Monitoring dam performance through neural networks and the Box Jenkins approach

J.L.C. Gutierrez[1], C. Romanel[1]

Summary

In this paper, artificial neural networks (ANN) and the Box Jenkins approach [1] are used for modeling time series consisted of a sequence of water discharges, measured along several years of observation, through the foundation of a large Brazilian dam (Corumba-I dam). Results indicated that the ANN technique could be a powerful tool for early detection of abnormal conditions during operation of dams. In particular, the application of neural networks yielded quite useful water discharge forecasts since conventional methods of analysis would require three-dimensional models and a detailed investigation of the difficult subsoil conditions at Corumba-I dam.

Box Jenkins approach

The field of statistics concerned with analysis of data possessing spatial and temporal dimensions is known as time series analysis. Forecast of future developments is the most widespread application of time series analysis. Quite often a classical statistical approach known as the AR[p] model is used, based on the linear autoregressive equations

$$Z(t) = \sum_{i=1}^{p} \alpha_i Z(t-i) + \varepsilon(t) = \phi_L\big(Z(t-1),\ldots,Z(t-p)\big) + \varepsilon(t) \qquad (1)$$

where the prediction of variable Z at time t depends on a linear combination of p previous series observations, including the noise term ε.

Finding an adequate AR[p] model means selecting the p appropriate terms and estimating the corresponding α_i coefficients through a least-square optimization technique. This method is rather limited, since it assumes a linear relationship among the sequence elements and it is based on the hypothesis that the times series is stationary, i.e. the mean and the standard deviation of the measured observations do not vary over time.

Another approach for modeling time series is to assume the series being generated through a linear combination of q noise signals, referred in the literature as the *moving average* or the MA[q] model, or a combination of the AR[p] and MA[q] components giving rise to the so-called ARMA[p,q] models. The aforementioned limitations of the autoregressive AR[p] model, concerning restrictions on linearity and stationarity of the

[1] Pontifical Catholic University of Rio de Janeiro (PUC-Rio), Department of Civil Engineering, Rio de Janeiro – RJ, Brazil, romanel@civ.puc-rio.br

phenomenon being modeled, are also applicable to the MA[q] and ARMA[p,q] time series models.

The time series for Corumbá-I dam consisted of 130 patterns, measured at time intervals of 15 days, beginning in March 1997 and extending to December 2002. Two data sets were considered: the first one composed of 104 samples, used during the training phase, and the second data set (26 patterns) reserved for the validation process.

The null of stationarity of the time series was determined with basis on the null of a unit root [2], while the sample autocorrelation function and the sample partial correlation function indicated that the sequence could be conveniently described by a linear autoregressive AR[1] model. Results for the univariate forecast performance are presented in table 1, while figures 1 and 2 compare the predicted and actual values for the training and validation phases.

There are a number of error measurements that allow comparison between predicted and observed values. The two most commonly used are the Root Mean-Square Error (RMSE) and the Mean Absolute Percentage Error (MAPE), considering f_i the predicted result, t_i the actual value and N the number of patterns.

$$RMSE = \sqrt{\frac{\sum_{i=1}^{N}(f_i - t_i)^2}{N}} \qquad MAPE = \frac{1}{N}\sum_{i=1}^{N}\left|\frac{f_i - t_i}{t_i}\right| \times 100\% \tag{2a}$$

The disadvantage of RMSE is that it is sensitive to outliers and the drawback of the MAPE is that it puts a heavier penalty on predictions that exceed the actual values than on those that fall behind. To help judgement about the performance of a regression model, Theil [3] proposed the following index ,

$$\mathrm{U-Theil} = \frac{\sqrt{\frac{1}{N}\sum_{i=1}^{N}(f_i - t_i)^2}}{\sqrt{\frac{1}{N}\sum_{i=1}^{N}f_i^2} + \sqrt{\frac{1}{N}\sum_{i=1}^{N}t_i^2}} \tag{2b}$$

Values close to zero mean a good forecast accuracy while large values indicate a rather easy interpretation of the series behavior. The smaller the U-Theil index, the better the model performs compared to a naive prediction of no-change over the sample data.

Table 1 – Forecast errors using AR[1] model.

Training			Validation		
MAPE (%)	RMSE	U-THEIL	MAPE (%)	RMSE	U-THEIL
35.440	40.863	0.950	10.867	34.210	1.022

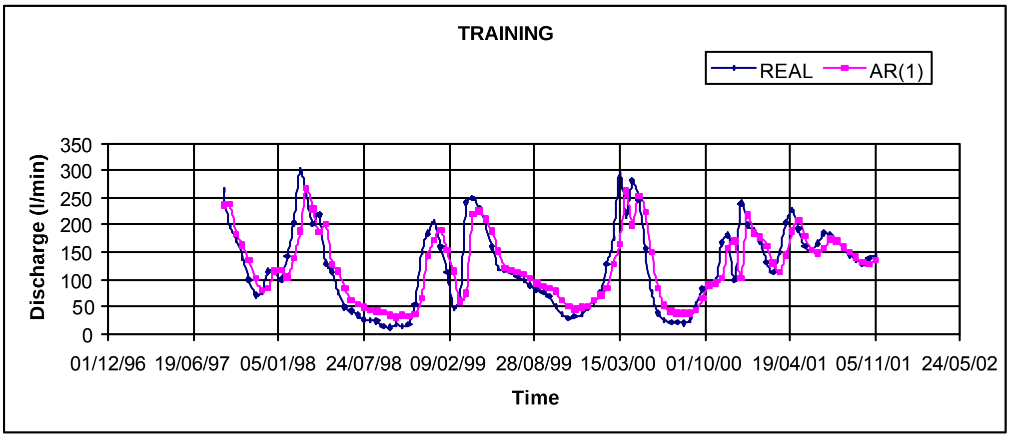

Figure 1: Comparison between real and predicted values during the training phase of the AR[1] model for Corumbá-I dam.

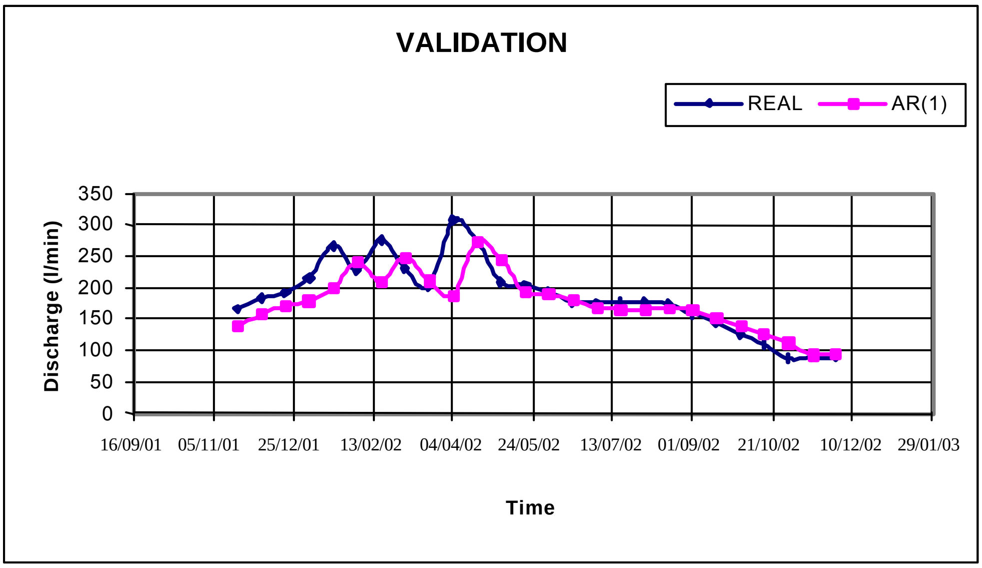

Figure 2: Comparison between real and predicted values during the validation phase of the AR[1] model for Corumbá-I dam.

Artificial neural networks - ANN

A multilayer feedforward neural network could be used to replace the linear function ϕ_L in equation (1) by a non-linear function ϕ_{NL} estimated from a learning technique such as the backpropagation or the conjugent gradient methods. Making ϕ_{NL} dependent on the p previous sequence elements is equivalent to use p input units being fed with p adjacent sequence elements [4].

The sequence of water discharges through the foundation of Corumba-I dam was also modelled using a multilayer feedforward network, a very common network architecture composed by an input layer of several source nodes, a hidden layer containing 1 to 10 neurons and an output layer with 1 neuron (the water discharge forecast). The algorithm for training the neural networks was the backpropagation algorithm, with the descendent gradient method. Several ANN were trained and tested in order to assess their influence on the network model and, in this way, to find out the most adequate neural netwok for the flow conditions observed in field. Each neural network was initialized 5 times, aiming to reduce the effects of non-optimum local minimum. The stop criterion for the training algorithm was the early stop criterion [5].

Eight ANN were organized according to the number and type of input parameters, as shown in table 2, where the symbol Q represents the measured water discharge, L the reservoir water level and P the piezometer readings. The output value refers to water discharge at time t, being Δt equal to 15 days, i.e. the time interval between consecutive measurements. Each configuration was trained and tested several times, admitting 2, 3, 5, 8 and 10 neurons in the hidden layer. All networks have the discharge value $Q_{t-\Delta t}$ as an input parameter, given the previous (and good) results obtained with the AR[1] model.

Table 3 presents a summary of the computed ANN errors, where the values a/b/c in the topology column mean the number of input parameters, the number of hidden neurons and the number of output neurons, respectively. ANN1 with topology 1/3/1 was chosen as the "best" ANN for water discharge forecasts. This choice was made based on the magnitude of the computed validation errors as well as on the concept of parsimony, since ANN1 is a model that yields good predictions with a small number of neurons.

Figures 3 and 4 compare the actual measured values with respective forecasts using ANN1 1/3/1.

Conclusion

ANN were used to control performance monitoring of the Corumba -I dam. Results indicate that this model can be successfully extended to other dams where the early detection of abnormal conditions of water flow is of fundamental importance.

The Box-Jenkins approach is a helpful tool for preliminary analyses, permitting *a priori* estimates on the required number of sequence elements for the ANN models. When ANNs involve several variables as input parameters, such as temperature, reservoir water levels, piezometric readings, etc., it is clear that the choice of the number and type of parameters must be made based on a good understanding of the physical phenomena being modeled.

Table 2 - Eight ANNs with different number and type of input parameters.

Neural Network	Input Parameters
ANN1	$Q_{t-\Delta t}$
ANN2	$Q_{t-\Delta t}$, $Q_{t-24\Delta t}$, $Q_{t-25\Delta t}$, $Q_{t-26\Delta t}$
ANN3	$Q_{t-\Delta t}$, $Q_{t-23\Delta t}$, $Q_{t-24\Delta t}$, $Q_{t-25\Delta t}$
ANN4	$Q_{t-\Delta t}$, $L_{t-\Delta t}$
ANN5	$Q_{t-\Delta t}$, $Q_{t-24\Delta t}$, $L_{t-\Delta t}$, $L_{t-24\Delta t}$
ANN6	$Q_{t-\Delta t}$, $Q_{t-24\Delta t}$, $Q_{t-25\Delta t}$, $Q_{t-26\Delta t}$, $L_{t-\Delta t}$, $L_{t-24\Delta t}$, $L_{t-25\Delta t}$, $L_{t-26\Delta t}$
ANN7	$Q_{t-\Delta t}$, $P_{t-\Delta t}$
ANN8	$Q_{t-\Delta t}$, $Q_{t-16\Delta t}$, $Q_{t-17\Delta t}$, $Q_{t-18\Delta t}$, $Q_{t-19\Delta t}$, $P_{t-\Delta t}$, $P_{t-16\Delta t}$, $P_{t-17\Delta t}$, $P_{t-18\Delta t}$, $P_{t-19\Delta t}$

Table 3- Summary of the neural networks performance.

ANN	Topology	Training error			Validation error		
		MAPE (%)	RMSE	U-THEIL	MAPE (%)	RMSE	U-THEIL
1	**1/3/1**	**33.559**	**41.452**	**0.962**	**8.165**	**27.921**	**0.852**
2	4/3/1	55.490	44.774	1.041	7.098	29.039	0.839
3	4/3/1	44.590	38.672	0.904	7.277	30.114	0.871
4	2/8/1	25.721	34.313	0.800	9.731	30.638	0.915
5	8/3/1	41.462	40.481	0.938	10.190	30.897	0.893
6	4/10/1	33.537	35.453	0.833	10.302	30.377	0.855
7	2/5/1	31.304	36.558	0.851	9.041	32.405	0.966
8	10/10/1	33.990	33.846	0.810	8.854	34.773	0.991

Reference

1 Box, G.E.; Jenkins, G.M. (1970): *Time Series Analysis*, Holden-Day.

2 Dickey, D. A; Fuller, W.A (1979): "Distribution of the Estimates for Autoregressive Time Series with a Unit Root", *Journal of the American Statistical Association*, vol. 74, pp. 427-431.

3 Theil, H.(1966): *Applied Economic Forecasting*, North-Holland, Amsterdam.

4 Dorffner, G. (1996): "Neural Networks for Time Series Processing", *Neural Network World*, pp. 447-468.

5 Haykin, S. (1994): *Neural Networks: A Comprehensive Foundation*, MacMillan.

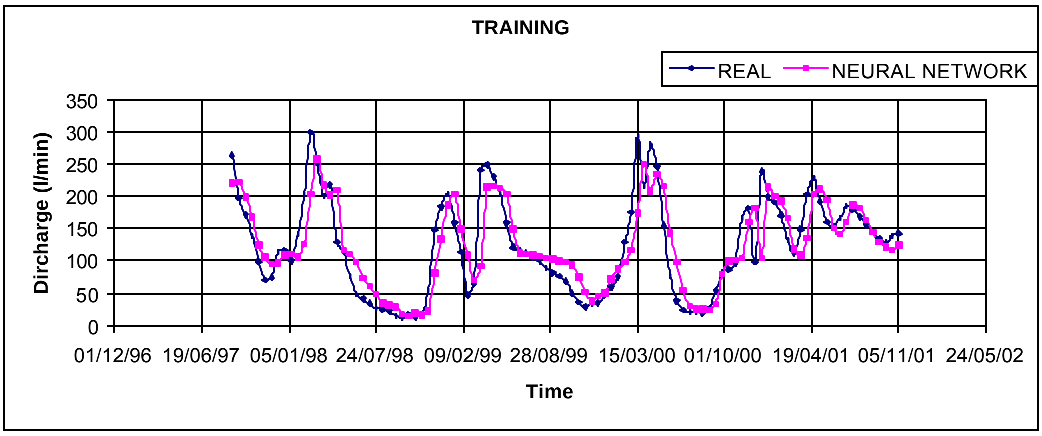

Figure 3: Comparison between real and predicted values during the training phase.

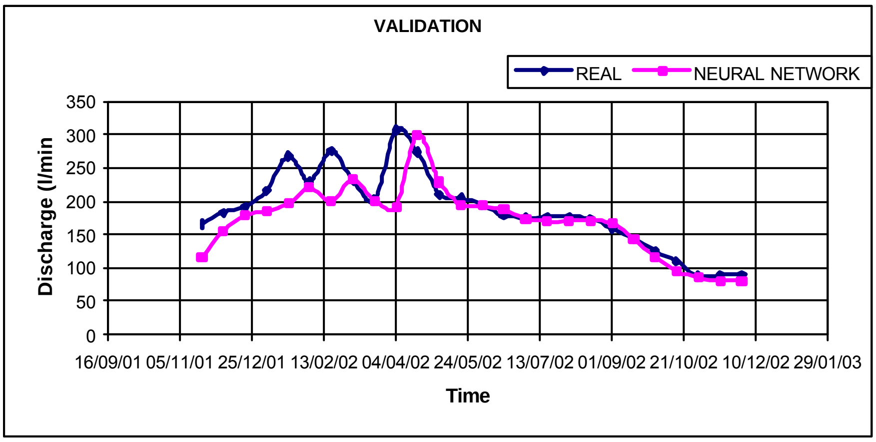

Figure 4: Comparison between real and predicted values during the validation phase.