1] and [3], which shows better performance than Karnik-Mendel algorithms proposed by Nilesh Karnik and Jerry Mendel in [4], [5], [6] and Jia Zeng & Zhi-Qiang Liu [7], improving computing time. A computational intelligence approach to IT2TSK systems was given by Mendez and Castillo in [8] applied to a temperature control problem.

The paper is organized as follows: Section I presents the introduction of the study, Section II presents the case study and their statistical analysis. Section III presents the designed Genetic-IT2TSK system. Section IV presents the IT2TSK results. Section V presents the residual analysis of the method and finally we suggest some concluding remarks.

II. US Dollar - Colombian Peso Exchange Rate

US Dollar - Colombian Peso (USD-COP) exchange rate is a time series that presents deviations from linearity and volatile behavior. The Figure 1 shows the original time series which has no constant mean and variance, therefore we differentiate the series to obtain a constant mean process as is shown in the Figure 2.

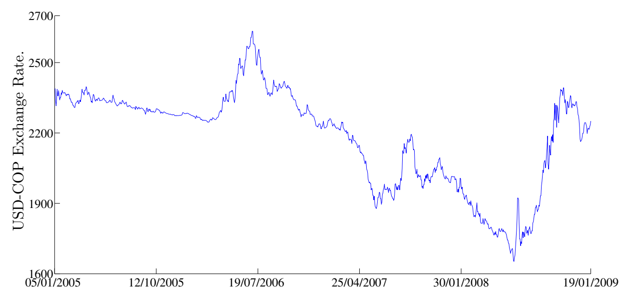

Fig. 1. USD-CP Exchange Rate Data.

Clearly, the structure of the variance of this time series is not constant and non-linear for which we perform two suitable heteroscedasticity tests: The Golfeld-Quant [9], ARCH [10] and the Ljung-Box tests [11] with an $\alpha = 0.05$ are summarized in the Table I, Table II and Table III respectively.

These tests suggest that the series does not present a linear behavior neither a suitable structure for linear statistical modelling, other tests about normality and randomness can be used for determining whether the suppositions for the classical analysis are true or not, which are summarized in Table IV.

TABLE I
Goldfeld-Quant Results.

Date Interval	F-test	p-value
04/01/2005 - 08/05/2006 ; 18/09/2007- 19/01/2009	9.786	≈ 0

TABLE II
ARCH Test Results.

Lag	ARCH Stat	Critical Value	p-value
1	79.99	3.84	≈ 0
2	91.05	5.99	≈ 0
3	222.97	7.81	≈ 0
4	235.84	9.48	≈ 0
5	241.88	11.07	≈ 0
6	244.15	12.59	≈ 0

978-1-4244-2794-9/09/$25.00 ©2009 IEEE

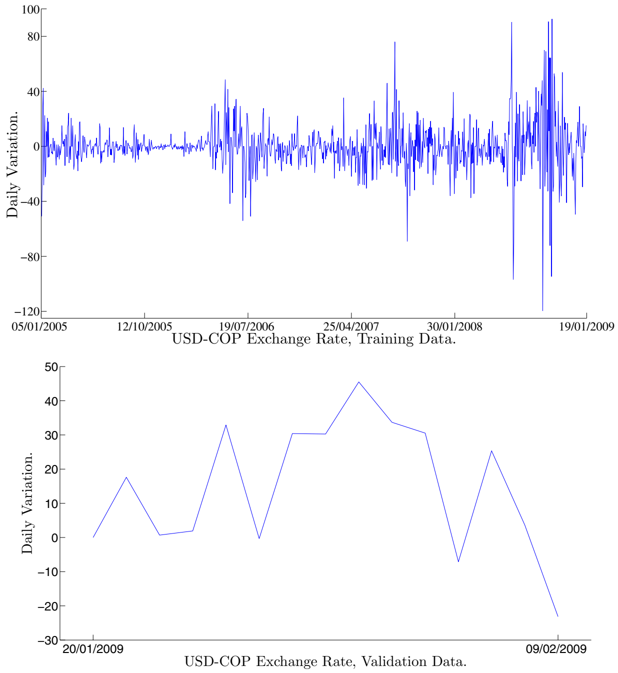

Fig. 2. USD-CP Differentiated Exchange Rate Data.

TABLE III
LJUNG-BOX TEST RESULTS.

Lag	L-B Stat	Test Stat	p-value
20	71.05	31.41	≈ 0
25	73.17	37.65	≈ 0

TABLE IV
RANDOMNESS AND NORMALITY TESTS.

Test	Stat	Test Stat	p-value
Randomness Tests			
Runs Test	4.5155	1.9599	≈ 0
Turning Points	1.9591	1.9599	0.0501
Normality Tests			
Shapiro-Wilks Test	0.978	0.854	≈ 0
Kolmogorov-Smirnov Test	0.207	0.144	≈ 0
Anderson - Darling Test	40.14	2.49	≈ 0

In Tables I to IV, Stat refers to the obtained statistic of each test, Test Stat is the theoretical value of the Test, that is, its critical value and the p-value is the significance of the test, in other words the probability of fulfillment of the test.

All tests for randomness and normality suggest that the time series is deviated from normality and presents some kind of relation among other variables because is not a random sample. This asseveration is logical since those financial time series are correlated with hundreds of economic and instrumental variables present in a typical economic scheme.

Brockwell and Davis in [12] suggest that both Runs and Turning point Tests are useful to verify the randomness of the series, the Shapiro-Wilks and Kolmogorov-Smirnov Tests are used by Thode in [13] for contrasting the normality hypothesis in time series. With these statistical evidences it is clear that classical linear models and ARIMA methodologies are not good candidates for modeling the series.

A. Training and Testing Data.

The training data set is defined as follows: 1070 bussines days, removing Saturdays and Sundays because in these days the USD-CP Exchange rate has no measurement. A test data set is defined on 15 days to test the generalization capability of the model. We use only 15 days since Financial time series have high sensibility in front to ergodicity, that is, the information of older than 30 days is not very useful to forecast the present value, so a larger dataset is not a good choice because it would be too old.

An ARIMA(5,1,12) model for Train and Test Data Sets was estimated. The obtained model is summarized as follows:

$$\begin{aligned} X(t) &= -.1029X_{t-1} - .04219X_{t-2} + .1621X_{t-3} + .1217X_{t-4} \\ &+ .2269X_{t-5} + .1804X_{t-6} + .2981X_{t-7} - .1866X_{t-8} \\ &+ .02603X_{t-9} + .1133X_{t-10} - .04630X_{t-11} - .1464X_{t-12} \\ &- .05899X_{t-13} + .01842X_{t-14} + .08551X_{t-15} + .06141X_{t-16} \\ &+ .01925X_{t-17} - .1253X_{t-18} + Z_t + .1973Z_{t-1} \\ &+ .01166Z_{t-2} - .2791Z_{t-3} - .07899Z_{t-4} - .2644Z_{t-5} \\ &- .3758Z_{t-6} - .2440Z_{t-7} + .03843Z_{t-8} + .04368Z_{t-9} \\ &- .03501Z_{t-10} - .04579Z_{t-11} + .3938Z_{t-12} + .1785Z_{t-13} \end{aligned} \quad (1)$$

Here, X_{t-h} are the autoregressive values and Z_{t-p} are the moving average values of the model.

Some important statistics of the model are derived from their residuals, as well as normality and stability conditions. Now, ARIMA and GARCH models has no good results due to the high volatility of data and the analysis of their residuals shows worst results.

III. INTERVAL TYPE-2 FUZZY LOGIC SYSTEM DESIGN METHODOLOGY.

The classical modelling scenario for Time Series identification uses the General Linear Model:

$$X_t = \widehat{X}_t + e_t \quad (2)$$

Where X_t is an observation of the process in the time instant "t", $\widehat{X}_t$ is an *Estimation of* X_t and e_t is the model residual or residual estimation of $\widehat{X}_t$.

Methodological issues conduce us to refer the classical modelling procedure proposed by Box and Jenkins in [14] which is focused on a statistical point of view. In this paper, $\widehat{X}_t$ is obtained by using a IT2TSK and its goodness of fit is evaluated by means of classical residual analysis theory.

General facts about the proposed fuzzy inference engine are:

- Fuzzy input sets are interval Type-2 fuzzy sets and Outputs are TSK sets.
- Both upper and lower membership functions have the same rule base.
- Operators: *and:* min, *or:* max, *Aggregation:* max.
- Type reduction: *IASCO* algorithm proposed by Melgarejo in [1].

Now, we define 6 input vectors in the form of (1), and 2 sigmoidal IT2 Fuzzy sets per input which result on 12 IT2 fuzzy sets and a complete combination of 64 rules which yield into 64 TSK output sets with 7 parameters (An intercept and one weight per input). By using classical autocorrelation criterions, we decide to use the following 6 inputs: $x_{t-1}, x_{t-2}, x_{t-5}, x_{t-30}, z_{t-2}$ and z_{t-5} so the j_{th} TSK polynomial output Y^j is defined as:

$$Y^j = a_0^j + a_1^j x_{t-1} + a_2^j x_{t-2} + a_3^j x_{t-5} + \\ + a_4^j x_{t-30} + a_5^j z_{t-2} + a_6^j z_{t-5} \quad (3)$$

For each input, we use two linguistic variables: *Positive variation*, p, and *Negative Variation*, n, and each IT2 fuzzy set is designed around these linguistic variables. In addition, the lower and upper membership functions of each $\tilde{p}(x_i)$ and $\tilde{n}(x_i)$ fuzzy set are $\bar{\mu}_{\tilde{p}}(x_i), \underline{\mu}_{\tilde{p}}(x_i), \bar{\mu}_{\tilde{n}}(x_i)$ and $\underline{\mu}_{\tilde{n}}(x_i)$. We use quadratic membership functions for each $\bar{\mu}_{\tilde{p}}(x_i), \underline{\mu}_{\tilde{p}}(x_i), \bar{\mu}_{\tilde{n}}(x_i)$ and $\underline{\mu}_{\tilde{n}}(x_i)$ set, in any of the following forms:

$$\mu_p(x_i) = 2\left(\frac{x_i - x_0}{x_1 - x_0}\right)^2 \quad (4)$$

$$\mu_n(x_i) = 1 - 2\left(\frac{x_i - x_0}{x_1 - x_0}\right)^2 \quad (5)$$

Where x_1 and x_0, $x_0 < x_1$ are parameters of form. A graphical representation of $\tilde{p}(x_i)$ and $\tilde{n}(x_i), \bar{\mu}_{\tilde{p}}(x_i), \underline{\mu}_{\tilde{p}}(x_i), \bar{\mu}_{\tilde{n}}(x_i)$ and $\underline{\mu}_{\tilde{n}}(x_i)$ is depicted in the Figure 4.

In this way, each rule R^j can be represented as Mendel defined in [6]:

$$R^j = \text{IF } x_{t-1} \text{ is } \tilde{n}_{t-1}^j \text{ and } \cdots \text{ and } z_{t-5} \text{ is } \tilde{p}_{t-5}^j, \\ \text{THEN } Y^j = a_0^j + a_1^j x_{t-1} + a_2^j x_{t-2} + \cdots + a_6^j z_{t-5} \quad (6)$$

This configuration of IT2FS, rules and TSK outputs yield on 48 parameters for the input sets according to (4), (5) and 448 parameters for the TSK outputs which will define the size of each individual and its number of genes to construct the genetic structure.

A comprehensive diagram which describes the methodology used in this work is shown in the Figure 3.

IV. Evolutionary Optimizer.

The selected strategy to synchronize the IT2TSK system is an Evolutionary Optimizer which is an appropriate way to find solutions by using genetic structures. Some of most important issues of the designed algorithm are described next.

A. Fitness Function Operator.

The main idea of the fitness function $\mathcal{F}$ is to minimize the Mean Squared Error (MSE) between the original series and the output of the IT2TSK after Type-reduction and deffuzification, as is presented:

$$\mathcal{F} = \sum_{i=1}^{n} \frac{(x_i - \hat{x}_i)^2}{n-1} \quad (3)$$

Where $\hat{x}_i$ is the estimate of x_i through the IT2TSK.

Thus, the main goal is to find an evolutionary estimate of the parameters of the IT2 fuzzy input sets and the parameters of the TSK output sets, as were defined in above Section. Therefore, the principal objective of the genetic structure is to minimize $\mathcal{F}$.

B. Individuals, Population Size and Number of Generations.

An *Individual* is defined as a *chromosome* of size $m = 496$, where each gen or cell is a set of parameters of the IT2TSK system: The first 48 genes are the parameters of the IT2 fuzzy input sets and the remaining 448 are the parameters of the TSK output sets.

The size of the population is defined by two values: The m parameters and a pre-selected k number of individuals, creating a matrix called $P_{k,m}^g$ where g is the *Generation index*. It is defined that $m = 496$ and $k = 100$.

The generation index g is a genetic operator which is used as stop criterion. In this paper, the maximum number of generations is defined as $G = 1500$.

C. Population Random Generator.

The selected Random Generator is called R_j. The original *pdf* of the series $\{x_i\}$ can be used as a random number generator, but in practice, the uniform generator which is presented next, exhibits better results:

$$R_j(a, b) = a + r_j(b - a)I_{[0,1]}(r_j) \quad (4)$$

Law and Kelton in [15] did an important discussion about pseudo-randomly number generation showing the uniform generator as a proper method for simulating variables. Therefore it is possible to use it instead of the sample distribution.

D. Mutation and Crossover Operators.

- The selected Mutation strategy is described below:
  1) Select a random position for each orderly individual in $P_{k,m}^g$ by $\mathcal{F}$.
  2) Replace the selected position with a new individual by using (4).
  3) Repeat Step 3 for the c_1 better individuals of each population $P_{k,m}^g$.

- The selected Crossover strategy is described below:
  1) Select the c_2 first individuals in the orderly Population $P_{k,m}^g$ by $\mathcal{F}$.
  2) Generate a new individual by replacing all even genes with its respective even gene located in the next individual.
  3) Generate a new individual by replacing all odd genes with its respective odd gene located in the next individual for each one.
  4) Repeat Step 3 for the c_2 better individuals of each population $P_{k,m}^g$.

To complete the population, an additional set of individuals is generated by replacing worst individuals with new ones, trying an exhaustive search. First, the best four individuals are preserved for the next generation and later a complementary

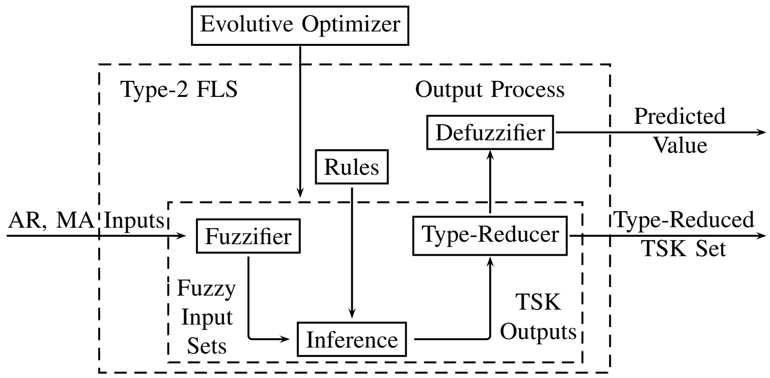

Fig. 3. Evolutive Type-2 TSK FLS design methodology.

population of size $\{k-4-c_1-c_2\times m\}$ is obtained by using (4).

E. Finalization and stopping strategy.

Two stopping criterions are used to stop the algorithm: A first one is by defining a maximum number of iterations called G, that is $g \to G$, and the second one stops when $\mathcal{F}$ has no significant improvement through a specific number of iterations.

Finally, the best individual is selected by $\mathcal{F}$ and it is used to generate an IT2TSK, whose outputs Y^j are deffuzified to obtain an estimation $\hat{x}_i$. A brief description of the procedure is presented next.

Procedure 1 Genetic Algorithm.

Set $k=100, m=496, G=1500, c_1=4, c_2=4$
Generate an initial population $P^1_{k,m}$ {Use (4).}
Use each individual to generate an IT2TSK
return Y^j and $\hat{x}_i$, $i=1,2,\cdots,1055$
Compute $\mathcal{F}$
for $g=1\to G$ **do**
  Order the population by $\mathcal{F}$
  Select the best 4 individuals
  Apply Mutation Operator
  Apply Crossover Operator
  Generate a complimentary population $P^g_{k-4-c_1-c_2,m}$ {Use (4).}
  Use each individual to generate an IT2TSK
  return Y^j and $\hat{x}_i$, $i=1,2,\cdots,1055$
  Compute $\mathcal{F}$
end for
Order the final population by $\mathcal{F}$
Set $P^G_{1,m}$ as final solution.
return $\tilde{p}(x_i)$, $\tilde{n}(x_i)$, Y^j and $\hat{x}_i$, $i=1,2,\cdots,1055$
Compute e_i, $i=1,2,\cdots,1055$

V. EVOLUTIVE INTERVAL TYPE-2 FUZZY LOGIC SYSTEM RESULTS.

After the implementation of the Evolutive IT2TSK, we obtain successful results in terms of a reduced $\mathcal{F}$ and the properties of the residual of the model, improving the results of the best ARIMA model shown in (1) which was obtained by using *RATS*©. An example of the obtained IT2 fuzzy inputs are shown in the Figure 4.

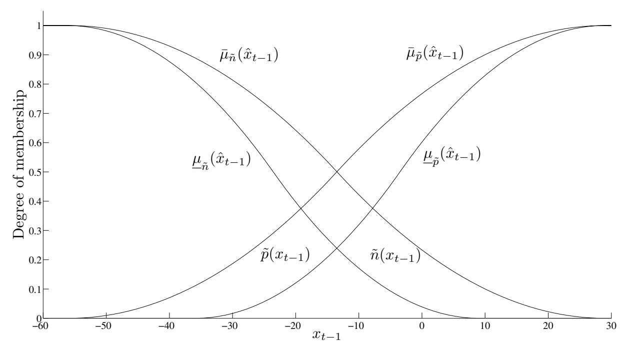

Fig. 4. IT2 Fuzzy Input sets for x_{t-1}.

And the obtained residuals are shown in the Figure 5.

VI. RESIDUAL ANALYSIS.

The residual analysis is based on statistics derived from vector e_t in (2) as Akaike Information criteria in [16] and Schwarz Information criteria (Often known as Bayesian Information criteria BIC in [17]) and Hannan-Quinn (HQC) information criterion in [18] and [19] which estimates the dimension of the model. The Ljung-Box and Goldfeld-Quant tests [20] are used to verify that residuals do not have serial correlation in a selected number of lags[1], Runs Test and Turning Point test

[1]20 Lags are used in this work

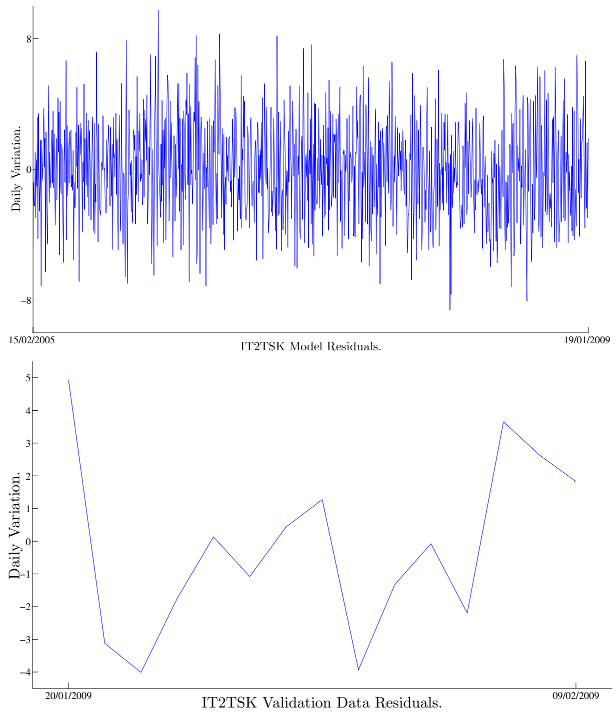

Fig. 5. Residuals of the Evolutive IT2TSK model.

are used to verify the randomness. Arch effect are tested to verify heteroscedasticity in residuals.

The Table V shows some tests on randomness, AIC, BIC and HQC information criterions of the IT2TSK residuals. The normality tests are summarized in Table VI, the Goldfeld-Quant test is shown in the Table VII, the ARCH test is summarized in Table VIII and the Ljung-Box test is shown in the Table IX.

TABLE V
RESIDUAL ANALYSIS FOR ARIMA AND IT2TSK MODELS

IT2TSK Results.		
Tests on Randomness.		
Test	Stat	p-value
Runs Test	1.813	0.069
Turning point	0.222	0.823
Model Adequation Statistics.		
AIC Criteria = 2251.39		
BIC Criteria = 2280.99		
HQC Criteria = 2322.58		
MSE = 8.8887 ; SSE = 9110.9		

These tests suggest that there no exists correlation in residuals of the model, which is a clear sign of linearity. The Ljung-Box and Goldfeld-Quant tests confirm the idea that IT2TSK is a good method for identification and forecasting volatile time series. The effect of volatility is low in the model since it no presents correlation on errors or effect of heteroscedasticity.

Figueroa and Soriano in [2] show that ARIMA models have worst results than computational intelligence techniques as ANFIS, DBR and ANN. Given that antecedent, ARIMA model is not evaluated in this paper.

TABLE VI
NORMALITY TESTS ON RESIDUALS.

Test	Stat	Test Stat	p-value
Shapiro-Wilks	1.134	0.998	0.226
Kolmogorov-Smirnov	0.423	0.018	0.87
Anderson - Darling	0.505	2.49	0.742
Chi-Squared	8.27	21	0.764

TABLE VII
GOLDFELD-QUANT RESULTS.

Date Interval	F-test	p-value
15/02/2005 - 19/06/2006 ; 18/09/2007- 19/01/2009	1.045	0.659

TABLE VIII
ARCH TEST RESULTS.

Lag	ARCH Stat	Test Stat	p-value
1	0.2295	3.84	0.6319
2	1.4698	5.99	0.4796
3	1.6048	7.81	0.6583
4	1.8571	9.48	0.7620
5	1.8832	11.07	0.8651
6	1.9468	12.59	0.9245

TABLE IX
LJUNG-BOX TEST RESULTS.

Lag	L-B Stat	Test Stat	p-value
20	23.1142	31.41	0.2832
25	28.2073	37.65	0.2984

Goldfeld-Quant, ARCH and Ljung-Box Tests confirm that the idea that there are no serial correlation among residuals. AIC, BIC, HQC and MSE criteria lead us to think that residuals has no significant information inside it.

VII. CONCLUDING REMARKS.

The proposed model obtains successful results for volatile time series identification. Their residuals show good properties, improving its applicability. The validation data shows small differences regarding original data.

The proposed genetic structure is computationally expensive, so the design of more efficient genetic structures is a challenge for upcoming fuzzy models. In this way Figueroa and Soriano in [2] presented a complementary lecture in this topic.

The analyst must consider the computing cost of finding solutions by using genetic structures on IT2TSK systems. In addition to, Type-reduction methods increase the computing time since it is an expensive process of Interval Type-2 fuzzy sets, so it is an interesting field to be covered, moreover if any hybrid method is used to optimize fuzzy systems.

Finally, the IT2TSK model outperforms classical techniques, obtaining good identification results in a Volatile Time Series context. The presented algorithm synchronizes a FLS with successful results. The reader must keep in mind that genetic structures applied to complex inference systems require huge computing resources. In this way, we selected the IASCO Type-reduction algorithm proposed by Melgarejo in [1] due to it is faster than other algorithms.

Dedicatory

Juan Carlos Figueroa García dedicates his work to the memory of Maria Ninfa García Bautista (*Rest in peace my angel*).

References

[1] M. Melgarejo, H. Bernal, and K. Duran, "Improved iterative algorithm for computing the generalized centroid of an interval type-2 fuzzy set," in *2008 Annual Meeting of the North American Fuzzy Information Processing Society (NAFIPS)*, vol. 27. IEEE, 2008, pp. 1–6.

[2] J. C. Figueroa and J. J. Soriano, "A comparison of anfis, ann and dbr systems on volatile time series identification," in *2007 Annual Meeting of the North American Fuzzy Information Processing Society*, IEEE, Ed., vol. 26. IEEE, 2007, pp. 321–326.

[3] M. A. Melgarejo, "Implementing Interval Type-2 Fuzzy processors," *IEEE Computational Intelligence Magazine*, vol. 2, no. 1, pp. 63–71, 2007.

[4] J. M. Mendel and F. Liu, "Super-exponential convergence of the Karnik-Mendel algorithms for computing the centroid of an interval type-2 fuzzy set," *IEEE Transactions on Fuzzy Systems*, vol. 15, no. 2, pp. 309–320, 2007.

[5] N. N. Karnik and J. M. Mendel, "Centroid of a type-2 fuzzy set," *Information Sciences*, vol. 132, no. 1, pp. 195–220, 2001.

[6] J. Mendel, *Uncertain Rule-Based Fuzzy Logic Systems: Introduction and New Directions*, N. Upper Saddle River, Ed. Prentice Hall, 1994.

[7] J. Zeng and Z.-Q. Liu, "Enhanced Karnik-Mendel algorithms for Interval Type-2 fuzzy sets and systems," in *Annual Meeting of the North American Fuzzy Information Processing Society (NAFIPS)*, vol. 26. IEEE, 2007, pp. 184–189.

[8] G. Mendez and O. Castillo, "Interval type-2 tsk fuzzy logic systems using hybrid learning algorithm," in *Proceedings of IEEE FUZZ Conference*, vol. 27. IEEE, 2005, pp. 230–235.

[9] S. Goldfeld and R. Quandt, *Nonlinear Methods in Econometrics*, N. Holland, Ed. North Holland - Amsterdam, 1972.

[10] R. Engel, "Autoregressive conditional heteroscedasticity with estimates of the variance of united kingdom inflation," *Econometrica*, vol. 1, no. 50, pp. Pg 987–1007, 1982.

[11] P. Brockwell and R. Davis, *Time Series: Theory and Methods*, Springer, Ed. Springer-Verlag, 1998.

[12] ——, *Introduction to Time Series and Forecasting.*, Springer, Ed. Springer-Verlag, 2000.

[13] H. C. J. Thode, *Testing For Normality*, M. Drecker, Ed. Marcel Drecker Inc. New York, 2002.

[14] G. Box and G. Jenkins, *Time Series Analysis: Forecasting and Control.*, H. Day, Ed. Holden Day Publishing, 1970.

[15] A. Law and D. Kelton, *Simulation System and Analysis*, M. G. Hill, Ed. Mc Graw Hill International, 2000.

[16] G. M. Ljung and G. E. P. Box, "Information theory and the extension of the maximum likelihood principle." *Biometrika*, vol. 65, no. 1, pp. Pg 553–564, 1978.

[17] G. Schwarz, "Estimating the dimension of a model." *Annals of Statistics*, vol. 6, no. 1, pp. Pg 461–464, 1978.

[18] E. J. Hannan, "The estimation of the order of an arma process," *Annals of Statistics*, vol. 8, no. 5, pp. 1071–1081, 1981.

[19] E. J. Hannan and B. G. Quinn, "The determination of the order of an autoregression," *Journal of the Royal Statistical Society*, vol. 41, no. 1, pp. 190–195, 1979.

[20] G. M. Ljung and G. E. P. Box, "On a measure of lack of fit in time series models," *Biometrika*, vol. 65, no. 1, p. 297, Fall 1978.