COMPARISON BETWEEN MULTIPLE LINEAR REGRESSION AND FEED FORWARD BACK PROPAGATION NEURAL NETWORK MODELS FOR PREDICTING EFFLUENT WATER QUALITY IN A TREATMENT PLANT SYSTEM.

[1*]HOWARD, C.C., [2] ETUK, E. H. AND [3]HOWARD, I. C.

[1]Department of Mathematics, Faculty of Science, University of Africa Toru-orua, Sagbama. Bayelsa State, Nigeria. [2]Department of Mathematics, Faculty of Science, Rivers State University Port Harcourt, Nigeria. [3] Department of Chemistry/Biochemistry Federal Polytechnic, Nekede, Owerri. Imo State, Nigeria

ABSTRACT

In this study effluent water quality in a treatment plant system located at the Gulf of Guinea, Nigeria is presented. The time series data used were generated by a standard laboratory that actually carried out the field and laboratory analysis which involves weekly water quality data obtained directly from a flow station for the period of five years. Two major effluent water quality parameters; biochemical oxygen demand (BOD_5) and chemical oxygen demand (COD) are considered in this study. The result from multiple linear regression (MLR) analysis using stepwise forward selection method for BOD_5 and COD showed that two parameters for BOD_5 (COD and dissolved oxygen (DO) and one parameter for COD (BOD_5) have significant impact. The main objective of this study is to compare the accuracy of Artificial Neural Network (ANN) and MLR models for prediction of effluent water quality. We have compared MLR with ANN of three layer feed- forward network with sigmoid function in the second and third layer using a resilient back propagation algorithm. The performance of MLR was found to be better than the ANN model for both BOD_5 and COD predictions with least values of the statistical error measures viz. root

mean square error (RMSE), mean absolute error (MAE) and mean absolute percentage error (MAPE) used for the comparisons of the models prediction abilities.

***Keywords**: Prediction, ANN, MLR, effluent water quality*

Introduction

Many authors have carried out comparative studies between traditional techniques and ANNs. It has been recognized in the literature that regression and neural network methods have become competitive modeling methods. Although artificial neural networks have the advantages of accurate prediction, their performance in some specific situations is not consistent (Khashei and Bijari, 2011a).

Despite the fact that many studies that have shown that ANNs are considerably better than the conventional linear models and their prediction is considerably and consistently more accurate, however other investigations have come up with some inconsistent results. For instance Forster *et al.*, (1992) observed that ANNs approaches are considerably less efficient than linear regression and a simple average of exponential smoothing methods. Brace *et al.* (1991) also, observed that the capabilities of ANNs are lower than that of other methods of statistical analysis that are often used in predicting load. Denton (1995) in his work made available data through different experimental conditions for which in an ideal situation, having in mind the assumptions of regression analysis, indicates a non significant differences in the predictability of ANNs and Linear regression, however in less ideal circumstances like outliers and multicollinearity as a wrong model specification ANNs performs better than MLR. A comparative study of ANNs and MLR in exchange rate forecast was carried out by Ham and Steurer (1996). They reported that ANNs do not give accurate forecast which is unlike MLR when monthly data is used. This was corroborated by the study of Taskaya and Cassey (2005) where they compared

the capabilities of ANNs and Multiple linear regression models. In this case and in some other cases MLR surpasses neural networks (Khashei & Bijari, 2011b). Other investigators have made several comparisons between artificial neural network and the corresponding traditional linear regression techniques in their specific areas of application. Again Fishwick (1989) asserted that the performance capabilities of ANN are far less than the simple linear regression approaches; while responding to this assertion, Tang *et al.*, (1991) and Tang and Fiswick (1993) asked the question as to under what circumstances can ANN predictors predict better of than the MLR prediction methods for time series models. Some researchers are of the school of thought that certain circumstances where ANNs predicts worse than MLR models, could be due to the fact that the variables are linear with very minimal disruption; ANNs cannot be expected to outperform linear models for linear relationships (Zhang *et al*. 1998). For whatever reason, the use of ANNs to model linear problems has yielded mixed results and therefore; one must chose wisely and not to blindly apply ANNs to every type of data. Therefore there is a need to compare models to know which model that best fits the data. Hence this study compares the accuracy performance of MLR and ANN models for the prediction of BOD_5 and COD.

MATERIALS AND METHODS

The data (five years, total observation of 260, 2007-2011) for this study was generated by a standard laboratory that actually carried out the field and laboratory work which involves collection of weekly effluent samples for the analysis of principal parameters BOD_5, COD, Dissolved oxygen, conductivity, temperature, etc. from a waste water treatment plant located between Longitude $4^o34.276'$ and Latitude 8^o 25.557′ at the Gulf of Guinea. The models were built using the Times Series Forecasting System tool of the R software package.

Multiple Linear Regressions

Multiple regression analysis is meant to include multiple independent parameters at the same time for predicting the importance of a dependent

parameter. In this study, several linear regression equations associated with weekly produced water quality parameters behave as a dependent parameter and are generalized as five other independent parameters, given below**.** Independent parameters and stepwise regression analysis were used to identify critical parameters to predict dependent parameters based on five of them.

$$Y = \beta_o + \beta_1 X_1 + \beta_2 X_2 + \ldots\ldots + \beta_6 X_6 + \varepsilon \qquad (3.1)$$

Whereβ_o = Intercept,

β_1 = regression coefficient of i^{th} independent parameters (i=1,2,….,6),

ε = error term,

X_1= i^{th} effluent water quality parameters selected using stepwise forward regression method.

Artificial Neural Network

ANN captures the domain knowledge; it can handle continuous as well as discrete data and have good generalization capability as with fuzzy expert systems. An ANN is a computational model of the brain and it is composed of an input layer, one or more hidden layer, and an output layer. They assume that the computation is distributed over several simple units called neurons, which are interconnected and operate in parallel thus known as parallel distributed processing systems.

Implicit knowledge is built into a neural network by training it. The number of hidden layer and the neuron number in the hidden layer are determined by trials. In the hidden layer, inputs and relevant weights are multiplied, and then the results are transmitted to transfer function (Yongjae and Sehun, 2005).

Results and Discussion

The water quality data (five years, a total of 260 observations) were divided into two data sets. The first data set containing former 4-year (2007-2010) records were used as the training data for model development; the second data set

containing the remaining year (2011) records were used as the testing data to evaluate capabilities of the established model.

Multiple Linear Regression

Biochemical Oxygen Demand (BOD_5)

The multiple linear regression model is fitted to predict the weekly BOD_5 as a dependent parameter taking the other weekly independent parameters as Chemical Oxygen Demand (COD), Dissolved Oxygen (DO). The most significantly contributed parameters are selected using forward stepwise selection having the smallest BIC value (see Table 1). Therefore, more emphasis should be laid on these parameters during selection for further improvement of BOD_5. The best fit multiple regression model is given as equation (3.1):

$$BOD = 98.0893 + 0.4630COD – 16.8979DO + \mu_i \quad (3.1)$$

The parameters temperature, pH and conductivity seem to have the least control over BOD_5 and hence did not appear in the proposed multiple regression models.

Table 1: The best subset with BIC values for weekly BOD_5 parameter prediction

Number iof ivariables	Best iSubset	BIC
1	COD	-89.476
2	COD, iDO	-145.541
3	COD, iDO, iTEMP	-145.042
4	COD, iDO, iTEMP, ipH	-139.775
5	COD, iDO, iTEMP, ipH, iCOND	-134.457

Artificial Neural Network

The ANN model building process was performed using the five water quality parameters. The best-suited architecture of Feed Forward Neural Network

Model for our weekly BOD_5 data was selected by comparing methods and changing the layer and number of neurons in each network. This proposed model had an input environment with significant water quality parameters, one hidden layer with 6 neurons and one neuron in the output layer (see Figure 1). The number of neurons in the hidden layer was varied, as shown in Table 2. The number of hidden neurons (four) with the smallest value of RMSE is the best fit. A resilient back propagation with weight backtracking algorithm was used for training this multilayer perceptron until the best combination was achieved. A sigmoid activation function was used in the hidden layer and output layer. A set of random values distributed uniformly from 0 to 1 utilised to start the weight of the neural network model. The best-fitting network selected, is composed of five input neurons, six neurons in the hidden layer and one output neuron (in abbreviated form), N (5×6×1), see Figure 1.

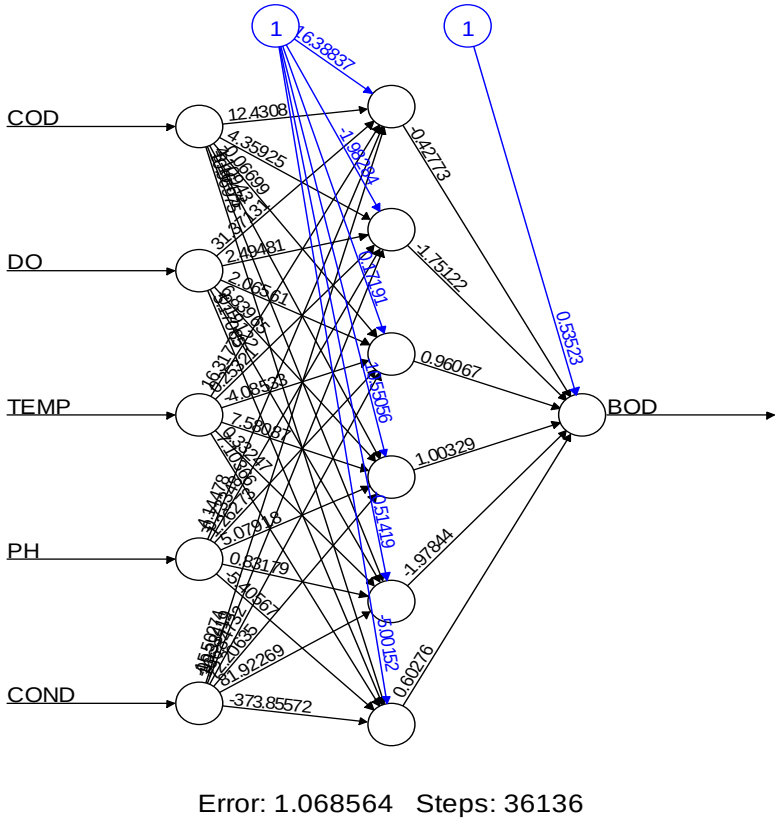


Figure 1: Artificial neural network structure for weekly BOD_5 prediction parameter

Table 2: Neural network performance using different number of hidden neurons Note: RMSE = Root Mean Square Error

Number iof I hidden ineurons	Training i i(RMSE)
4	28.09486
5	30.46437
6	26.81779
7	33.83007
8	31.51882
9	30.18726
10	30.9331

Chemical Oxygen Demand (COD)

Multiple Linear Regression (MLR)

The multiple regression models is fitted to predict the weekly chemical oxygen demand (COD) as a dependent parameter taking the other weekly independent parameter as BOD_5. The most significantly contributed parameter (BOD_5) is selected using stepwise forward regression analysis as the best subset having the smallest BIC value (see Table 3). The best fit multiple regression model is given as equation (3.2):

$$COD = 27.1735 + 0.65481BOD_5 + \mu_i \quad (3.2)$$

Table 3: The best subset with BIC values for weekly COD parameter prediction

Number iof iVariables	Best iSubset	BIC
1	BOD	-89.476
2	BOD, ipH	-85.1797
3	BOD, ipH, iCOND	-80.9352
4	BOD, iDO, ipH, iCOND	-76.5339
5	BOD, iDO, iTEMP, ipH, iCOND	-71.2331

Artificial Neural Network

An ANN model building process was performed using all the same water quality parameters. The best-suited architecture of Feed Forward Neural Network Model was selected for our weekly COD data by comparing methods and changing the layer and number of neurons in each network. The number of

neurons in the hidden layer was varied, as shown in Table 4. The number of hidden neurons (seven) with the smallest value of RMSE is the best fit. Resilient backpropagation with weight backtracking algorithm was used for training this multilayer perceptron until the best combination was achieved. A sigmoid activation function is used in the hidden layer and output layer. A set of random values distributed uniformly varies from 0 to 1 and is used to start the weight of the neural network model. The best-suited architecture of Feed Forward Neural Network selected, is composed of five input neurons, seven neurons in the hidden layer and one output neuron (in abbreviated form, N (5×7×1), see Figure 2.

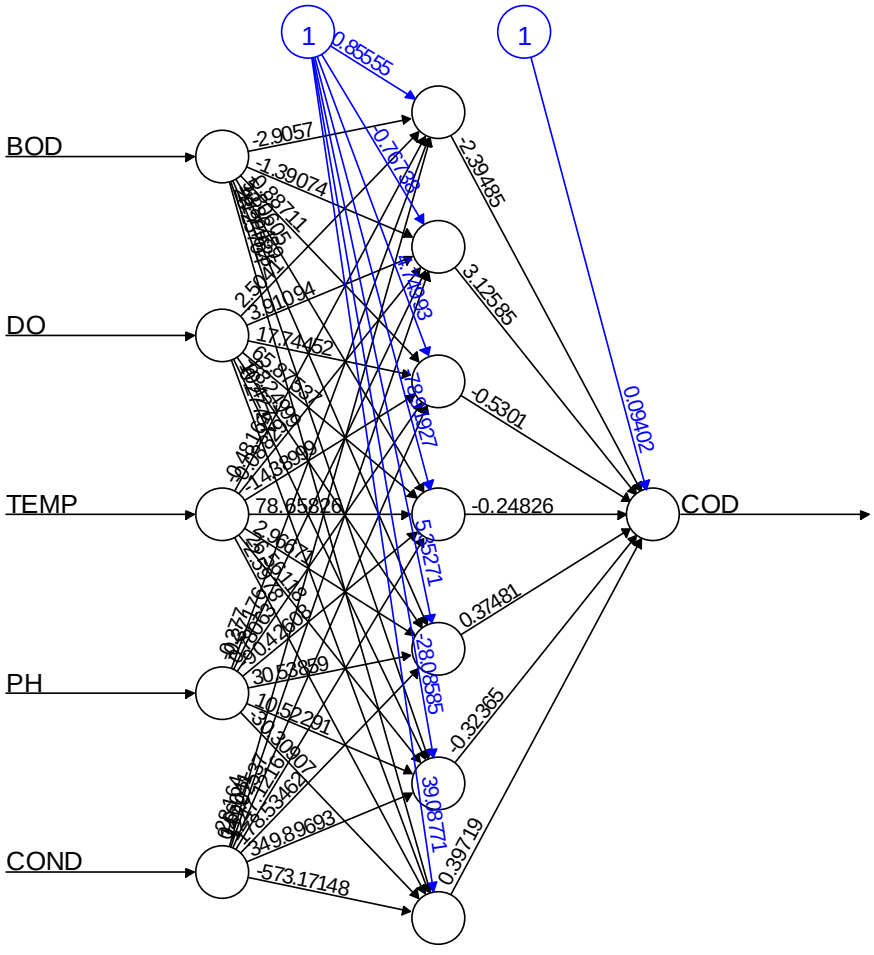


Figure 2: Artificial neural network structure for weekly COD prediction parameter

Table 4: Neural network performance using different number of hidden neurons Note: RMSE = Root Mean Square Error

Number iof I hidden ineurons	RMSE
4	24.29487
5	25.56701
6	23.62751
7	21.94817
8	24.95865
9	32.14235
10	32.14235

Comparison of MLR and ANN

A comparison of the efficiency of actual weekly BOD_5 and COD with their predicted value using MLR and ANN models is presented, graphically in Figure 3 and 4. The linear scale plots of the observed BOD_5 and COD with the predicted $BOD_5$and COD using MLR and ANN models shows that predicted values by MLR models are tending more towards the actual values of weekly BOD_5 and COD. The estimates viz. mean absolute error (MAE), root mean square error (RMSE) and mean absolute percentage error (MAPE) are presented in Table 5 and 6 also show that MLR model has the least values for the testing data set. The graphical representations as well as the numerical estimates both favored the MLR model as a preferred model in comparison to the ANN model, concluding that this MLR technique can be used as an effective BOD_5 and COD forecasting tool in the flow station.

Table 3: Comparison of the performance of forecasting models for BOD_5 parameter

Techniques	RMSE	MAE	MAPE
MLR	27.220180	23.311300	25.680630
ANN	36.792200	26.885900	31.708000

Table 4: Comparison of the performance of forecasting models for COD

Techniques	RMSE	MAE	MAPE
MLR	18.25344	17.14506	19.03417
ANN	41.72618	28.90650	36.07570

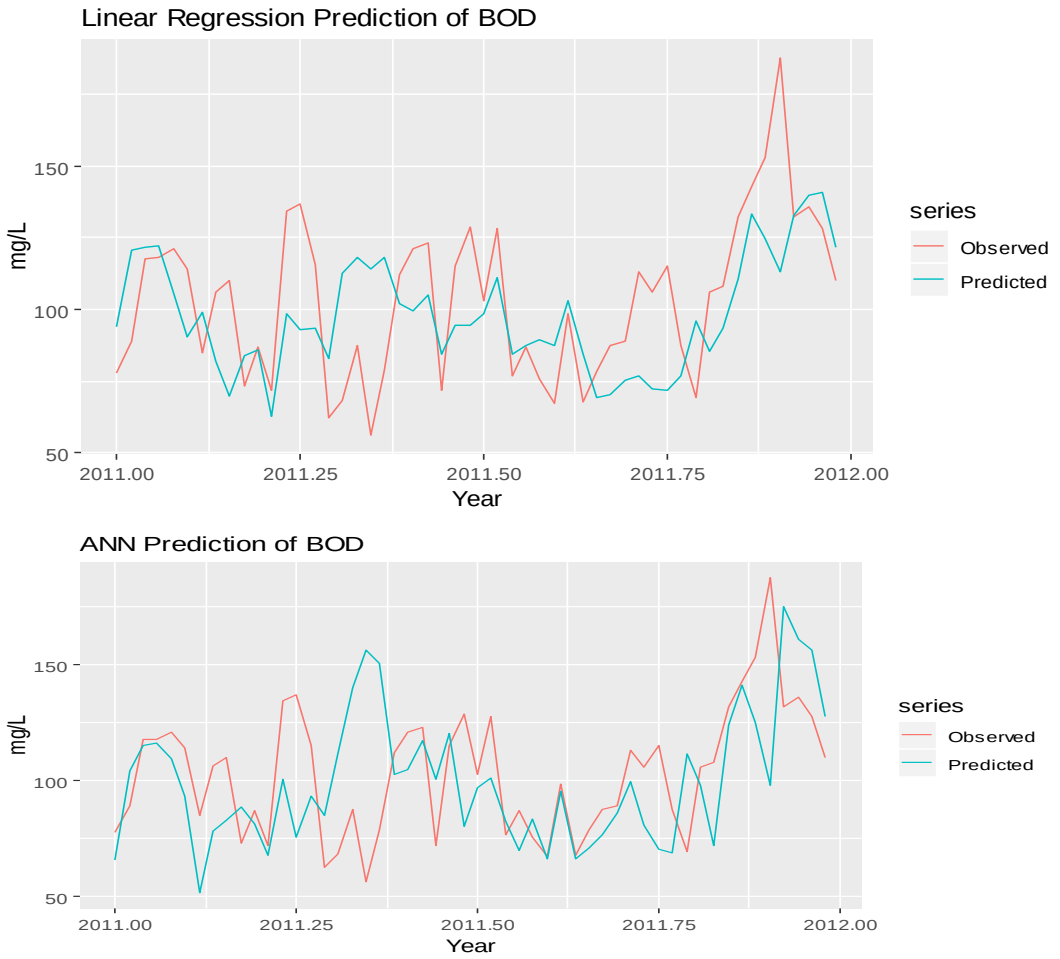
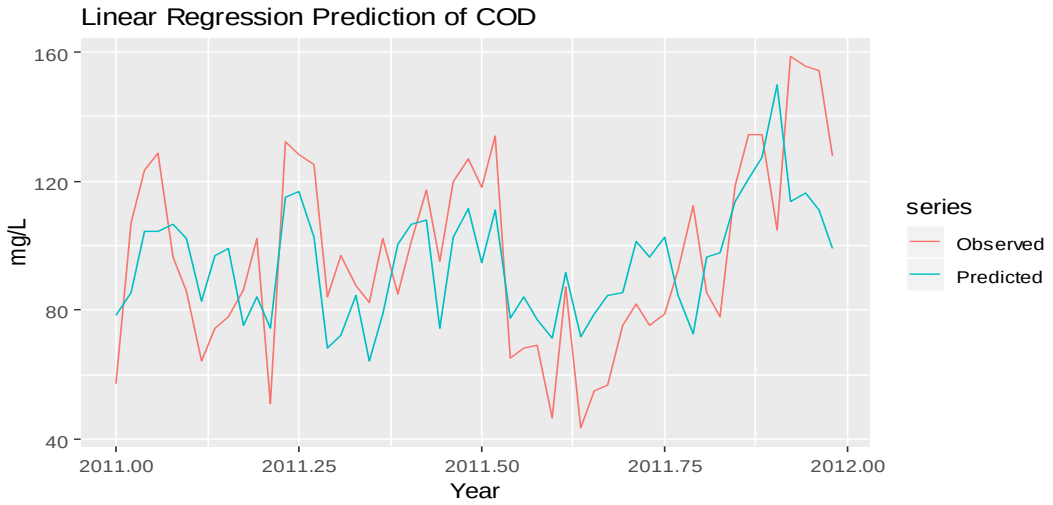


Figure 3: Linear scale plot of the observed BOD_5 with the predicted BOD5 of MLR and ANN models



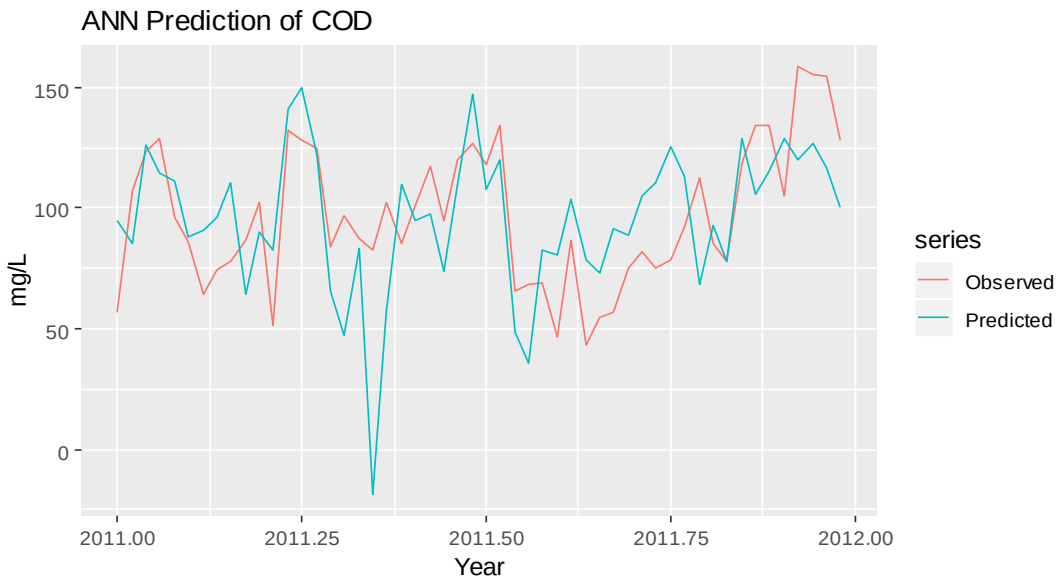


Figure 4. Linear scale plot of the observed COD with the predicted COD of MLR and ANN models.

Conclusion

In the present study, the efficiency of MLR and ANN models were investigated in prediction of two major water quality parameters, BOD and COD, in the waste water treatment plant. Performance of the models was evaluated using three statistical error measures mentioned above. The results indicated that the MLR model with significant input parameters selected using forward stepwise regression (COD and DO) and BOD_5 could be successfully used for predicting BOD_5 and COD concentrations respectively. Comparison of the ANN and MLR models showed that the MLR model performed much better than the ANN since MLR shows minimum values in all the error estimates. The MLR model developed in this study can be more useful in effluent water quality management efforts to ensure that water resource is sustainable for the future.

References

Brace M. C., Schmidt J. & Hadlin M., (1991). Comparison of the Forecasting Accuracy of Neural Networks With Other Established Techniques. In: Proceedings of the First Forum On *Application of Neural Network to Power Systems*, Seattle, WA., 31-35.

Denton J.W. (1995), How good are neural networks for causal forecasting? *The Journal of Business Forecasting* 14(2); 17–20.

Fishwick P.A. (1989). Neural network models in simulation: A comparison with traditional modeling approaches; *Proceedings of Winter Simulation Conference*, Washington D.C.; 702–710

Foster, W.R., Collopy F.& Ungar L.H. (1992). Neural network forecasting of short, noisy time series; *Computers and Chemical Engineering* 16(4); 293– 297.

Hann T.H.& Steurer E. (1996), Much ado about nothing? Exchange rate forecasting: neural networks vs. linear models using monthly and weekly data; *Neurocomputing* 10, 323–339.1

Khashei M.& Bijari M. (2011b), A novel hybridization of artificial neural networks and ARIMA models for time series forecasting; *Applied Soft Computing* 11, 2664–2675.

Khashei, M.& Bijari, M. (2011a) Which Methodology is Better for Combining Linear and Nonlinear Models for Time Series Forecasting? *Journal of Industrial and Systemsi Engineering* 4(4); 265-285

Tang Z.& Fishwick P.A. (1993), Feed forward neural nets as models for time series forecasting; *ORSA Journal on Computing* 5(4); 374–385.

Tang Z., Almeida C.& Fishwick P.A. (1991), Time series forecasting using neural networks vs. Box-Jenkins methodology; *Simulation* 57(5); 303–310.

Taskaya T.& Casey M. C. (2005), A comparative study of autoregressive neural network hybrids; *Neural Networks* 18; 781–789.

Yongjae, K., and Sehun, R. 2005. Arc sensor model using multiple-regression analysis and a neural network. *Pro. Q. Sci. J*. 219:431–47.

Zhang, G., Patuwo, B.E.& Hu M.Y. (1998), Forecasting with artificial neural networks: The state of the art; *International Journal of Forecasting* 14. i35– i62.