Khabur River Flow Modeling using Artificial Neural Networks

Jowhar R. Muhamad
Lecturer
Department of Water Resources
University of Duhok

Jwan N. Hassan
Assistant Lecturer
Department of Civil Engineering
University of Duhok

Abstract:

Modeling a hydrologic time series has been one of the most complicated tasks owing to the wide range of data, the uncertainties in the parameters influencing the time series and also due to the non availability of adequate data. Recently, Artificial Neural Networks (ANNs) have become quite popular in time series forecasting in various fields. This paper demonstrates the use of ANNs to forecast Khabur monthly river flows for flow data from January 1958 to December 1975. Using the feed forward network. The network is trained using the lagged or delayed variables from SARIMA model as an input variables for the network. ANN model for mouthy flow gives better result in comparison with Traditional ANN models and SARIMA model.

[illegible]

[illegible]

[illegible]

[illegible] (ANN) [illegible] (ANN) [illegible] 1958 [illegible] 1975 [illegible] (SARIMA) [illegible] (ANN) [illegible] .(SARIMA)[illegible]

Submitted 6 th Oct. 2004 Accepted 16th Feb 2005

Introduction

Operation and planning studies in a river basin require the ability to simulate realistic streamflow variability [8].
Stochastic streamflow models were traditionally developed in both Integrated Autoregressive and Moving Average ARIMA,SARIMA and PAR models [14,4,9,13].

Kernal-based non-parametric models developed recently for streamflow simulation [11],Streamflow disaggregation [6].

Since the early nineties, ANNs have been successfully used in hydrology-related areas such as rainfall-runoff modeling, Streamflow forecasting, Ground-water modeling, Hydrologic time series.[2,3].
N.Karunanithi et al.[7] uses ANN for streamflow forecasting of the Huron river in Michigan. Amir F. Atiya et al.[1] uses the average daily flow of nile river and apply the ANN model for two time series: the ten day previous and the monthly time series.

Sana Buttamra[10] introduce the ANN approach for water consumption time series data of Kuwait, and show that if the delayed variables are identified based on Box-Jenkins analysis, then better ANN model is obtained than one based on traditional methods.

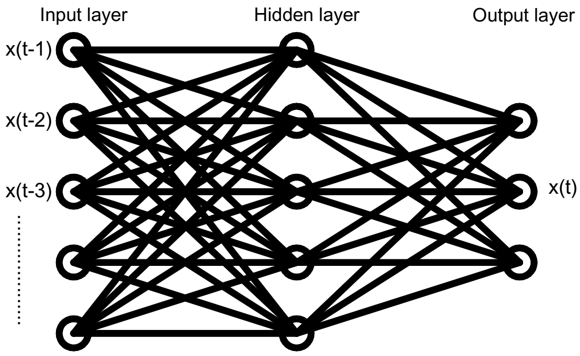
B.Sivakumar et al.[12] make a comparison between two non-linear black-box approaches, Phase-space reconstruction (PSR) and artificial neural networks(ANN), for forecasting daily river flow from the Nakhan Sawan station at Chao phraya river basin in Thailand for 1-day and 7-day ahead forecasting. The result shows good performance of both 1-day and 7-day ahead forecasting. However, the performance of the PSR approach is found to be consistently better than that of ANN.

In this paper Seasonal ARIMA (SARIMA) approach on the monthly flow of Khabur river time series from 1958-1975 is presented and an ANN model uses the delayed variables from SARIMA best model for input variables.

Artificial Neural Network (ANN)

ANN is one of the artificial intelligence algorithms that pertrains to the class of machine learning. ANN mimics the human brain process of acquiring and retrieving knowledge. It models the biological neuron that consists of nodes (cells) and links (axons). It is defined as “A computing system made up of a number of simple, highly interconnected processing elements, which process information by its dynamic state response to external input”[5].

A neural network structure consists of processing elements (nodes), links or interconnections between elements, and information processing.Neural network’s structure includes defining the number of layers, the number of nodes in each layer, and the interconnection scheme between nodes. Fig(1) shows a neural network for a three layer network with fully connected nodes of different layers. Selection of the number of layers is controlled by the learning algorithm. The training algorithm used in this paper is back-propagation algorithm which requires an interconnection between the nodes of the input and the hidden layers and nodes of the hidden and the output layers. The transfer function used is sigmoid function.

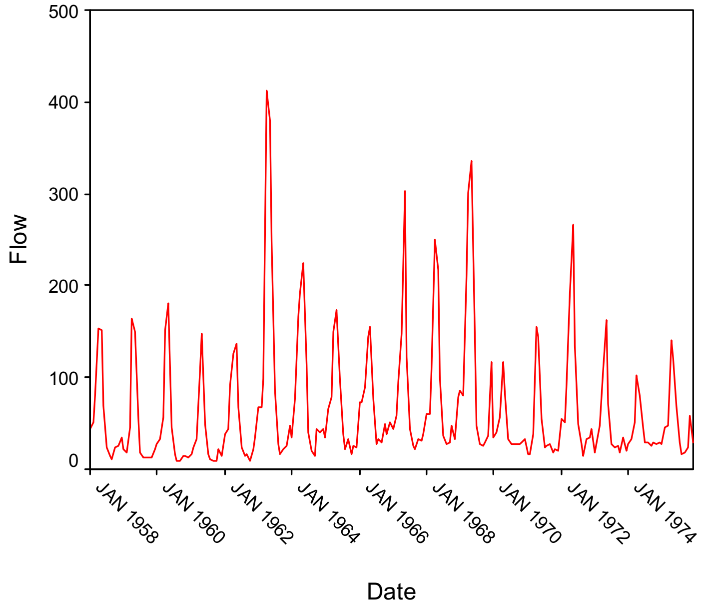


Fig(1) Neural Network Structure

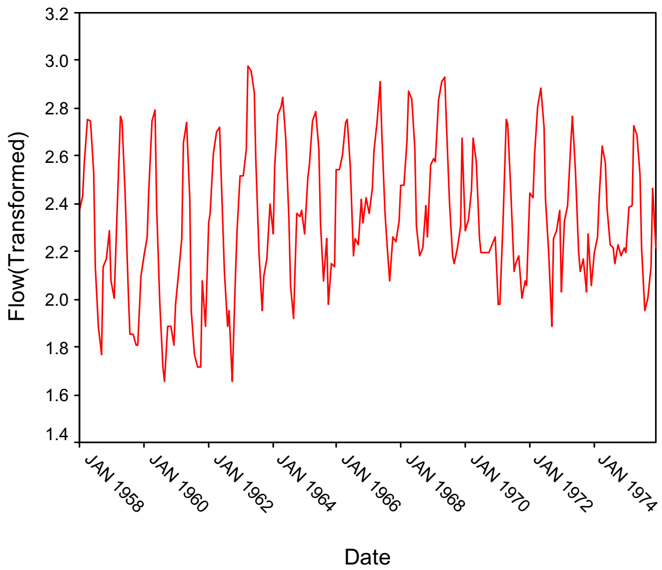
SARIMA Model

Let y_t represents the monthly flow of Khabur river during the period from January 1958 to December 1975. Fig(2) represents the time series plot of y_t which shows nonstationarity for the series. The standard SARIMA models involves taking a transformation of the data followed by seasonal and nonseasonal differences to make the data stationary. Among logarithmic, square root and box-cox transformations that we tried, a box-cox power transform $x_t = \frac{y_t^{\lambda} - 1}{\lambda}$ with $\lambda = -0.27$ gives best results, Table 1 shows the statistic properties of y_t and x_t in which it can be seen that the transform removes the skewness from the original data. Fig(3) shows time series plot of x_t which seems to be approximately constant size of variation around the mean. Fig(4) and Fig(5) shows the frequency polygon of the original and transformed data illustrates that the distribution is some what near to the normal distribution after transformation.

Since one of the main aims of this research is to construct an ANN model for the monthly Khabur river flow, therefore , the SARIMA approach is used to point which lagged variables are necessary to present as an inputs to the ANN model. Using the SARIMA technique on the transformed data, the best model found is of order $(2,1,2)(1,1,2)_{12}$ depending on minimum statistics AIC(Akaiki information criterion) and SBC (Schwarz's Baysian information criterion) which have values AIC=2008.1 and SBC=2034.6 (Table 2).



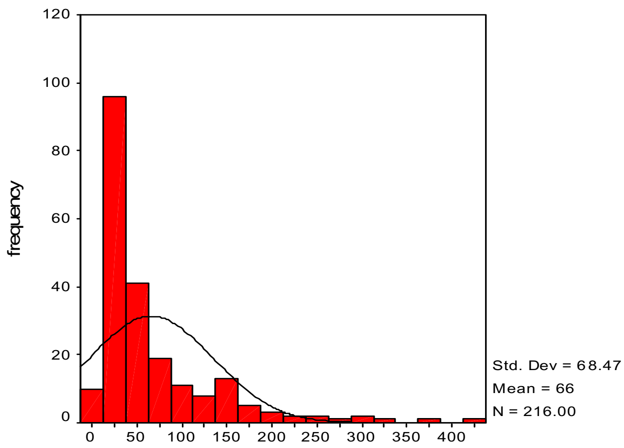
Fig(2) Time Series Plot Of Origin Data



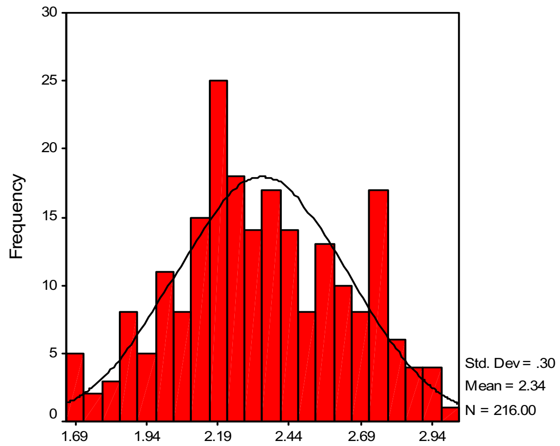
Fig(3) Time Series Plot Of Transformed Data

Table(1) Statistics of y_t and x_t

Data	y_t	x_t
Mean	65.995	2.341
Std. Deviation	68.474	.299
Skewnes	2.320	.000
Kurtosis	6.390	-.640



Fig(4) Frequency Polygon For Original Data



Fig(5) Frequency Polygon For Transformed Data

Table(2) Tested Parameters For Various Model Orders

Model Order	AIC	SBC
$(1,1,0)(1,1,0)_{12}$	2090.6	2100.5
$(1,1,1)(1,1,0)_{12}$	2066.4	2079.6
$(1,1,1)(1,1,1)_{12}$	2025.8	2042.4
$(2,1,1)(1,1,1)_{12}$	2023.3	2043.2
$(2,1,1)(2,1,1)_{12}$	2026.2	2049.4
$(2,1,2)(1,1,2)_{12}$	2008.1	2034.6
$(2,2,2)(1,1,2)_{12}$	2038.8	2065.2
$(2,1,2)(1,2,2)_{12}$	2017.3	2043.3

ANN Model

The available monthly data divided into three parts for training, testing and cross validation. The first twelve year (144 month) were used for training, Three year (36 month) for testing and last three year (36 month) for cross validation. The cross validation data computes the error in a special test set at the same time as the network in being trained with the training set to avoid the over-fitting of a network.

Two approaches were used for the input matrix preparation; the first approach is accomplished by setting the input layer nodes equal to the number of the lagged variables x(t-1),x(t-2), x(t-3).......x(t-n), where n is time delay. The time delay is taken 1,3,5,8,10 and 12 months.

The second approach is by setting the input layer nodes equal to the number of lagged variables from SARIMA model. Since the SARIMA model of order (2,1,2)(1,1,2)$_{12}$, (means autoregressive, differencing and moving average of order 2 ,1,2 respectively and seasonal autoregressive, differencing and moving average of order 1 ,1,2 respectively) then the model can be written as

$$\phi(B^{12})\varphi(B)\nabla_{12}^{1}\nabla^{1}x_t = \Theta(B^{12})\theta(B)\varepsilon_t \qquad \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots 1$$

where $\phi(B^{12})$ and $\varphi(B)$ are polynomials of degree 1 and 2 respectively

$\Theta(B^{12})$ and $\theta(B)$ are polynomials of degree 2 and 2 respectively

ε_t is independently-distributed random variable.

B^{τ} defined by $B^{\tau}\, x_t = x_{t-\tau}$, ∇_{τ} is the differencing operator defined by

$$\nabla_{\tau}\, x_t = x_t - x_{t-\tau} = \left(1 - B^{\tau}\right) x_t$$

$$\nabla\, x_t = x_t - x_{t-1} = \left(1 - B\right) x_t$$

then eq. 1 can be expressed as

$$(1+\phi_{1,1}B^{12})(1+\varphi_{2,1}B+\varphi_{2,2}B^{2})(1-B^{12})(1-B)x_t = \Theta(B^{12})\theta(B)\varepsilon_t \ldots\ldots\ldots\ldots 2$$

the above equation can be rewritten as

$$\begin{aligned} x_t &= A_1x_{t-1} + A_2x_{t-2} + A_3x_{t-3} + A_4x_{t-12} + A_5x_{t-13} + A_6x_{t-14} + A_7x_{t-15} \\ &\quad + A_8x_{t-24} + A_9x_{t-25} + A_{10}x_{t-26} + A_{11}x_{t-27} + G(\varepsilon) \qquad \ldots\ldots\ldots\ldots 3 \end{aligned}$$

where $A_1 = (1-\varphi_{2,1})$, $A_2 = (\varphi_{2,1}-\varphi_{2,2})$, $A_3 = \varphi_{2,2}$, $A_4 = (1-\phi_{1,1})$,

$A_5 = (-1+\varphi_{2,1}+\phi_{1,1}-\varphi_{2,1}\phi_{1,1})$, $A_6 = (\phi_{1,1}(\varphi_{2,1}-\varphi_{2,2})+\varphi_{2,1}+\varphi_{2,2})$,

$A_7 = \varphi_{2,2}(\phi_{1,1}-1)$, $A_8 = \phi_{1,1}$, $A_9 = \phi_{1,1}(\varphi_{2,1}-1)$, $A_{10} = \phi_{1,1}(\varphi_{2,1}+\varphi_{2,2})$,

$A_{11} = -\phi_{1,1}\varphi_{2,2}$ and $G(\varepsilon) = \Theta(B^{12})\theta(B)\varepsilon_t$

So

$$x_t = f(x_{t-1}, x_{t-2}, x_{t-3}, x_{t-12}, x_{t-13}, x_{t-14}, x_{t-15}, x_{t-24}, x_{t-25}, x_{t-26}, x_{t-27}) \ldots\ldots\ldots\ldots 4$$

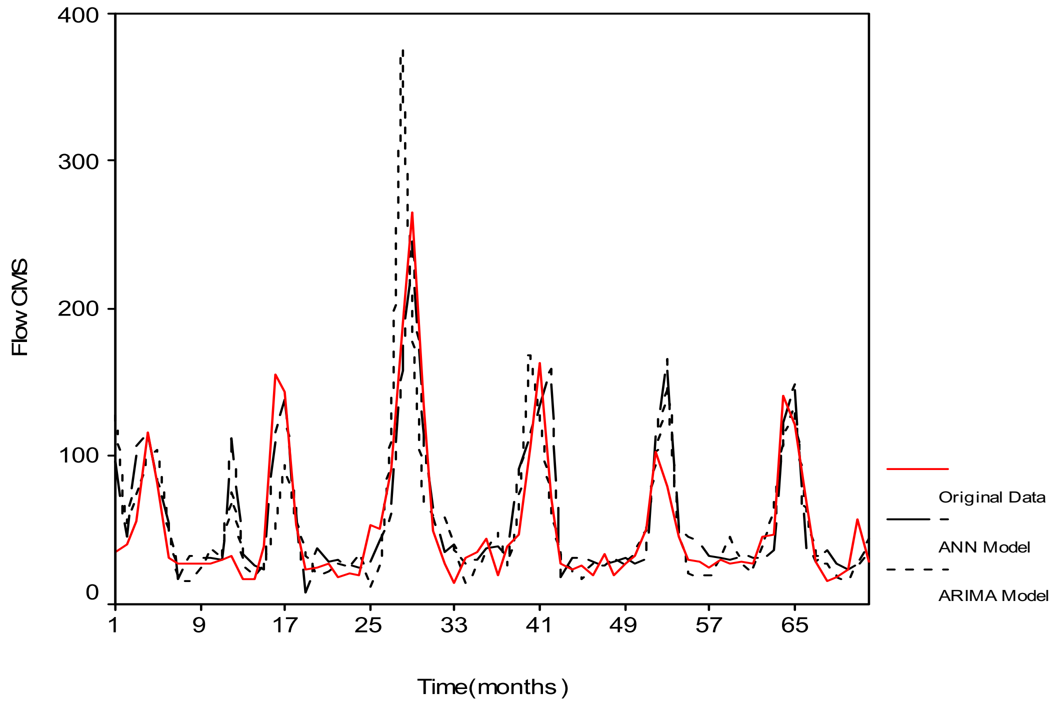
Eq. 4 represent the input lagged variables in second approach.
The total number of ANN models will be seven carried out on a PC using the commercially available software package MATLAB 6.5 . A variety of different architectures was examined, and the best result was obtained based on RMS error with a fully connected multilayer neural network containing 24 neurons in each of two hidden layers. The network was trained for about 17,000 epochs using a momentum of 0.5.

Results

The evaluation of models is performed using correlation coefficient (r) and root mean square error (RMSE). Table (3) resumes the results obtained from models (1) to (7) computed over the testing phase of ANN. It is clear that the best performance is achieved by model (7) which representing the second approach. Fig(6) shows a comparison between ANN model and SARIMA model performances as compared to the actual data for the last six years. It is clear that the fitted model is performing better than SARIMA model especially at the tails of the series, where as seen, that the second tends to divert more, Specifically at the tails of the series.

Table(3) Statistics of models

Model	Output variable	Correlation coefficient r	RMSE
1	x_{t-1}	0.813	5.44
2	x_{t-1}, x_{t-2}, x_{t-3}	0.874	4.61
3	x_{t-1}, x_{t-2}, x_{t-3},x_{t-4}, x_{t-5}	0.896	3.78
4	x_{t-1}, x_{t-2}, x_{t-3},x_{t-4}, x_{t-5}, x_{t-6}, x_{t-7}, x_{t-8}	0.912	2.15
5	x_{t-1}, x_{t-2}, x_{t-3},x_{t-4}, x_{t-5}, x_{t-6}, x_{t-7},x_{t-8}, x_{t-9}, x_{t-10}	0.932	1.24
6	x_{t-1}, x_{t-2}, x_{t-3},x_{t-4}, x_{t-5}, x_{t-6}, x_{t-7},x_{t-8}, x_{t-9}, x_{t-10}, x_{t-11}, x_{t-12}	0.954	0.85
7	x_{t-1},x_{t-2}, x_{t-3},x_{t-12}, x_{t-13}, x_{t-14}, x_{t-15}, x_{t-24}, x_{t-25}, x_{t-26}, x_{t-27}	0.978	0.22



Fig(6) Time Series Plot Of Proposed Model Output

CONCLUSIONS

In this paper, ANN method was applied to model the monthly flow of Khabur river. The results reveal that ANN model produced fairly significant prediction when it was used in conjunction with SARIMA approach. Using the lagged variables from SARIMA model as an input layer to ANN model, the correlation coefficient (r) and root mean square error (RMSE) for different models indicate that the proposed modification gives best results with r=0.938 and RMSE=0.22.

References

1. Amir F. A., Samir I. S., “A Comparison Between Neural-Network Forecasting Techniques-Case Study: River Flow Forecasting”, IEEE Transactions On Neural Networks, Vol 10, No. 2, 1999.

2. ASCE Task Committee,“Artificial Neural Networks in Hydrology I: Preliminary Concepts”, Journal of Hydrologic Engng., ASCE, Vol. 5, No. 2, P 115-123, 2000.
3. ASCE Task Committee,“Artificial Neural Networks in Hydrology II: Hydrologic Applications”, Journal of Hydrologic Engng., ASCE, Vol. 5, No. 2, P 124-137, 2000.
4. Bras R., Rodriquez-Itube I.,”Random Functions and Hydrology Reading”, MA: Addison-Wesley, 1985.
5. Caudill, M., “Neural Networks Primer”, Al Expert, 1987.
6. David G. Tarboton, Ashish Sharma, Upmanu Lall, “ Disaggrigation Procedures For Stochastic Hydrology Based On Nonparametric Density Estimation”, Water Resources Research, Vol 34, No. 1, P 107-119, 1998.
7. Karunanithi N., Grenney W.J., Whitley D., Bovee K., “ Neural Networks For River Flow Prediction”, Journal of Comp. Civ. Engng., ASCE, Vol 8, No. 2, P 201-220, 1994.
8. McMahon T.A., P.B.Pretto,F.H.S. Chew, “ Synthetic Hydrology. Where’s the hydrology. In: Stochastic Hydraulics ”, Edited K.S. Tickle, I.C. Gowter, Netherlands, 1996.
9. Salas J.D., Deller J.W., Yevjevich V., Lane W.L., “ Applied Modeling Of Hydrologic Time Series”, Water Resources Publications, Littleton, Colorado, 1980.
10. Sana BuHamra, Nejib Smaoui, Mahmoud Gabr, “ The Box-Jenkins analysis and neural networks: prediction and time series modeling ” , Applied Mathematical Modeling Vol 27, P 805-815, 2003.
11. Sharma A., David G. Tarboton, Upmanu Lall, “Streamflow Simulation: A Nonparametric Approach ”, Water Resources Research, Vol 33, No. 2, P291-308, 1997.
12. Sivakumer B.,Jayawardena A.W., T.M.K.G Fernando, “River Flow Forecasting: Use Of Phase-Space Reconstruction And Artificial Neural Networks Approaches”, Journal of Hydrology, Vol 26, No. 5, P 225-245, 2002.
13. Stedinger J.R., M.R. Taylor, “Synthetic Streamflow Generation, 1:Model Verification and Validation”, Water Resources Research, Vol 18, No. 4, P 909-918, 1982.
14. Yevjevich V., “Stochastic Process In Hydrology”, Water Resources Publications, Fortcollins, Colorado, 1972.