Applied Mathematics, 2015, 6, 1611-1619
Published Online August 2015 in SciRes. http://www.scirp.org/journal/am
http://dx.doi.org/10.4236/am.2015.69143

Scientific Research Publishing

Forecasting Short Time Series with Missing Data by Means of Energy Associated to Series

Cristian Rodríguez Rivero[1], Julián Pucheta[1], Sergio Laboret[1], Daniel Patiño[2], Víctor Sauchelli[1]

[1]Department of Electronic Engineering, Universidad Nacional de Córdoba, Córdoba, Argentina
[2]Institute of Automatic, Universidad Nacional de San Juan, San Juan, Argentina
Email: crodriguezrivero@efn.uncor.edu, jpucheta@efn.uncor.edu, slaboret@yahoo.com.ar, victorsauchelli@gmail.com, dpatino@inaut.unsj.edu.ar

Received 18 January 2014; accepted 21 August 2015; published 24 August 2015

Abstract

In this work an algorithm to predict short times series with missing data by means energy associated of series using artificial neural networks (ANN) is presented. In order to give the prediction one step ahead, a comparison between this and previous work that involves a similar approach to test short time series with uncertainties on their data, indicates that a linear smoothing is a well approximation in order to employ a method for uncompleted datasets. Moreover, in function of the long- or short-term stochastic dependence of the short time series considered, the training process modifies the number of patterns and iterations in the topology according to a heuristic law, where the Hurst parameter *H* is related with the short times series, of which they are considered as a path of the fractional Brownian motion. The results are evaluated on high roughness time series from solutions of the Mackey-Glass Equation (MG) and cumulative monthly historical rainfall data from San Agustin, Cordoba. A comparison with ANN nonlinear filters is shown in order to see a better performance of the outcomes when the information is taken from geographical point observation.

Keywords

Artificial Neural Networks, Rainfall Forecasting, Energy Associated to Time Series, Hurst's Parameter

1. Introduction

Time series forecasting recently has a preponderant significance in order to know the best behavior of a system

How to cite this paper: Rodríguez Rivero, C., Pucheta, J., Laboret, S., Patiño, D. and Sauchelli, V. (2015) Forecasting Short Time Series with Missing Data by Means of Energy Associated to Series. *Applied Mathematics*, **6**, 1611-1619.
http://dx.doi.org/10.4236/am.2015.69143

in study such as the availability of estimated scenarios for water predictability [1], the rainfall forecast problem [2] [3] in some geographical points of Cordoba, the energy demand purposes [4], and the guidance of seedling growth [5]. For general feed-forward neural networks [6]-[8], the computational complexity of these solutions grows exponentially with the number of missing features. In this paper we describe an approximation for the problem of missing information that is applicable to a large class of learning algorithms [9] [10], including ANN's. One major advantage of the proposed solution is that the complexity does not increase with an increasing number of missing inputs. The solutions can easily be generalized to the problem of uncertain (noisy) inputs.

The problem of missing data poses a difficulty to the analysis and decision making processes which depend on this data, requiring methods of estimation which are accurate and efficient. Various techniques exist as a solution to this problem, ranging from data deletion to methods employing statistical and artificial intelligence techniques to impute for missing variables [11]. However, in this work, a linear estimation is employed making assumptions about the data that may not be true, affecting the quality of decisions made based on this data. The estimation of missing data in vector elements in real-time processing applications requires a system that possesses the knowledge of certain characteristics such as correlations between variables, which are inherent in the input space [12]. Those are taken from the Mackay Glass benchmark equation and cumulative historical rainfall whose forecast is simulated by a Monte Carlo approach employing ANN. The main contribution here is the design of a forecast system that uses incomplete data sets for tuning its parameters, and at the same time the historical recorded data are relatively short. The filter parameter is put in function of the roughness of the short time series, between its smoothness. In addition, this forecasting tool is intended to be used by agricultural producers to maximize their profits, avowing profit losses over the misjudgment of future movements to maximize their utilities. A one-layered feed-forward neural network, trained by the Levenberg-Marquardt algorithm is implemented in order to give the next 15 values. The paper is organized as follows: Section 2 introduces the data used for the algorithm. Section 3 describes an important issue to forecast with small datasets. Section 4 summarizes the implementation of the energy associated approach. In Section 5, prediction results are shown for a class of high roughness time series, namely short-term chaotic time series with a forecast horizon of 15 steps. Lastly, in Section 6 some discussion and conclusion are drawn.

2. Data Treatment

2.1. Rainfall Data and Neural Network Pattern Modeling

In this work the Hurst's parameter is used to determine the long-short term stochastic dependence of the rainfall time series. Besides, the neural network algorithm modifies in the learning process on-line the number of patterns, the number of iterations, and the number of filter's inputs. The definition of the Hurst's parameter appears in the Brownian motion from generalizing the integral to a fractional one. The Fractional Brownian Motion (*fBm*) is defined in the pioneering work by Mandelbrot [13] through its stochastic representation:

$$B_H(t)=\frac{1}{\Gamma\left(H+\frac{1}{2}\right)}\left(\int_0^{\infty}\left((t-s)^{H-\frac{1}{2}}-(-s)^{H-\frac{1}{2}}\right)+\int_0^{\infty}(t-s)^{H-\frac{1}{2}}\mathrm{d}B(s)\right)dB(s) \tag{1}$$

The *fBm* is self-similar in distribution and the variance of the increments is defined by:

$$Var\left(B_H(t)-B_H(s)\right)=\nu\left|t-s\right|^{2H} \tag{2}$$

where ν is a positive constant.

2.2. San Agustin Rainfall Data

The dataset chosen is from historical data 2004 to 2011 from San Agustin, located at Cordoba, Argentina shown in **Figure 1**. The original dataset (AGUS) used is incomplete and contains 51 data of cumulative monthly rainfall data, in which there are 14 months values incomplete resulting in a non-determinist series, respectively. This kind of behavior is difficult to predict because seasonality is not well-determined by few data. For the sake of making a fair prediction, a linear smoothing was employed to replace the incomplete data. This consists of av-

eraging on vertical column shows in **Figure 2**, the prior and posterior value that corresponds to the same year.

2.3. Mackay-Glass Time Series

The second benchmark of series is obtained from solution of the MG equation. This equation serves to model natural phenomena and has been used in earlier work to implement different methods of comparison to make forecast [14], which is explained by the time delay differential MG equation [15], defined as

$$\dot{y}(t) = \frac{\alpha y(t-\tau)}{1+y^{c}(t-\tau)} - \beta y(t) \tag{3}$$

where α, β varies and $c = 10$ are parameters and $\tau = 100$ is the delay time. According as τ increases, the solution turns from periodic to chaotic. Thereby, a time series with a random-like behavior is obtained, and the long-term behavior changes thoroughly by changing the initial conditions to obtain the stochastic dependence of the deterministic time series according to its roughness [16].

In this work the Hurst's parameter is used in the learning process to modify on-line the number of patterns, the number of iterations, and the number of filter's inputs of the ANN. This H serves to have an idea of roughness

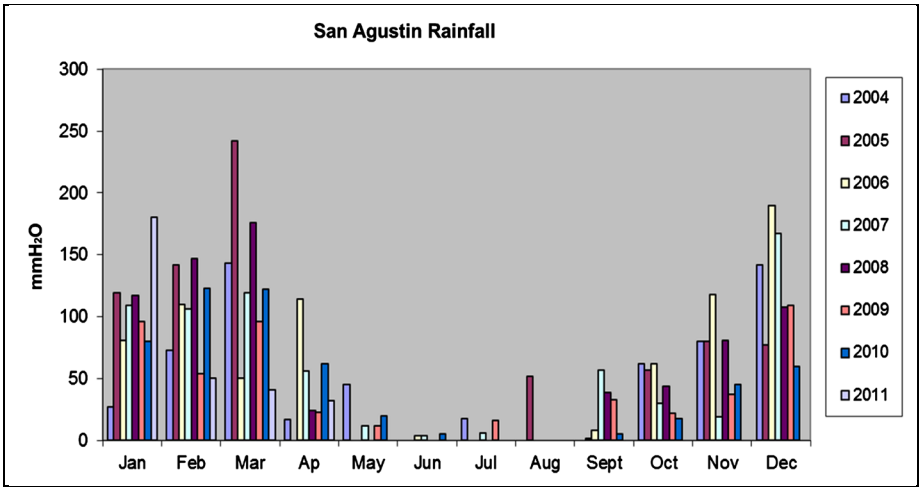


Figure 1. Cumulative monthly rainfall of san agustin (AGUS) with incomplete data, Cordoba, Argentina.

	Jan	Feb	Mar	Ap	May	Jun	Jul	Aug	Sept	Oct	Nov	Dec	Annual
2004	x	73	x	x	45	x	18	x	x	x	80	142	**358**
2005	119	142	242	x	x	x	x	52	x	57	80	77	**769**
2006	81	110	50	114	x	x	x	x	8	x	118	190	**671**
2007	109	106	119	56	12	x	x	x	57	30	x	167	**656**
2008	117	x	x	x	x	x	x	x	39	44	81	108	**389**
2009	96	54	96	23	x	x	16	x	33	22	x	109	**449**
2010	80	123	122	x	20	x	x	x	x	18	45	60	**468**
2011	180	50	41	32	x	x	x	x	x	x	x	x	**303**
Average	**112**	**98**	**112**	**56**	**16**	**0**	**16**	**52**	**34**	**34**	**81**	**122**	**732**

Figure 2. Average technique adopted to complete the rainfall dataset.

of a signal [17] [18] and the time series are considered as a trace of an *fBm* depending on the so-called Hurst parameter $0 < H < 1$ [19]. The MG benchmark chosen are called MG17 with $\tau = 17$, MG085, MG1.6 and MG1.9.

3. Problem Formulation

The main issue when forecasting a time series is how to retrieve the maximum of information from the available data [20]. In this case, the lack of data in the dataset is taken into account in order to predict one step ahead for the filter based on ANN. It is proposed to fill these empty values by using prior and posterior data. Four dataset are built following **Figure 2**. In the first one, the lack data is completed by taking the same ensemble of data of the past year. The second one by using the same ensemble of the next year, the third one is completed with zeros and lastly is filled in by averaging the prior and posterior year. The same analogy is used to construct MG17, MG0.85, MG1.9 and MG1.6 dataset solution of (3).

The coefficients of the ANNs filter are adjusted on-line in the learning process, by considering a criterion that modifies at each pass of the time series the number of patterns, the number of iterations and the length of the tapped-delay line, in function of the Hurst's value (H) calculated from the time series according to the stochastic behavior of the series, respectively.

In this work, the present value of the time series is used as the desired response for the adaptive filter and the past values of the signal serve as input of the adaptive filter [21]. Then, the adaptive filter output will be the one-step prediction signal. In the block diagram of the nonlinear prediction scheme based on an ANN filter is shown. Here, a prediction device is designed such that starting from a given sequence $\{x_n\}$ at time n corresponding to a time series it can be obtained the best prediction $\{x_e\}$ for the following sequence of 15 values. Hence, it is proposed a predictor filter with an input vector l_x, which is obtained by applying the delay operator, Z^{-1}, to the sequence $\{x_n\}$. Then, the filter output will generate x_e as the next value, that will be equal to the present value x_n. So, the prediction error at time k can be evaluated as

$$e(k) = x_n(k) - x_e(k) \tag{4}$$

4. Proposed Approach to Calculate the Energy Associated of Series

Approximation by Primitive of Integration

The area resulting of integrating the data time series of MG and rainfall data series is the primitive, that is obtained by considering each value of time series its derivate [22];

$$\int_{t_k}^{t_{k+1}} y_t \mathrm{d}t \cong y_t\left(t_{k+1} - t_k\right) \tag{5}$$

where y_t is the original value time series. The area approximation by its periodical primitive is:

$$I_{t_n} = \int_{t_n}^{t_{n+p}} y_t \mathrm{d}t = Y_t\Big|_{t_n}^{t_{n+p}}, n = 1, 2, \cdots, N. \tag{6}$$

During the learning process, those primitives are calculated as a new entrance to the ANN, in which the prediction attempts to even the area of the forecasted area to the primitive real area predicted. The real primitive integral is used in two instances, firstly from the real time series an area is obtained and run by the algorithm proposed. The H parameter from this time series is called H_A. On the other hand, the data time series is also forecasted by the algorithm, so the H parameter from this time series is called H_S. Finally, after each pass the number of inputs of the nonlinear filter is tuned—that is the length of tapped-delay line, according to the following heuristic criterion. After the training process is completed, both sequences—$\{\{I_n\},\{I_e\}\}$ and $\{\{y_n\},\{y_e\}\}$—in accordance with the hypothesis that should have the same H parameter. If the error between H_A and H_S is greater than a threshold parameter θ the value of l_x is increased (or decreased), according to $l_x \pm 1$. Explicitly,

$$l_x = l_x + sign(x) \tag{7}$$

Here, the threshold θ was set about 1%.

5. Prediction Results

5.1. Generations of Areas from Benchmark

Primitives of time series are obtained from sampling the MG equations with parameters shown in **Table 1**, with $\tau = 100$, $c = 10$ and varying β, α. This collection of coefficients was chosen to generate time series whose H parameters vary between 0 and 1 (**Table 2**). In fact, the chosen ones were selected in accordance with their roughness.

5.2. Performance Measure for Forecasting

In order to test the proposed design procedure of the ANN-based nonlinear predictor, an experiment with time series obtained from the MG solution was performed. The performance of the filter is evaluated using the Symmetric Mean Absolute Percent Error (*SMAPE*) proposed in the most of metric evaluation, defined by

$$SMAPE_S = \frac{1}{n}\sum_{t=1}^{n}\frac{|X_t - F_t|}{(X_t + F_t)/2}\times 100 \tag{8}$$

where t is the observation time, n is the size of the test set, **s** is each time series, X_t and F_t are the actual and the forecasted time series values at time t respectively. The *SMAPE* of each series s calculates the symmetric absolute error in percent between the actual X_t and its corresponding forecast value F_t, across all observations t of the test set of size n for each time series **s**.

5.3. Forecasting Results

Each time series is composed by samples of MG solutions and San Agustin rainfall time series. Three classes of data sets are used. The first one is the original time series used by the algorithm to train the predictor filter, which comprises 35 values. The next one is the primitive obtained by integrating the original time series data. The last one is used to compare if the forecast is acceptable or not, in which the last 15 of 50 values can be used to validate the performance of the prediction system. A comparison of roughness measured by the H parameter is made between AGUS rainfall and MG series.

The Monte Carlo method was used to forecast the next 15 values from San Agustin rainfall series (AGUS), MG085, MG1.6, MG1.9 and MG17 time series and their primitive. Such outcomes are shown from **Figure 3** to **Figure 7**. The plot shown in **Figure 3(a)**, **Figure 4(a)**, **Figure 5(a)**, **Figure 6(a)** and **Figure 7(a)** are from H dependent ANN predictor filter. **Figure 3(b)**, **Figure 4(b)**, **Figure 5(b)**, **Figure 6(b)** and **Figure 7(b)** are obtained by the energy associated approach.

The algorithm achieves the long or short term stochastic dependence measured by the Hurst parameter in order to make more precisely the prediction. The forecasted time series area is put as a new entrance to the ANN and serves to be compared with the real primitive obtained of the time series.

The figures show a class of high roughness time series selected from a benchmark of MG Equation and compared with AGUS rainfall series. These are classified by their statistically dependency, so the algorithm is adjusted by depending on the H parameter. At **Table 3** and **Table 4** shows a good performance seen from the

Table 1. Parameters to generate the mg times series.

Series No.	β	α	c	H
MG0.85	0.85	20	10	0.23
MG1.6	1.6	20	10	0.26
MG1.9	1.6	30	10	0.47

Table 2. Parameter of San Agustin rainfall series.

Series No.	H
AGUS	0.28

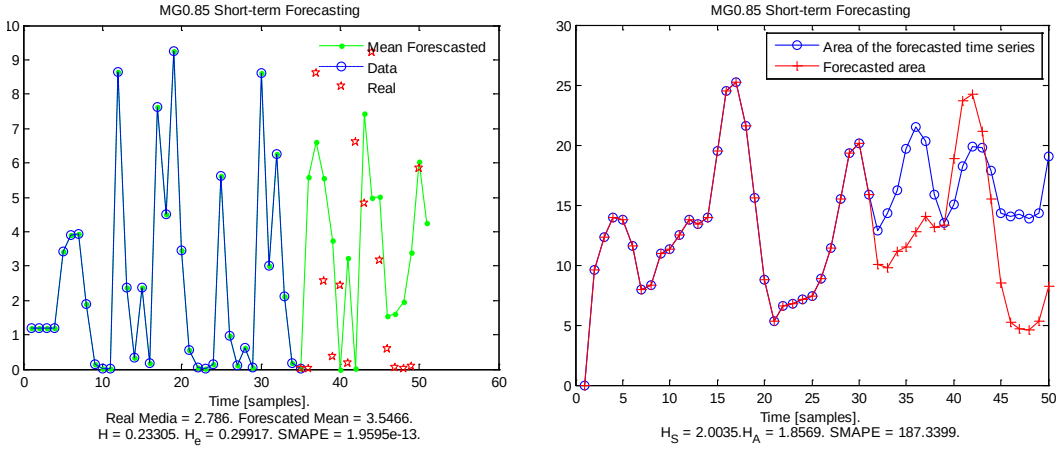


Figure 3. Non-linear autoregressive predictor filter. (a) *H* dependent neural network algorithm for MG085; (b) Energy asscociated approach for MG085.

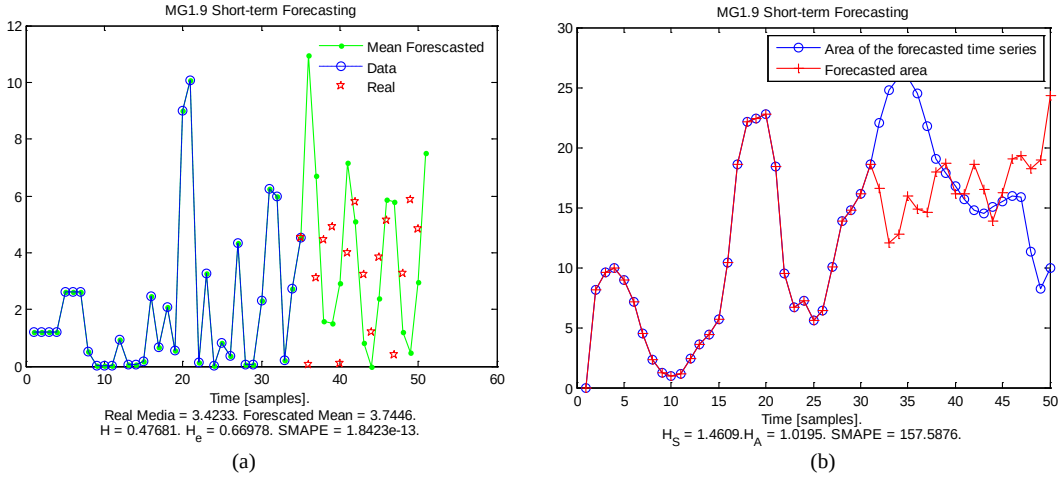


Figure 4. Non-linear autoregressive predictor filter. (a) *H* dependent neural network algorithm for MG1.9; (b) Energy asscociated approach for MG1.9.

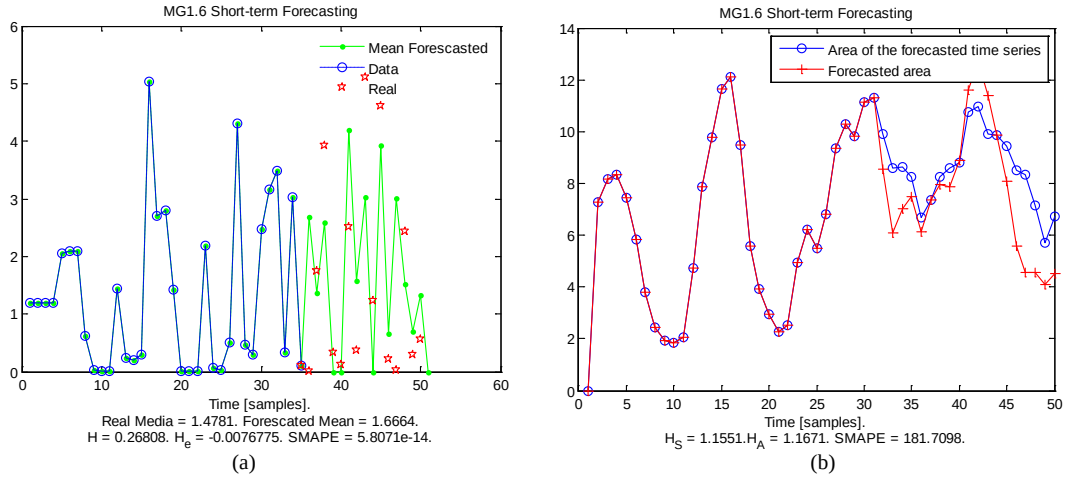
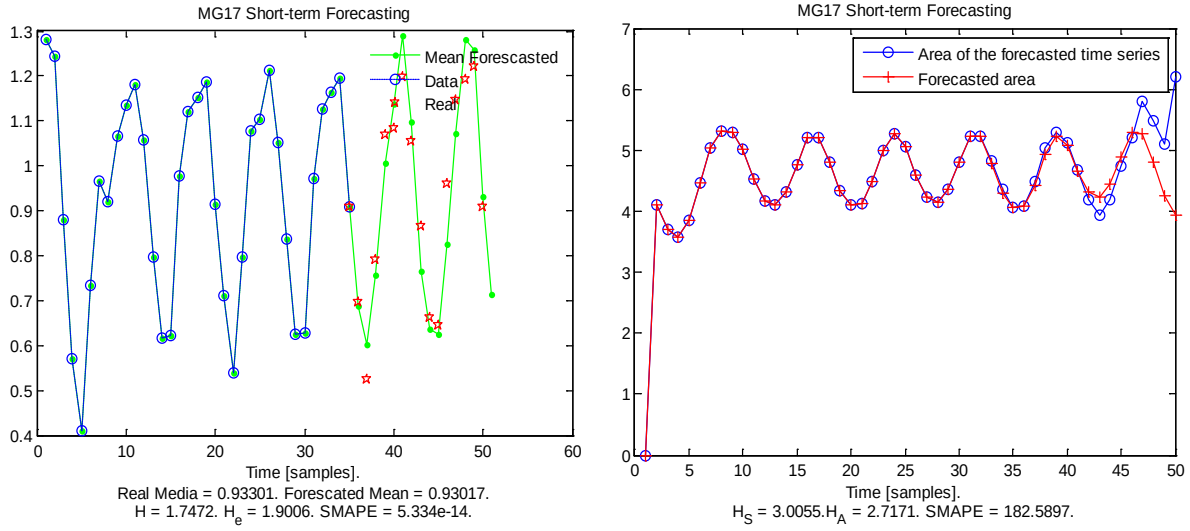
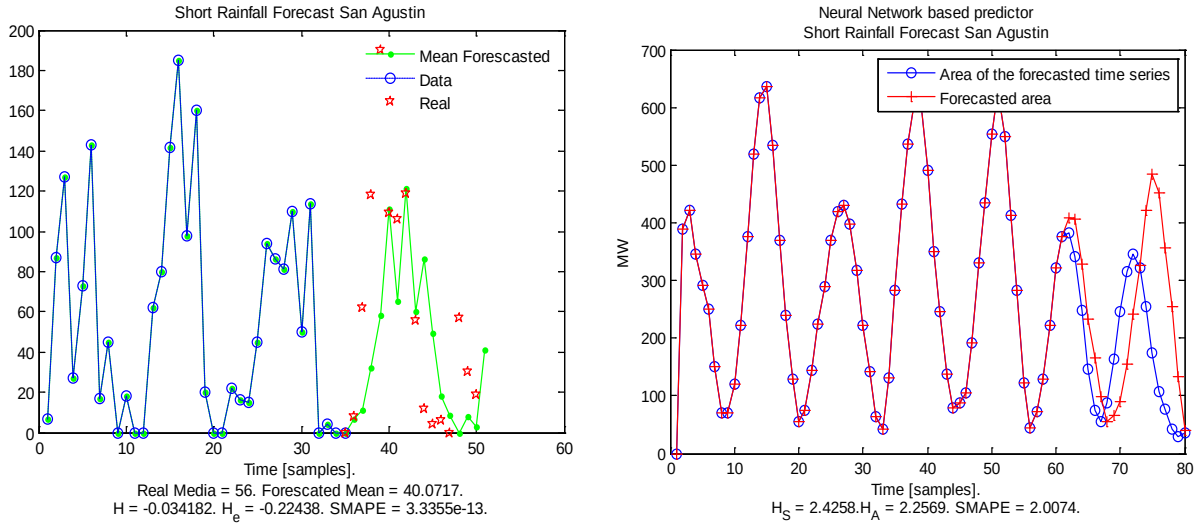


Figure 5. Non-linear autoregressive predictor filter. (a) *H* dependent neural network algorithm for MG1.6; (b) Energy asscociated approach for MG1.6.



(a) (b)

Figure 6. Non-linear autoregressive predictor filter. (a) *H* dependent neural network algorithm for MG17; (b) Energy asscociated approach for MG17.



(a) (b)

Figure 7. Non-linear autoregressive predictor filter. (a) *H* dependent neural network algorithm for AGUS rainfall series; (b) Energy asscociated approach for AGUS rainfall series.

Table 3. Comparisons obtained by the neural network *H* dependant predictor filter.

Series No.	*H*	H_e	SMAPE
Figure 3(a)	0.23	0.29	1.95e−13
Figure 4(a)	0.47	0.66	1.84e−13
Figure 5(a)	0.26	0.007	5.8e−14
Figure 6(a)	1.74	1.90	5.33e−14
Figure 7(a)	0.03	0.22	3.33e−13

Table 4. Comparisons obtained by the energy associated approach.

Series No.	H_S	H_A	SMAPE
Figure 3(b)	2.00	1.85	187.33
Figure 4(b)	1.46	1.01	157.58
Figure 5(b)	1.15	1.16	181.70
Figure 6(b)	3.00	2.63	182.58
Figure 7(b)	2.42	2.25	2.00

SMAPE index of in AGUS rainfall series and MG1.6 series when they take into accounts the roughness of the series considering the use of the stochastic dependence measured by the *H* parameter.

6. Discussion and Conclusions

In this work, short-term rainfall time series prediction with incomplete data by means of energy associated of series was presented. The learning rule proposed to adjust the ANNs weights is based on the Levenberg-Marquardt method and energy associated to series as a new input. Likewise, in function of the short-term stochastic dependence of the time series evaluated by the Hurst parameter *H*, the performance of the proposed filter shows that even the short dataset is incomplete, besides a linear smoothing technique employed, the prediction is almost good. The major result shows that the predictor system based on energy associated to series has an optimal performance from several samples of MG equations and, in particular, MG1.6 and AGUS rainfall time series. These were considered as a path of a fractional Brownian motion [19] whose *H* parameter measured is a high roughness signal, which is assessed by H_S and H_A, respectively. Although the comparison was only performed on ANN-based filters, the experimental results confirm that the energy associated to series method can predict short-term rainfall time series more effectively in terms of SMAPE indices when compared with other existing forecasting methods in the literature.

This approach encourages forecasting meteorological variables such as moisture soil series, daily and hour rainfall and water runoff when the observations are taken from a single point.

Acknowledgements

This work was supported by Universidad Nacional de Córdoba (UNC), FONCYT-PDFT PRH No. 3 (UNC Program RRHH03), SECYT UNC, Universidad Nacional de San Juan—Institute of Automatics (INAUT), National Agency for Scientific and Technological Promotion (ANPCyT) and Departments of Electronics—Electrical and Electronic Engineering—Universidad Nacional of Cordoba.

References

[1] Vamsidhar, E., Varma, K.V.S.R.P., Rao, P. and Satapati, R. (2010) Prediction of Rainfall Using Backpropagation Neural Network Model. *International Journal on Computer Science and Engineering*, **2**, 1119-1121.

[2] Wu, C.L. and Chau, K.W. (2013) Prediction of Rainfall Time Series Using Modular Soft Computing Methods. *Engineering Applications of Artificial Intelligence*, **26**, 997-1007. http://dx.doi.org/10.1016/j.engappai.2012.05.023

[3] Rodríguez Rivero, C., Herrera, M., Pucheta, J., Baumgartner, J., Patiño, D. and Sauchelli, V. and Laboret, S. (2013) Time Series Forecasting Using Bayesian Method: Application to Cumulative Rainfall. *IEEE Latin America Transactions*, **11**, 359 364. http://dx.doi.org/10.1109/TLA.2013.6502830

[4] Gonzalez-Romera, E., Jaramillo-Moran, M.A. and Carmona-Fernandez, D. (2006) Monthly Electric Energy Demand Forecasting Based on Trend Extraction. *IEEE Transactions on Power Systems*, **21**, 1946-4953. http://dx.doi.org/10.1109/TPWRS.2006.883666

[5] Pucheta, J., Patiño, H., Schugurensky, C., Fullana, R. and Kuchen, B. (2007) Optimal Control Based-Neurocontroller to Guide the Crop Growth under Perturbations. Dynamics of Continuous, Discrete And Impulsive Systems Special Volume Advances in Neural Networks-Theory and Applications. *DCDIS A Supplement, Advances in Neural Networks*, **14**, 618-623.

[6] Zhang, G., Patuwo, B.E. and Hu, M.Y. (1998) Forecasting with Artificial Neural Networks: The State of Art. *Journal*

of International Forecasting, **14**, 35-62. http://dx.doi.org/10.1016/S0169-2070(97)00044-7

[7] Pucheta, J., Rodríguez Rivero, C.M., Herrera, M., Salas, C., Patiño, D. and Kuchen, B. (2011) A Feed-Forward Neural Networks-Based Nonlinear Autoregressive Model for Forecasting Time Series. *Revista Computación y Sistemas*, *Centro de Investigación en Computación-IPN*, **14**, 423-435.

[8] Khosravi, A., Nahavandi, S., Creighton, D. and Atiya, A.F. (2011) Comprehensive Review of Neural Network-Based Prediction Intervals and New Advances. *IEEE Transactions on Neural Networks*, **22**, 1341-1356.

[9] Bishop, C. (2006) Pattern Recognition and Machine Learning. Springer, Boston.

[10] Bishop, C. (1995) Neural Networks for Pattern Recognition. University Press, Oxford.

[11] Tresp, V. and Hofmann, R. (1998) Nonlinear Time-Series Prediction with Missing and Noisy Data. *Neural Computation*, **10**, 731-747. http://dx.doi.org/10.1162/089976698300017728

[12] Markovsky, I., Willems, J.C. and De Moor, D. (2005) State Representation from Finite Time Series. *Proceedings of the 44th IEEE Conference on Decision and Control*, *and the European Control Conference* 2005, Seville, 12-15 December 2005, 832-835.

[13] Mandelbrot, B.B. (1983) The Fractal Geometry of Nature. W. H. Freeman, San Francisco.

[14] Rodriguez Rivero, C., Pucheta, J., Patiño, H., Baumgartner, J., Laboret, S. and Sauchelli, V. (2013) Analysis of a Gaussian Process and Feed-Forward Neural Networks Based Filter for Forecasting Short Rainfall Time Series. 2013 *International Joint Conference on Neural Networks*, Dallas, 4-9 August 2013, 1-6. http://dx.doi.org/10.1109/IJCNN.2013.6706741

[15] Glass, L. and Mackey, M.C. (1998) From Clocks to Chaos, the Rhythms of Life. Princeton University Press, Princeton.

[16] Pucheta, J., Patiño, H.D. and Kuchen, B. (2007) Neural Networks-Based Time Series Prediction Using Long and Short Term Dependence in the Learning Process. *International Symposium on Forecasting* (*ISF*'07), NN3 Forecasting Competition, New York.

[17] Abry, P., Flandrin, P., Taqqu, M.S. and Veitch, D. (2003) Self-Similarity and Long-Range Dependence through the Wavelet Lens. In: Doukhan, P., Oppenheim, G. and Taqqu, M., Eds., *Theory and Applications of Long-Range Dependence*, Birkhäuser, 527-556.

[18] Flandrin, P. (1992) Wavelet Analysis and Synthesis of Fractional Brownian Motion. *IEEE Transactions on Information Theory*, **38**, 910-917. http://dx.doi.org/10.1109/18.119751

[19] Dieker, T. (2004) Simulation of Fractional Brownian Motion. The Netherlands MSc Theses, University of Twente, Enschede.

[20] Pucheta, J., Patino, D. and Kuchen, B. (2009) A Statistically Dependent Approach for the Monthly Rainfall Forecast from One Point Observations. In: Li, D. and Zhao, C., Eds., *Computer and Computing Technologies in Agriculture II*, Vol. 2, IFIP Advances in Information and Communication Technology, Vol. 294, Springer, Boston, 787-798.

[21] Pucheta, J., Rodríguez Rivero, C., Herrera, M., Salas, C., Sauchelli, V. and Patiño, H.D. (2012) Non-Parametric Methods for Forecasting Time Series from Cumulative Monthly Rainfall. In: Martín, O.E. and Roberts, T.M., Eds., *Rainfall*: *Behavior*, *Forecasting and Distribution*, Nova Science Publishers, Inc., New York.

[22] Rodríguez Rivero, C., Herrera, M., Pucheta, J., Baumgartner, J., Patiño, D. and Sauchelli, V. (2012) High Roughness Time Series Forecasting Based on Energy Associated of Series. *Journal of Communication and Computer*, **9**, 576-586.