Journal of Applied Mathematics

volume III (2010), number III

THE IMPLEMENTATION OF HYBRID ARIMA-NEURAL NETWORK PREDICTION MODEL FOR AGREGATE WATER CONSUMTION PREDICTION

ŠTERBA Ján, (SK) , HIĽOVSKÁ Katarína (SK)

Abstract. Typical time series prediction methods used in many real-world applications – ARIMA models and Neural networks, achieve good prediction performance, hovewer, both of them fits better the different type of time series. The ARIME models are generally better in prediction of linear time series, while Neural Networks are superior in predicting nonlinear time series. To create one, superior prediction method suited for prediction of general real-world time series, containing both linear and nonlinear parts, individual models can be hybridized to create single, ARIMA-Neural Network hybrid model. Using this approach, models can complement each other in capturing patterns and internal dependencies of time series. To examine the performance, proposed hybrid prediction method is used for prediction of water consumtion based on time series collected from 1984 to 2007. As can be observed from results achieved from 12-step ahead prediction, the hybrid neural network outperforms the individual forecasting model.

Keywords. ARIMA-Neural Network, hybrid system, time series prediction

Mathematics Subject Classification: Primary 60G25; Secondary 62M20.

1 Introduction

The prediction of some real-world time series is a very important component for urban and industrial planning on both national and municipal level. Hence, it is critical to realize high-precision forecasting models to obtain reliable data. The state of the art of the time series forecasting methods in recent years can be divided into two areas. Ones are forecasting models based on traditional mathematical models, such as ARIMA model, Parametric Regressive model, Kalman filter model, Exponential Smoothing model, etc. Others are forecasting methods and models which does not pay attention to rigorous mathematical derivations and clear physical meaning, but emphasize on whether the model can fit the underlying relations of the investigated problems closely, including artificial neural network models, nonparametric regressive models, KARIMA algorithms, spectral basis analyses and others.

The accuracy of water consumption forecast has significant impact on planning and decision making of water facilities. Accurate water prediction is therefore important, especially with present rapid changes in water supplies and sources availability. Absolute necessity of exact water levels and water significance for industrial and municipal areas, and high costs for obtaining additional supplies makes accurate forecasts necessary. Therefore, it is important to search for models improving the performance of prediction and reducing the forecast error.

Box and Jenkins developed the autoregressive moving average to predict time series [1]. The ARIMA model is used for prediction non-stationary time series when linearity between variables is supposed. However, in many practical situations supposing linearity is not valid. For this reason, ARIMA models do not produce effective results when used for explaining and capturing nonlinear relations of many real world problems, what results in increased forecast error.

Artificial neural networks (ANN) models are part of an important class that has attracted a considerably attention in many applications. The use of ANN in many applied works is generally motivated by empirical results showing that under certain conditions, even simple ANN are able to approximate any measurable function to any degree [2-4]. As artificial neural networks are used as universal function approximations [5], they are very often used as to predict nonlinear time series [6,7]. Existing ANN models for forecasting generally use Multilayer Perceptron networks, which parameters - number of hidden layers, number of neurons in the layers and transfer function are often chosen trough trial and error method with aim of finding the most feasible model for specific application [8].

However, the real-world time series problems are not absolutely linear or nonlinear – they often contain both linear and nonlinear parts. Furthermore, real time problems are often affected by irregularities and infrequent events, which make time series forecasting complicated and difficult [9]. Thus, using a single model for forecasting is not the best approach. Although both ARIMA and ANN models have achieved success in their own linear and nonlinear domains, neither ANN or ARIMA can adequately model and predict time series since the linear model cannot deal with nonlinear relationships, while ANN models alone are not able to handle both linear and nonlinear patterns equally well.

As a result, several researchers have proposed hybridizing ARIMA and ANN models, since different forecasting models can complement each other in capturing patterns of data set and time series. Both theoretical and empirical studies have revealed that a hybridization forecast outperforms individual forecasting models [7,10]. The merging of this structure can help the researchers in modelling complex structures in real-world time series more effectively. Moreover, by using ARIMA and ANN in single model can significantly assist in generating lower generalization variance of error.

To test the hybrid ARIMA-ANN prediction error on water consumption time series, data set consisting of 278 observation points was used, using aggregate water consumption ranging from January 1984 to June 2007.

The remaining of this paper is organized as follows. In section 2, description of ARIMA models is presented. Section 3 describes the hybrid ARIMA-ANN model and neural network used. Detailed discussion on the results is given in section 4, and finally in section 5 is a conclusion and projection of a future work.

2 ARIMA Models for Modelling of Time Series

ARIMA models, also known as Box-Jenkins models, are classical time series analyses method, which is being generally used for time series prediction. An Autoregressive Moving Average ARMA(p,q) is defined as

$$y_t(t) = \sum_{i=1}^{p} \phi_i y_{t-i} + \sum_{j=o}^{q} \theta_j \varepsilon_{t-j}, \tag{1}$$

where y_t is the time series value lagged by time moments $t = 1, 2, \ldots, l$, the ϕ_i and the θ_j are the auto-regressive and moving averages model parameters, and the ε_t is purely a random process with zero mean and variance σ^2.

One of the necessary conditions for applying ARMA model is the stationarity of the time series, which in practice, is very rarely met. For this reason, extension of ARMA model exist, which allows to apply model even on non-stationary time series, called autoregressive integrated moving average process (ARIMA). This extension transforms the time series by differencing them by the order of d, thus insuring stationarity of the time series. In other words, if the series y_t is non-stationary, but the d-th difference, $\Delta^d y_t = (1-B)^d y_t$, is stationary.

The ARIMA(p,d,q) model is then defined as

$$\phi(B)[(1-B)^d y_t - \mu] = \theta(B)\varepsilon_i \tag{2}$$

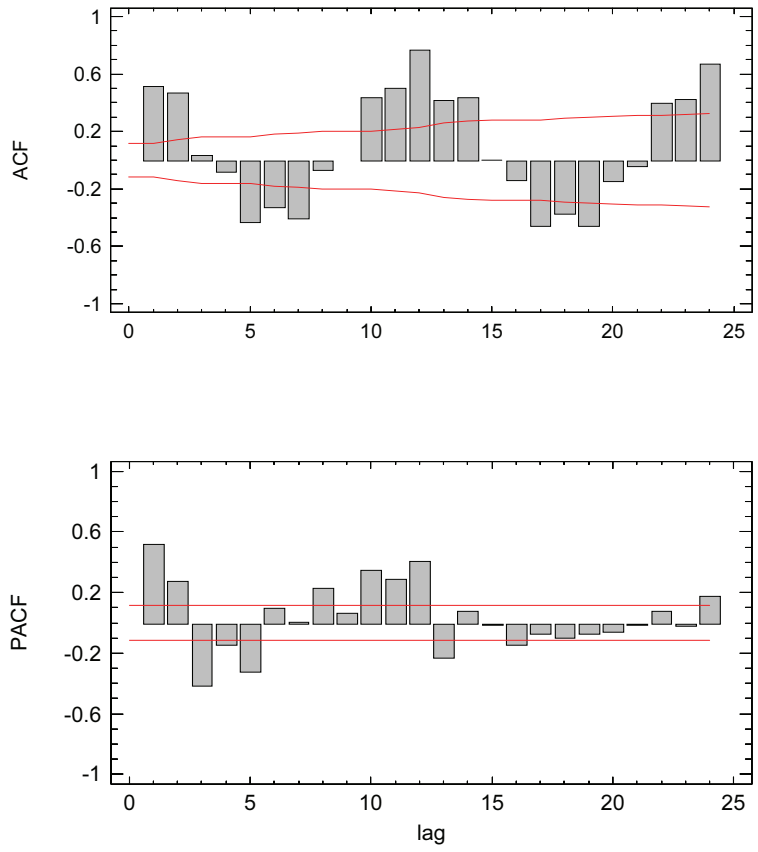

Figure 1. The autocorrelation function (ACF) and partial autocorrelation function (PACF) of analyzed time series

Thus, ARIMA model of order (p,d,q), often simply denoted as ARIMA(p,d,q) can be uses to predict the values of non-stationary time series.

Because real-world time series often fluctuates with seasonal patterns, the above mentioned ARIMA cannot model time series successfully, especially when seasonality presents a dynamic pattern. Therefore, ARIMA(p,d,q) model can be extended, forming the seasonal ARIMA (p,d,q) (P,D,Q) model of time series. For more information regarding seasonal ARIMA models of time series and their practical applications, see [1].

Process of evaluation of model parameters p,q,d, and seasonal parameters P,Q and D if they are present, is often referred to as Model Identification. The identification of ARIMA model is usually based on analyses of auto-correlation function (ACF) and Partial auto-correlation function (PACF), or alternatively it can be based on AIC criterion (Akaike Information Criterion) or FPE (Final Prediction Error) criterion. The PACF and ACF of the time series evaluated in this article can be seen in Fig. 1.

For more detail information regarding model identification, see [11].

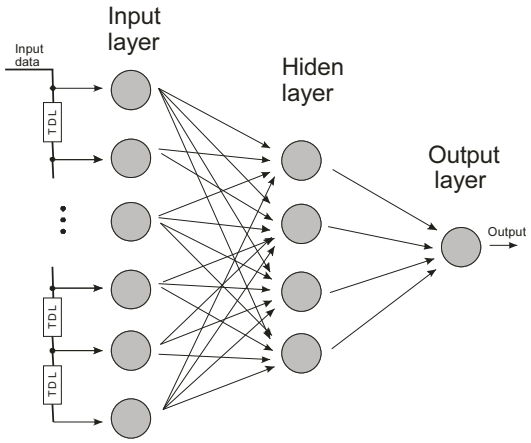

Figure 2. Artificial neural network with one hidden layer and tapped delay lines (TDL) at the input layer

TABLE I.
PERFORMANCE COMPARISON OF PREDICTION METHODS

Prediction method	RMSE	MAD	MAPE
ARIMA	8.1658 e^{+4}	5.4648	3.1638 %
ARIMA-NN	6.3163 e^{+4}	4.2099	2.4479 %

3 ARIMA-Neural Network Hybrid System for Time Series Prediction

Hybrid system is the combination of two or more than two systems in a one functioning system. Our hybrid system was obtained by combining neural networks with ARIMA time series model. Figure 2 demonstrates the framework of employed hybrid system.

The first step of investigated hybrid system involves usage of seasonal ARIMA model to model the linear part of the time series, and to create the ARIMA forecast. As the ARIMA model is based on linear relationship of system parameters, ARIMA can forecast linear relationships with high performance, but the performance often fail when the time series have non-linear relationship. On the other hand, the artificial neural networks have proven good performance when modelling non-linear relationships. For this reason, neural network is engaged in the following step to model the non-linearity, and all the remaining relationships which have not been absorbed by ARIMA model. Therefore in the second step, the ARIMA forecasts and times series data are used as inputs for artificial neural network, and trained using the known input and output training data to model the system responsible for creation of the time series. In the third and last stage, the neural network

is used to predict the future values of investigated time series 12-step ahead. As the network is trying to predict values in an interval out of known input values, the output of neural network from one time spot is used as an input of neural network in the following time spot. Thus, output from artificial neural network is used as estimate of time series value for the next forecast.

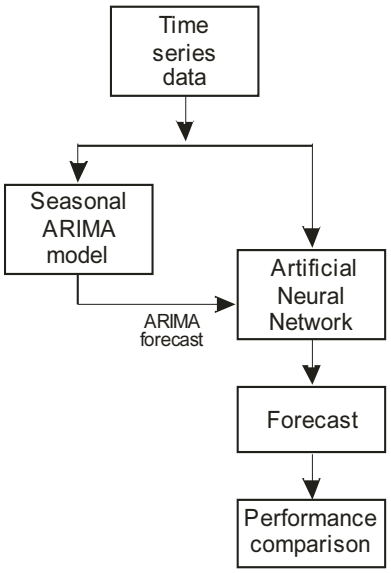

Figure 3. Block type pilot arrangement

TABLE II.
PERFORMANCE OF BEST NEURAL NETWORKS

Number of neurons in input layer	Number of neurons in hidden layer	RMSE (validation set)	MAPE (validation set)
14	4	7.01 e^{+04}	2.447 %
5	2	7.03 e^{+04}	2.827 %
5	1	7.06 e^{+04}	3.316 %
4	1	7.08 e^{+04}	3.751 %
10	2	7.11 e^{+04}	4.127 %

Three statistical tests are used to evaluate the performance of ARIMA and ARIMA-ANN hybrid models. These tests were realized using the well-known error functions as Root Mean Square Error (RMSE), Mean Absolute Deviation (MAD) and Mean Absolute Percentage Error (MAPE). These tests are defined as

$$RMSE = \sqrt{\frac{1}{n}\sum_{t=1}^{n}(v_t - p_t)^2} \tag{3}$$

$$MAD = \frac{1}{n}\sum_{t=1}^{n}\frac{|v_t - p_t|}{n} \tag{4}$$

$$MAPE = \sum_{t=1}^{n}\left|\frac{v_t - p_t}{v_t}\right| \cdot \frac{100}{n} \tag{5}$$

where n is number of forecasting period, v_t is actual time series value at time t and p_t is the predicted value of time series.

Prior to prediction, the known time series data (Fig. 4) are first divided into training set, which is being used to train the neural network and ARIMA model, and into validation set, which is used only for evaluation of forecast performance, and thus data from validation set are not used as inputs in third, and all prior stages of system usage.

In investigated hybrid system, time-delayed feed-forward neural network was employed with one hidden layer, which general topology is illustrated in Fig. 3. Input data is send through set of tapped delay lines (TDL) into neurons in the first layer. The number of input neurons depends on the number of inputs and the values of input delays. Outputs from the neurons of input layer are

then connected with every neuron in hidden layer. Similarly, the outputs from neurons in hidden layer are connected to every neuron in output layer. As for a time series prediction, we are interested only in one forecasted value, therefore the number of neurons in the output layer is 1. The optimal number of neurons in hidden layer is obtained trough trial and error method, e.g. using the computer simulations and choosing the neural network topology with the best performance.

Prior to prediction, near optimum network architecture should be found. In our experiments, three layered neural network has been used with different lengths of input data and number of neurons in hidden layer. Using different network architectures, we first trained all the networks using the back-propagation algorithm and then chose the network with best performance. The table 1 shows the results for few of the best neural networks.

4 Forecast Results

To show the efficiency of hybrid prediction model, the forecast performance is evaluated and compared with traditional ARIMA model. The time series data used for prediction shows significant seasonality tendency. For this reason, we modeled time series with seasonal ARIMA model, and determinated the best model based on Akaike's information criterion as ARIMA (1,0,1) (1,1,1) 12 model. The output from ARIMA model was then used as input into artificial neural network. The optimal number of input delays of neural network and number of neurons in hidden layer was identified using the computer simulations and based on comparison of performance on validation data. The best neural network topology was identified as 14 4 1 topology, that is neural network with 14 neurons in the input layer, 4 neurons in the hidden layer and 1 neuron in the output layer. The neural network was trained using the back-propagation algorithm using the Levenberg-Marquardt optimization for updating weights and biases values, with tan-sigmoid transfer function in hidden layers and linear transfer function in the output layer. The lags {1,2, … 14} were used for the input data. The performance comparison of prediction models is based on Root Mean Square Error (RMSE), Mean Absolute Deviation (MAD) and Mean Absolute Percentage Error (MAPE). The results shown in Table 2 shows that hybrid model outperforms the ARIMA model. Figure 5 shows the forecast results for ARIMA and ARIMA-NN model, with real values of time series, and residuals of the models are shown in Fig. 6.

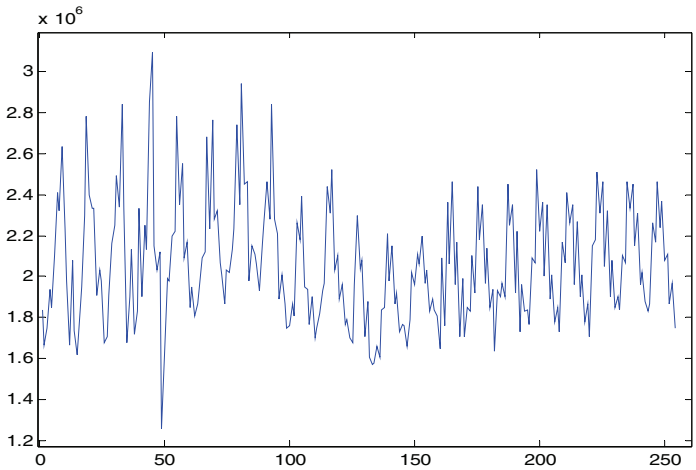

Figure 4. Time series data used for prediction

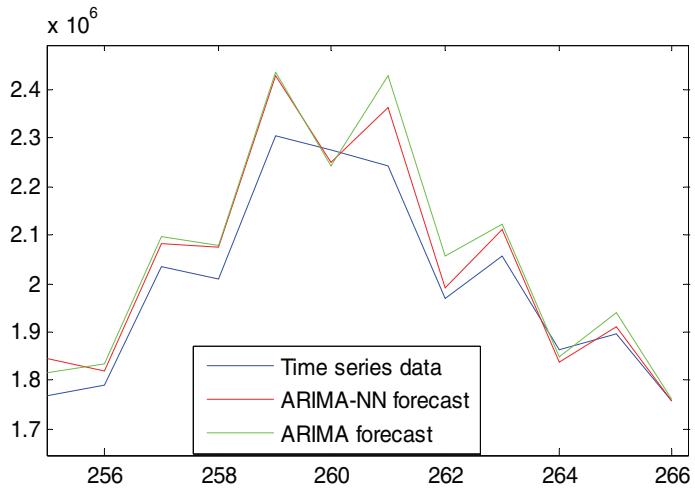

Figure 5. 12-step ahead forecast using ARIMA and ARIMA-NN hybrid system

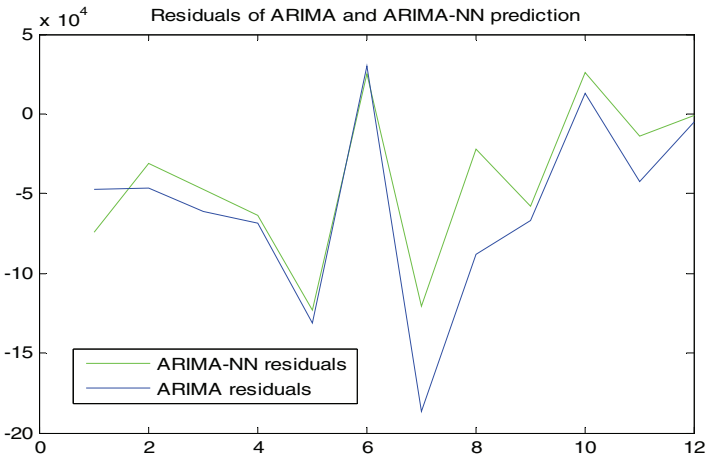

Figure 6. Residuals of ARIMA and ARIMA-NN hybrid system prediction

5 Conclusion

In this article we provided a forecast comparison of traditional seasonal ARIMA model and proposed ARIMA-NN hybrid model, based on time series data of aggregate water consumption of Spain. This case proved that ARIMA-NN hybrid prediction methods provide superior results than traditional ARIMA time series prediction model. The hybrid method takes advantage of the unique strength of ARIMA and NN in linear and nonlinear modelling of time series. The linear ARIMA model and nonlinear NN model are used jointly, aiming to improve the forecast prediction. For complex problems that have both linear and nonlinear relationships, the hybrid model can provide more accurate results. The results for water consumption prediction shows that this approach presents a superior and reliable alternative to traditional methods when choosing the appropriate number of delays and lagged input variables.

An extension to this work would be to investigate different alternative topologies of hybrid models and provide their performance comparison. In addition, the forecast models should be tested on different data sets in the future.

References

[1] BOX, G.E.P., JENKINS, G.M.: *Time Series Analyses, Forecasting and Control*, San Francisco: Holden day, 1976.

[2] CYBENKO, G.: *Approximation by superposition of sigmoidal functions*, Mathematics of Control, Signal and Systems, 2,1989.

[3] WHITE, H.: *Connectionist nonparametric regression: Multilayer feedforward networks can learn arbitrary mappings*, Neural Networks, 3, pp. 535-550, 1990.

[4] GALANT, A.R., WHITE, H.: *On learning the derivatives of an unknown mapping with multilayer feedforward neural network*, Neural networks, 5, pp. 129-138, 1992.

[5] HORNIK, K., STINCHCOMBE M., WHITE, H.: *Multilayer feedforward network are universal approximators*, Neural Networks, 2, pp. 359-366, 1989.

[6] TANG, Z., FISHWICK P.A.: *Feedforward neural nets as models for time series forecasting*, Jounal of Computing, pp. 374-385, 1993.

[7] ZHANG, G.: *Time series forecasting using a hybrid ARIMA and neural network model*, Journal of Neurocomputing, 50, pp. 159-175, 2003.

[8] GOMES, G.S.S., MAIA, A.L.S., LUDERMIR, T.B., CARVALHO, F., ARAUJO, A.F.R.: *Hybrid model with dynamic architecture for forecasting time series*, International Joint Conference on Neural Networks, pp.3742-3747, 2006.

[9] SALLEHUDDIN, R., SHAMSUDDIN, S. M., ZAITON S., HASHIM, M.: *Hybridization Model of Linear and Nonlinear Time Series Data for Forecasting*, Second Asia International Conference on Modelling and Simulation, pp.597-602, 2008.

[10] JAIN, A. & KUMAR, A.M.: *Hybrid neural network models for hydrologic time series forecasting*, Applied Soft Computing, Vol. 7, pp. 585-592, 2007.

[11] AKAIKE H.: *A new look at the statistical model identification*, IEEE Transactions on Automatic Control, 19, pp. 716-723, 1974.

[12] SKOKAN, M., BUNDZEL, M., SINCAK, P.: *Pseudo-distance based artificial neural network training*, 6th International Symposium on Applied Machine Intelligence and Informatics, pp.59-62 2008.

[13] DEHUAI, Z., JIANMIN, X., JIANWEI, G., LIYAN, L., GANG, X.: *Short Term Traffic Flow Prediction Using Hybrid ARIMA and ANN Models*, Workshop on Power Electronics and Intelligent Transportation System, pp.621-625, 2008.

Current address

Ing. Ján Šterba
University of Economics in Bratislava, Department of Statistics, Dolnozemská cesta 1/b, 852 35 Bratislava, Slovak Republic, tel. number: +421 949 524 714,
e-mail: sterba.jan@gmail.com

Ing. Katarína Hiľovská
Tecnical university of Košice, Faculty of Economics, Department of Banking and Investment, B. Nemcovej 32, 040 01 Košice, Slovak Republic, tel. number: +421 55 602 3263,
e-mail: katarina.hilovska@tuke.sk