Forecasting Venezuelan Caroni river flow through support vector machines.

Authors: Cesar Seijas 1*, Sergio Villazana 1*, Jorge Guevara 2*, Edilberto Guevara 2*
[1] Escuela de Ingeniería Eléctrica, Facultad de Ingeniería, Valencia, Venezuela, [2] Escuela de ingeniería Civil, Facultad de Ingeniería, Valencia, Venezuela.

email: cseijas@uc.edu.ve ; svillaza@uc.edu.ve;

Abstract

This paper models the time series formed by the monthly flows of Caroni River in Venezuela, since 1950 to 2003 using regression based on Support Vector Machines (SVR). The mentioned time series was preprocessed by extracting the trend and seasonal components through Independent Component Analysis (ICA) and the resulting stochastic series was modeled as a Nonlinear Autoregressive Moving Averaging process (NARMA), of order defined by Singular Value Decomposition (SVD), using SVR. The model was validated by obtaining a prediction error in the flow, lower than traditional statistical models. The results of this study demonstrate the strength of nonlinear computational models to predict river flows.

Keywords: *regression with support vector machines, flow modeling, independent component analysis, singular value decomposition.*

1. Introduction

This article explores the implementation of statistical processes and high performance computing, such as Independent Component Analysis (ICA) [13] and emergent computing algorithms, specifically Support Vector Machines (SVM) as a computational process to modeling time series. The particular application of this research will focus on modeling the flow variation of watersheds, considering the importance that this type of tool has on the problem of environmental conservation [8], [10], and hydroelectric power generation [4]. Caroni River´s basin is the most strategically important one in Venezuela, in which Guri hydroelectric plant is located; Guri plant is the biggest in Venezuela and one of the largest in the world. Guri plant, belonging to the state hydroelectric generation company CVG-EDELCA, supplies more than 65% of the total energy consumed in the country and over 80% of hydropower generated in Venezuela [4]. Given the vital importance to the nation, CVG-EDELCA has a permanent monitoring of the basin's flow, and consequently has a period of relatively long hydrological records. A model of Caroni River's flow hydrological time series is presented in this research. It consists of daily records covering a period between 1952 and 2003. This model is based on a nonlinear autoregressive structure, implemented using a SVM, whose inputs are the independent components obtained from the application of ICA algorithm to the time series above. In [9] the same time series are modeled using the classical linear statistical model ARIMA. SVM are algorithms of emergent computation developed by Vapnik and coworkers [28] who have demonstrated outstanding performance in applications of regression of nonlinear time series of physiological [16], [17] and hydrology [3] signals. On the other hand, ICA has been used in problems of predicting financial time series based on regression using artificial neural networks (ANN) [6].

2. Theoretical fundaments

2.1 Regression using SVM

A *SVM* used as a regressor or *SVR* (Support Vector Regression) estimates a non-linear function using a set of linear functions defined in a hiperdimensional space. That is, for a set of data points $G = \{(x_i, d_i)\}_i^n$ (where x_i is the input vector, d_i is the expected output and n is the number of data patterns), *SVR* approximates the regression function using:

$$y = f(x) = w\,\phi(x) + b \tag{1}$$

$\phi(x)$ is the hiperdimensional feature space where input space **x** is nonlinearly mapped. Coefficients w and b are estimated minimizing:

$$R_{SVMs}(C) = C\frac{1}{n}\sum_{i=1}^{n} L_s(d_i, y_i) + \frac{1}{2}\|w\|^2 \ , \tag{2}$$

where L_s is the linear ε-insensitive loss function

$$L_s(d, y) = \begin{cases} |d - y| - \varepsilon & |d - y| \geq \varepsilon \\ 0 & \end{cases} \tag{3}$$

In the loss function given by (2), the term $C(1/n)\sum_{i=1}^{n} L_\varepsilon(d_i, y_i)$ is the empirical (risk) error while the term $(1/2).\|w^2\|$ is the regularization term. Parameter C is known as regularized constant or capacity of the *SVM* and determines the trade-off between the empirical risk and the regularization term. In (3), ε is known as size of the hyperdimensional cylinder that wraps the function and is equivalent to the approximation accuracy on the training data points. C and ε are parameters to be set by the designer in a tuning process during the training stage of the *SVM*.

To obtain the estimates of $\boldsymbol{w}$ and b, (2) is transformed to (4), using slack variables ξ_i y $\xi_i^{(*)}$ that represents upper and lower limits in the system output as shown in Fig. 2, that is, minimizing:

$$R_{SVMs}(w, \xi_i, \xi_i^{(*)}) = \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{n}(\xi_i + \xi_i^{(*)}) \ , \tag{4}$$

subject to constraints:

$$d_i - \mathbf{w}\,\phi(\mathbf{x}) - b_i \leq \varepsilon + \xi_i$$

$$w\,\phi(x) + b_i - d_i \leq \varepsilon + \xi_i^{(*)}, \qquad \xi_i^{(*)} \geq 0 \tag{5}$$

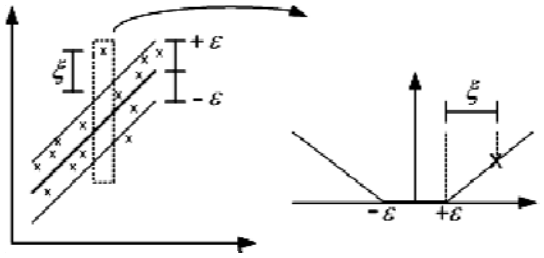

Figure 1. Pre-established error ε and limits ξ in the ε-insensitive function

Finally introducing the Lagrange multipliers α_i y α_i^* [28], the regression function (1) is transformed as indicated in (6):

$$f\left(x,\alpha_i,\alpha_i^*\right)=\sum_{i=1}^{n}\left(\alpha_i-\alpha_i^*\right)K\left(x,x_i\right)+b \quad (6)$$

$K\left(\mathbf{x_i},\mathbf{x_j}\right)$ is termed kernel function and α_i y α_i^* are the Lagrange multipliers meeting the constraints: $\alpha_i * \alpha_i^* = 0$, $\alpha_i \geq 0$ y $\alpha_i^* \geq 0$, $(i=1,\ldots,n)$, the latter are calculated maximizing the dual function of (6) having the form:

$$R\left(\alpha_i,\alpha_i^*\right)=\sum_{i=1}^{n}d_i\left(\alpha_i-\alpha_i^*\right)-\varepsilon\sum_{i=1}^{n}\left(\alpha_i+\alpha_i^*\right)-\frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n}\left(\alpha_i-\alpha_i^*\right)\left(\alpha_j-\alpha_j^*\right)K\left(x_i,x_j\right), \quad (7)$$

subject to: $\sum_{i=1}^{n}\left(\alpha_i-\alpha_i^*\right)=0$, $0\leq\alpha_i\leq C$, $0\leq\alpha_i^*\leq C$, $i=1,\ldots,n$

Only one certain number of these Lagrange multipliers, based on the Karush-Kuhn-Tucker (*KKT*) [28] quadratic programming conditions, will have values nonzero and the points or vector associated with them will have errors equal or greater than ε, these data points are called support vectors. It is evident from (6) that these support vectors define $f(\mathbf{x})$.

Kernel function is equal to inner product of two vectors x_i and x_j in the feature space $\phi(\mathbf{x}_i)$ and $\phi(\mathbf{x}_i)$, this is: $K\left(\mathbf{x_i},\mathbf{x_j}\right)=\phi\left(\mathbf{x_i}\right)*\phi\left(\mathbf{x_j}\right)$. The advantage of using kernel function, the computation is done in an arbitrary feature space without explicitly using $\phi(\mathbf{x})$. Any function that satisfies the Mercer's conditions [2] is candidate to be a kernel function, among which can mention, polynomial functions of the form $K(\mathbf{x},\mathbf{y}) = (\mathbf{x}*\mathbf{y} + 1)^d$, where d is the polynomial degree, Gaussian $K(\mathbf{x},\mathbf{y}) = exp(-1/\sigma^2(\mathbf{x} - \mathbf{y})^2)$, where σ is the widespread coefficient of the Gaussian, also known as radial basis function (*RBF*).

2.2 Independent Components Analysis

Given a set of N independent components s_j linear and non-overlapping Gaussian:

$$x_j = a_{j1}s_1 + a_{j2}s_2 + \cdots + a_{jn}s_n,\ j = 1,\ldots,N, \quad (8)$$

where it is assumed that each x_j (t) as well as each independent component $s_i(t)$ are samples of the corresponding random variables, a function of time, with zero mean.

Let be **X**, called mixed vector, the vector composed of random vectors x_j consisting of the superposition of random variables **S**, called source vector, consisting of random vectors s_i. Then the matrix model to all the following equation:

$$\mathbf{X} = \mathbf{A.S} \tag{9}$$

The above equation is known as Independent Component Analysis or ICA [13]. The problem is to find **A** and **S** from **X**, where it is assumed that **S** is independent and non-Gaussian distribution. The condition of non-Gaussian can be determined using higher order statistics (HOSA); in [18] it shows a "toolbox" for processing HOSA implemented in Matlab. Get "whitening" of the eigenvalues of covariance matrix:

$$\mathbf{VDV^T} = \mathbf{E}\left[\widehat{\mathbf{X}}\widehat{\mathbf{X}}^{\mathbf{T}}\right], \tag{10}$$

where **V** is the matrix of orthogonal eigenvectors and **D** the diagonal matrix of corresponding eigenvalues. The "whitening" is achieved by multiplication with the transformation matrix **P**:

$$\tilde{\mathbf{x}} = \mathbf{P}\hat{\mathbf{x}}$$

$$\mathbf{P} = \mathbf{V}\mathbf{D}^{-\frac{\mathbf{1}}{\mathbf{2}}}\mathbf{V}^{\mathbf{T}}, \tag{11}$$

Matrix to extract the independent components of **x** is **W** where:

$$\mathbf{W} = \widetilde{\mathbf{W}}\mathbf{P} \tag{12}$$

2.3 Gaussian and linear test

The Hinich algorithm [18] is a statistic test to verify signal linearity and/or Gaussian condition in signals. The mentioned algorithm is based on the detection of non skewness. It is basically supported on the fact that in a Gaussian process the cumulants whose order is higher than two are nil, hence the bi-spectre and the corresponding bi-coherence. Then, the null hypothesis of no Gaussian condition is established if the bi-specter is not zero. On the other hand, if the bi-coherence is not constant, it must be concluded that the process is non linear. A toolbox of free use (HOSA [18]), developed under mathematic software Matlab, implements the Hinich algorithm ("glstat" routine) by obtaining non biased consistent estimates of the bi-coherence, that is:

$$|\hat{bic}_{xxx}(f_1, f_2)|^2 = \frac{|\hat{S}c_{xxx}(f_1,f_2)|^2}{S_{2x}(f_1+f_2)S_{2x}(f_1)S_{2x}(f_2)} \tag{13}$$

3. Methodology and result analysis

This paragraph describes the methodology used to describe the time series designed by the daily flow of Caroni River, between 1952 and 2002. First, the nonexistent linearity and Gaussian condition of the series are determined. Then there is a segmenting of the series to get sub-series which exhibit linearity and non Gaussian condition. Third, these auto-regressive high order (10 is used as the initial pivotal value) term matrices are transformed (*AR(p)* model) due to the fact that the adequate model order has not yet been identified, so that they could be processed by ICA. Lastly, an *AR(p)* model [9] is developed, based on SVR. The inlet training matrix is formed by auto-regressive components derived from the independent components, and the objective vector is constituted by the future values (index: *p+1*) of the variable to be produced.

3.1 Time series

The time series to be analyzed is composed by daily hydrometric data registered between 1952 and 2002, on the flow of Caroní River basin, located to the South-West of Venezuela in Bolívar State. This basin feeds Guri dam, the main source of hydroelectric energy in Venezuela. The geographic coordinates are 3^o 40' and 8^o 40' N, 60^o 50' and 64^o 10' W and it has an extension of 96,000 Km^2. Figure 1 shows a Venezuelan map, locating the region where Caroní River basin is [16].

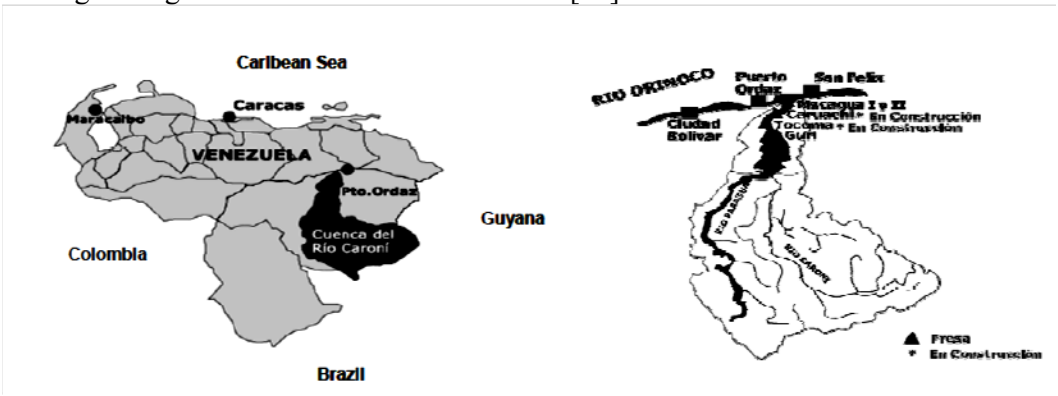

Figure 2. Location of Caroní River basin. (source: EDELCA 2004 [8])

The registered data show the daily flow between 1952 and 2002, which is a time series composed by 6,050 points with a daily frequency (dimension vector: 6206 X 1). Figure 2 shows the series under study.

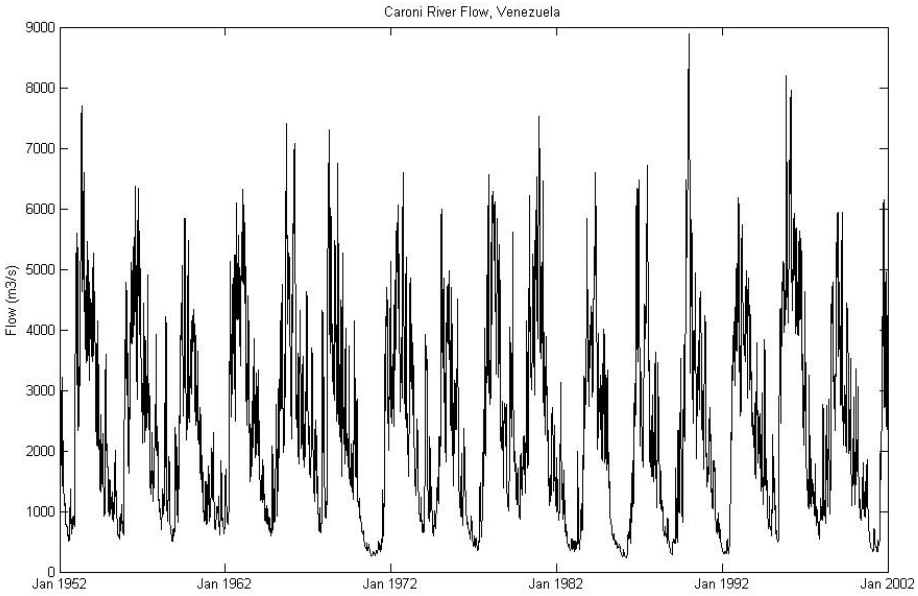

Figure 3. Daily flow of Caroní River basin

3.2 Processing

The time series was pre-processed taking its mean and normalizing its variance to the unit. Then, it was segmented in sub-series, using Hinich test [6]. The procedure is shown in Table 1. The mentioned procedure was applied to 50 sub-series having 1000 samples, in overlapping segments and already normalized.

Table 1. Gaussian and linearity test on time series

G	df	pfa
127.93±43.52	36	0
Rest	**λ**	**Rtheory**
7.47	5.52	6.51

This table summarizes, in the upper row, statistics of the Gaussian index, G, the degrees of freedom of the distribution χ^2, and predicted false alarm (p.f.a). The lower row contains the linearity test ("glstat" routine, toolbox HOSA [6]). The results show that the selected segments are non Gaussian and come from a linear system excited by white noise.

3.3 Organization of data

The processed sub-series are organized as auto-regressive arrangements *AR(p)*. The order *p* of the model that best represents any of the sub-series is not known at this point of the process. Therefore, a high order is assumed, as a pivotal element (in this case p=10 was used as pivot value).

3.4 ICA

Using toolbox FASTICAG (free academic use) [13], the initial ICA process of extraction of the eigenvector matrix was carried out ("Singular Value Decomposition", SVD [6]). The corresponding eigenvalues allowed the reduction of dimensionality of matrices in process, neglecting those of low power. Figure 3 shows that a dimensionality of order 3 retains almost all the signal power expressed in the first three self-values, so the order p=3 will be considered for the auto-regressive matrix model.

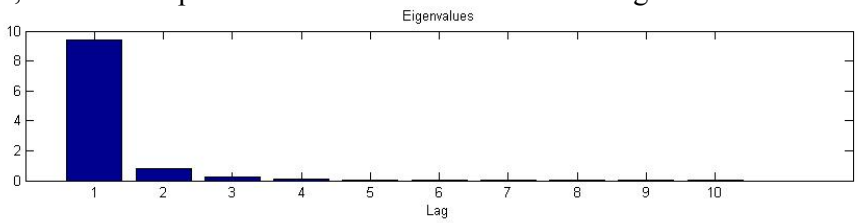

Figure 3. ICA eigenvalues

Figure 4 shows the three main components identified in the ICA process.

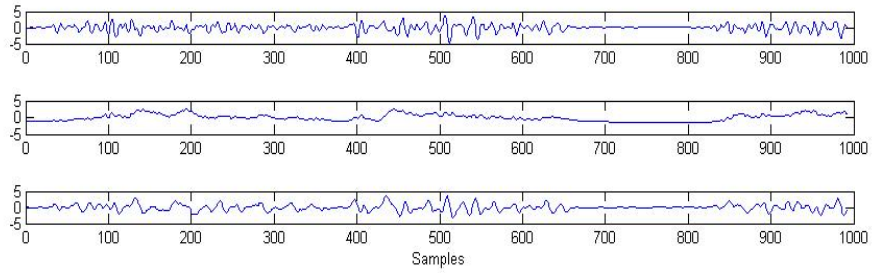

Figure 4. Source signals identified by ICA

4.5 SVR

Each matrix of order L x p (L: serie length - p), that belongs to each sub-series and obtained in the above step gave origin to two new matrices. These two new matrices contain 80% and 20% of randomly chosen rows (function "divideran", statistic toolbox-Matlab [19]); theses matrices are used as training and validation matrices, respectively, for SVR. The objective vector is the column vector concerning self-regressive terms constituted by the next value in the series o value to forecast, i.e. the forecasting horizon is established a step ahead. The tuning of the implemented SVR [21] concluded that the best performance kernel function was the radial base function (rbf), having similar parameters shown in the next table. FMSE (Forecasting Mean Square Error) is lower than other models of the same time series, implemented using ARIMA [9].

Table 2. Tuning parameters of SVR

σ	C	ε	T_e (sec)	NVS	FMSE
2.0	10^2	10^{-1}	87.2	437	7.8832

A comparative graph on the validation series vs. the simulation created by SVR is shown in Figure 5. This figure shows the approximation of SVR model to the series of validation data from the original processed series.

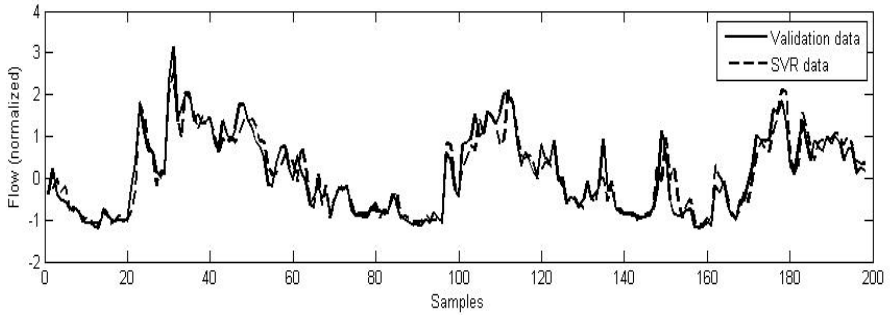

Figure 5. SVR vs. validation data

CONCLUSIONS:

The behavior of non Gaussian of time series was verified. Such series were derived from hydrographic records, hence the application of ICA to determine data to build training SVR matrices and implement self-regressive non linear models, NARMA type. The implemented SVR demonstrated a satisfactory behavior as backward agent, factor that shows the possibility of application of SVR in the modeling of hydrographic series. The best kernel function behavior occurred in the radial type function. The results shown in this paper supports the development of new accurate computing tools for forecasting hydrologic time series.

REFERENCES

[1] Akhtar M. K., G. A. Corzo, S. J. van Andel & A. Jonoski, *"River flow forecasting with artificial neural networks using satellite observed precipitation pre-processed with flow length and travel time information: case study of the Ganges river bas".* Hydrol. Earth Syst. Sci., 13, 1607–1618, 2009, www.hydrol-earth-syst-sci.net/13/1607/2009.

[2] Al-Zu'bi Y., A. Sheta & J Al-Zu'bi, *"Nile River Flow Forecasting based Takagi-Sugeno Fuzzy Model".* Journal of Applied Sciences 10 (4): 284-290, 2010.

[3] Clarke R., *"Hydrological Prediction in a non-stationary world"*. Hidrology, Earth and Systems Sciences.11-1, 408-414, www.hydrol-earth-syst-sci.net/11/408/ , 2007.

[4] Chacón I. & A. Leiro, *"Methodology for Setting a Reference System for Guri Dam Operation".* IEEE PES Transmission and Distribution Conference and Exposition Latin America, Venezuela, 2006.

[5] Dibike Y. B. & D. P. Solomatine, *"River Flow Forecasting Using Artificial Neural Networks".* EGS journal of Physics and Chemistry of the Earth, 2001.

[6] Górriz J., C.Puntonet, M. Salmerón & E.W. Lang, *"Time Series Prediction using ICA Algorithms".* IEEE International Workshop on Intelligent Data Acquisition and Advanced Computing Systems: Technology and Application, Lviv, Ukraine, 8-10 September 2003.

[7] Guenni L., E. Degryze & K. Alvarado, *"Análisis de la tendencia y la estacionalidad de la precipitación mensual en Venezuela".* Revista Colombiana de Estadística, volumen 31, no. 1, pp. 41 a 65, Junio 2008.

[8] Guevara E., "*The Influence of El Niño Phenomenon on the Climate of Venezuela*". Hydrology Days, 2006.

[9] Guevara E. & J. Guevara, *"Análisis estocástico de los caudales mensuales del río Caroní, Venezuelas".* XXIV Congreso Latinoamericano de Hidráulica, Punta del Este, Uruguay, noviembre 2010.

[10] Guevara E. & L. Alabano, *"Modelación de la función de distribución de frecuencias de los caudales maximos en la cuenca del Rio Caroni".* Revista Ingenieria UC, Vol. 13, Nro. 2, 2006.

[11] Han D., L. Chan & N. Zhu, *"Flood Forecasting using Support Vector Machines*". Journal of Hydroinformatics, 267, 09.4, 2007. Hidrology, Earth and Systems Sciences.11-1, 408-414, www.hydrol-earth-syst-sci.net/11/408/, 2007.

[12] Hastenrath S., L. Greischar, E. Colon & A. Gil, "*Forecasting the Anomalous Discharge of the Caroni River, Venezuela*". Journal of Climate, August 1999.

[13] Hyvärinen A. & E. Oja, "*Independent Component Analysis: Algorithms and Applications".* Neural Networks, 13(4-5):411-430, Neural Networks Research Centre, Helsinki University of Technology, Finland, 2000.

[14] Ismail S. R. Samsudin & A. Shabri, "*River Flow Forecasting: a Hybrid Model of Self Organizing Maps and Least Square Support Vector Machine".* Hydrol. Earth Syst. Sci. Discuss., 7, 8179–8212, 2010, www.hydrol-earth-syst-sci-discuss.net/7/8179/2010.

[15] Nayak P. C., K. P. Sudheer & K. S. Ramasastri, *"Fuzzy computing based rainfall–runoff model for real time flood forecasting".* Published online 16 September 2004 in Wiley InterScience www.interscience.wiley.com). DOI: 10.1002/hyp.5553*Hydrol. Process.* 19, 955–968 (2005).

[16] Seijas C., A. Caralli & S. Villazana, *"Estimation of Action Potential of the Cellular Membrane using Support Vectors Machines*". Proceedings of the 28th IEEE EMBS Annual International Conference, p 4200-4204, New York City, USA, Aug 30-Sept 3, 2006.

[17] Seijas C., A. Caralli & S. Villazana, *"Estimation of Brain Activity using Support Vector Machines"*. Proceedings of the 3rd IEEE EMBS International Conference on Neural Engineering, p 604-607, Kona, Kohala Coast, Big Island, Hawaii, USA, May 2-5, 2007.

[18] Swami A., J. Mendel & C. Nikias, "*Higher-Order Spectral Analysis Toolbox: User's Guide*". Version 6.0, The MathWorks Inc., Natick, USA, 2007.

[19] The MathWorks Team, "*Statistics Toolbox 6 User's Guide*". Version 6.0, The MathWorks Inc., Natick, USA, 2007.

[20] Vapnik V. N., "*The Nature of Statistical Learning Theory*", New York: Springer-Verlag, 1995.

[21] Villazana S. & G. Montilla, "*Un Toolbox para Procesamiento de Señales usando Máquinas de Vectores de Soporte*". Centro de Procesamiento de Imágenes, Facultad de Ingeniería, Universidad de Carabobo, Valencia, Venezuela, 2008.