

## Research Article

# A New Hybrid Methodology for Nonlinear Time Series Forecasting

**Mehdi Khashei and Mehdi Bijari**

*Department of Industrial Engineering, Isfahan University of Technology, Isfahan 84156-83111, Iran*

Correspondence should be addressed to Mehdi Khashei, khashei@in.iut.ac.ir

Received 7 March 2011; Revised 24 May 2011; Accepted 8 June 2011

Academic Editor: Andrzej Dzielinski

Copyright © 2011 M. Khashei and M. Bijari. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Artificial neural networks (ANNs) are flexible computing frameworks and universal approximators that can be applied to a wide range of forecasting problems with a high degree of accuracy. However, using ANNs to model linear problems have yielded mixed results, and hence; it is not wise to apply them blindly to any type of data. This is the reason that hybrid methodologies combining linear models such as ARIMA and nonlinear models such as ANNs have been proposed in the literature of time series forecasting. Despite of all advantages of the traditional methodologies for combining ARIMA and ANNs, they have some assumptions that will degenerate their performance if the opposite situation occurs. In this paper, a new methodology is proposed in order to combine the ANNs with ARIMA in order to overcome the limitations of traditional hybrid methodologies and yield more general and more accurate hybrid models. Empirical results with Canadian Lynx data set indicate that the proposed methodology can be a more effective way in order to combine linear and nonlinear models together than traditional hybrid methodologies. Therefore, it can be applied as an appropriate alternative methodology for hybridization in time series forecasting field, especially when higher forecasting accuracy is needed.

## 1. Introduction

Both artificial neural networks (ANNs) and autoregressive integrated moving average (ARIMA) models have achieved successes in their own linear or nonlinear domains. However, none of them is a universal model that is suitable for all circumstances. The approximation of ARIMA models to complex nonlinear problems as well as ANNs to model linear problems may be totally inappropriate and, also, in problems that consist of both linear and nonlinear correlation structures. Theoretical as well as empirical evidences in the literature suggest that, by using dissimilar models or models that disagree with each other strongly, the hybrid model will have lower generalization variance or error. In combined models, the aim is to reduce the risk of using an inappropriate model by combining several models to reduce the risk of failure and obtain results that are more accurate [1]. Typically, this is done because the underlying process cannot easily be determined. The motivation for using hybrid models comes from the assumption that either

one cannot identify the true data-generating process or that a single model may not be totally sufficient to identify all the characteristics of the time series [2].

Much effort has been devoted to develop and improve the hybrid time series models, since the early work of Reid [3] and Bates and Granger [4]. In pioneering work on combined forecasts, Bates and Granger showed that a linear combination of forecasts would give a smaller error variance than any of the individual methods. Since then, the studies on this topic have expanded dramatically. Combining linear and nonlinear models are one of the most popular and widely used hybrid models, which have been proposed and applied in order to overcome the limitations of each component and improve forecasting accuracy. Chen and Wang [5] constructed a combination model incorporating seasonal autoregressive integrated moving average (SARIMA) model and support vector machines (SVMs) for seasonal time series forecasting. Khashei et al. [6] presented a hybrid autoregressive integrated moving average and feedforward

neural network to time series forecasting in incomplete data situations, using the fuzzy logic.

Pai and Lin [7] proposed a hybrid method to exploit the unique strength of ARIMA models and support vector machines (SVMs) for stock prices forecasting. Yu et al. [8] proposed a novel nonlinear ensemble forecasting model integrating generalized linear autoregression (GLAR) with neural networks in order to obtain accurate prediction in foreign exchange market. Zhou and Hu [9] proposed a hybrid modeling and forecasting approach based on grey and Box-Jenkins autoregressive moving average models. Valenzuela et al. [10] presented a novel hybridization of intelligent techniques and ARIMA models for time series prediction. Aburto and Weber [11] presented a hybrid intelligent system combining autoregressive integrated moving average (ARIMA) models and neural networks for demand forecasting. Aladag et al. [12] proposed a new hybrid approach combining Elman's recurrent neural networks (ERNN) and ARIMA models for time series forecasting.

In all of these methods, the Zhang's hybrid methodology [13] is applied in order to combine the linear and nonlinear models together. This methodology includes three steps as follows. It must be first noted that, in the Zhang's hybrid methodology, a time series is considered to be composed of a linear autocorrelation structure and a nonlinear component as follows:

$$y_t = L_t + N_t, \quad (1)$$

where  $y_t$  denotes original time series,  $L_t$  denotes the linear component, and  $N_t$  denotes the nonlinear component.

(i) *Modeling the Linear Component.* In the first step, linear component is estimated by ARIMA model and residuals, which will contain only the nonlinear relationship, obtained from the ARIMA model as follows:

$$e_t = y_t - \hat{L}_t \quad (2)$$

where  $\hat{L}_t$  is the forecasting value for time  $t$  of the time series  $y_t$  estimated by ARIMA model. Zhang [13] claims that any ARIMA model can be selected for the data, as this does not affect the final forecast accuracy [12].

(ii) *Modeling the Nonlinear Component.* In the second step of the Zhang's hybrid methodology, a multilayer perceptron is used to model the ARIMA residuals. Zhang [13] claims that, by modeling ARIMA residuals by multilayer perceptron, nonlinear relationships can be discovered. With  $P$  input nodes, the ANN model for the residuals will be

$$\begin{aligned} e_t &= f(e_{t-1}, \dots, e_{t-p}) + \varepsilon_t \\ &= w_0 + \sum_{j=1}^Q w_j g \left( w_{0j} + \sum_{i=1}^p w_{i,j} e_{t-i} \right) + \varepsilon_t, \end{aligned} \quad (3)$$

where  $f$  is a nonlinear function determined by the multilayer perceptron and  $\varepsilon_t$  is the random error. Zhang [13] claims that if the model  $f$  is not an appropriate one, the error term is not necessarily random. Therefore, the correct model identification is critical.

(iii) *Combining the Linear and Nonlinear Components.* In the third step, the linear and nonlinear forecasting values obtained from (2) and (3) denoted as  $\hat{L}_t$  and  $\hat{N}_t$  respectively, are combined together as follows:

$$\hat{z}_t = \hat{L}_t + \hat{N}_t. \quad (4)$$

In the Zhang's hybrid methodology are jointly used the linear ARIMA and the nonlinear multilayer perceptrons models in order to capture different forms of relationship in the time series data. The motivation of the Zhang's hybrid methodology comes from the following perspectives. First, it is often difficult in practice to determine whether a time series under study is generated from a linear or nonlinear underlying process; thus, the problem of model selection can be eased by combining linear ARIMA and nonlinear ANN models. Second, real-world time series are rarely pure linear or nonlinear and often contain both linear and nonlinear patterns, where neither ARIMA nor ANN models alone can be adequate for modeling in such cases; hence, the problem of modeling the combined linear and nonlinear autocorrelation structures in time series can be solved by combining linear ARIMA and nonlinear ANN models. Third, it is almost universally agreed in the forecasting literature that no single model is the best in every situation, due to the fact that a real-world problem is often complex in nature and any single model may not be able to capture different patterns equally well. Therefore, the chance in order to capture different patterns in the data can be increased by combining different models.

Although the Zhang's hybrid methodology has been shown to be successful for single models in several studies, it has some assumptions [14] that will degenerate its performance if the opposite situation occurs; therefore, it may be inadequate in some specific situations. In this methodology, it is assumed that (1) the existing linear and nonlinear patterns in a time series can be separately modeled, (2) the residuals from the linear model contain only the nonlinear relationship, and (3) the relationship between the linear and nonlinear components is additive. However, these assumptions may underestimate the relationship between the components and degrade performance if the opposite situation occurs [15], for example, if we cannot separately model the linear and nonlinear patterns in a time series, the residuals of the linear component do not comprise valid nonlinear patterns, or there is not any additive association between the linear and nonlinear elements and the relationship is different (e.g., multiplicative).

In this paper, a new methodology is proposed in order to combine the linear and nonlinear models for better modeling both linear and nonlinear components, simultaneously. This methodology has no aforementioned assumptions of the Zhang's hybrid methodology. The proposed methodology consists of two stages. In the first stage, a linear model such as autoregressive integrated moving average (ARIMA), generalized linear autoregression (GLAR), and so forth is used in order to identify and magnify the existing linear component in data. In the second stage, a nonlinear model such as multilayer perceptron (MLP),

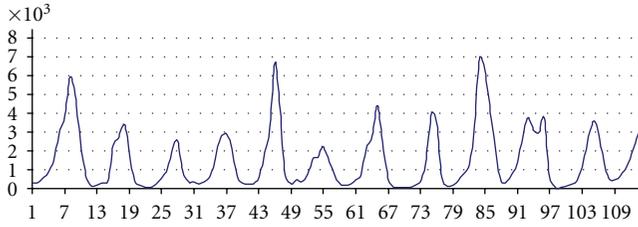


FIGURE 1: Canadian lynx data series (1821–1934).

support vector machine (SVM), Elman's recurrent neural network (ERNN), and so forth. is used in order to model the preprocessed data. The rest of the paper is organized as follows. In the next section, the formulation of the proposed hybrid methodology is introduced. In Section 3, in order to show the appropriateness and effectiveness of the proposed methodology, it is applied to the Zhang's model and Aladag's model, respectively, presented in [12, 13] and its performance is evaluated in comparison with the Zhang's hybrid methodology. Conclusions will be the final section of the paper.

## 2. The Proposed Hybrid Methodology

In our proposed methodology, a time series is considered as function of a linear and a nonlinear component. Thus,

$$y_t = f(L_t, N_t), \quad (5)$$

where  $L_t$  denotes the linear component and  $N_t$  denotes the nonlinear component. In the first stage, a linear model such as autoregressive integrated moving average (ARIMA), generalized linear autoregression (GLAR), and so forth is used in order to model the linear component. The residuals from the first stage will contain the nonlinear relationship, in which linear model is not able to model it, and maybe linear relationship [14]. Thus,

$$L_t = \hat{L}_t + e_t, \quad (6)$$

where  $\hat{L}_t$  is the forecasting value for time  $t$  estimated by the linear model and  $e_t$  is the residual at time  $t$  from the linear model. The forecasted values and residuals of linear modeling are the results of first stage that are used in next stage. In addition, the linear patterns are magnified by linear model in order to be applied in second stage.

In second stage, a nonlinear model such as multilayer perceptron (MLP), support vector machine (SVM), Elman's recurrent neural network (ERNN), and so forth is used in order to model the nonlinear and probable linear relationships existing in residuals of linear modeling and original data. Thus,

$$\begin{aligned} N_t^1 &= f^1(e_{t-1}, \dots, e_{t-n}), \\ N_t^2 &= f^2(z_{t-1}, \dots, z_{t-m}), \\ N_t &= f(N_t^1, N_t^2), \end{aligned} \quad (7)$$

where  $f^1$ ,  $f^2$ , and  $f$  are the nonlinear functions determined by the nonlinear model.  $N$  and  $m$  are integers and often referred to as orders of the model. Thus, the combined forecast will be as follows:

$$y_t = f(N_t^1, \hat{L}_t, N_t^2) = f(e_{t-1}, \dots, e_{t-n}, \hat{L}_t, z_{t-1}, \dots, z_{t-m}), \quad (8)$$

where  $f$  are the nonlinear functions determined by the nonlinear model.  $n_1 \leq n$  and  $m_1 \leq m$  are integers that are determined in design process of the nonlinear model. It must be noted that anyone of the aforementioned variables  $e_i$  ( $i = t-1, \dots, t-n$ ),  $\hat{L}_t$ , and  $z_j$  ( $j = t-1, \dots, t-m$ ) or set of them  $\{e_i (i = t-1, \dots, t-n)\}$  or  $\{z_i (i = t-1, \dots, t-m)\}$  may be deleted in design process of the nonlinear model. This may be related to the underlying data-generating process and the existing linear and nonlinear structures in data. For example, if data only consist of pure linear structure, then  $\{e_i (i = t-1, \dots, t-n)\}$  variables will be probably deleted against other of those variables. In contrast, if data only consist of pure nonlinear structures, then  $\hat{L}_t$  variable will be probably deleted against other of those variables.

## 3. Application of the Hybrid Methodology

In this section, the proposed hybrid methodology is applied to the Zhang's model and Aladag's model for Canadian lynx data forecasting. This data consists of the set of annual numbers of lynx trappings in the Mackenzie River District of North-West Canada for the period from 1821 to 1934. The Canada lynx data, which is plotted in Figure 1, was also examined by Kajitani et al. [16], beyond the other various studies in the time series literature with a focus on the nonlinear modeling [17]. Following other studies, the logarithms (to the base 10) of the data is used in the analysis. The proposed hybrid methodology is first applied to the Zhang's model as follows. It must be noted that all calculations are performed on a personal computer that has a "Pentium(R) 4 CPU 3.2 GHz, RAM 1.0 GB" processor and Windows XP.

*Stage I.* Using the Eviews package software, the established model is a autoregressive model of order twelve,  $AR(12)$ , which has also been used by many researchers [4].

*Stage II.* Using pruning algorithms [18] in MATLAB7 package software, the best-fitted network which is selected is composed of eight inputs, three hidden and one output neurons (in abbreviated form,  $N^{(8-3-1)}$ ). The structure of the best-fitted network is shown in Figure 2. The performance measures of the proposed methodology for the Zhang's process modeling for Canadian lynx data are given in Table 1. The estimated values of the proposed methodology for the Zhang's process modeling for Canadian lynx data set are plotted in Figure 3. In addition, the estimated values of ARIMA, ANN, and the proposed methodology for the Zhang's process modeling for the last 14 observations (test data) are, respectively, plotted in Figures 4, 5, and 6. In addition, a feedforward neural network, which is composed

TABLE 1: Comparison of the performance of the proposed methodology.

Model	MSE
Autoregressive integrated moving average (ARIMA)	0.020
Feedforward neural network (FNN)	0.020
Zhang's hybrid model	0.017
Proposed methodology (Zhang's process)	<b>0.009</b>
Aladag's hybrid model	0.009
Proposed methodology (Aladag's process)	<b>0.006</b>

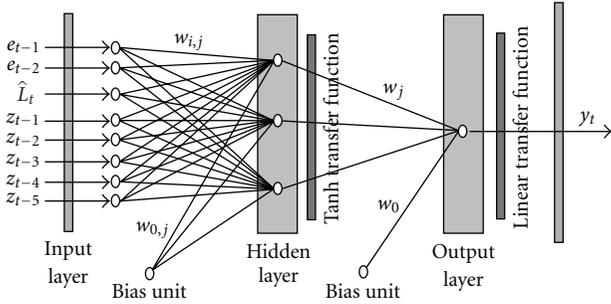


FIGURE 2: Structure of the best-fitted network (lynx data),  $N^{(8-3-1)}$ .

of seven inputs, five hidden and one output neurons ( $N^{(7-5-1)}$ ), has been designed to Canadian lynx data set forecast, as also employed by Zhang [13]. Similar to the above, the proposed methodology is applied to the Aladag's process modeling for Canadian lynx data forecasting. The performance measures of this case are also given in Table 1.

Numerical results from Table 1 show that both Zhang's hybrid model and Aladag's hybrid model by using the Zhang's methodology outperform the autoregressive integrated moving average (ARIMA) and feedforward neural network (FNN) models for Canadian lynx data. Moreover, Aladag et al. [12] can make a more accurate hybrid model, by replacing the feedforward neural network (FNN) model by Elman's recurrent neural network (ERNN) for Canadian lynx data. However, according to the obtained results, it can be seen that using the proposed hybrid methodology instead of using the Zhang's hybrid methodology for combination improves the accuracy in both Zhang's model and Aladag's model. Our proposed methodology indicates 47.06% and 33.34% improvement over the Zhang's hybrid methodology in Zhang's hybrid model and Aladag's hybrid model in MSE, respectively. In addition, the computation times of the constructed models based on the proposed methodology are as fast as models that are constructed based on the Zhang's methodology. These evidences clearly show that the performance of the proposed methodology is better than Zhang's methodology, for hybridization of linear and nonlinear models together for forecasting.

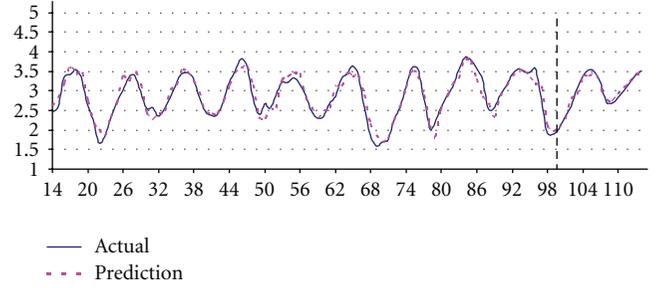


FIGURE 3: Results obtained from the proposed model for Canadian lynx data set.

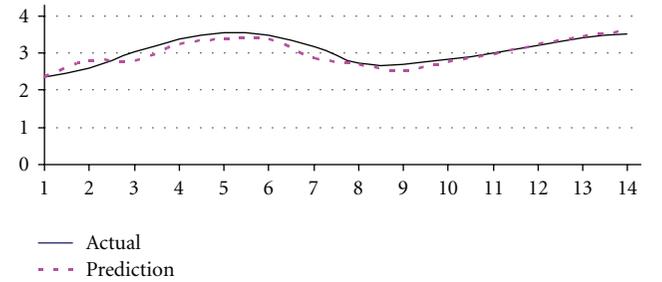


FIGURE 4: ARIMA model prediction of lynx data (test sample).

## 4. Conclusions

The Zhang hybrid methodology that decomposes a time series into its linear and nonlinear form is one of the most popular hybrid methodologies, which have been shown to be successful for single models. This methodology jointly uses the linear and nonlinear models in order to capture different forms of relationship in the time series data. However, some researchers believe that assumptions that are considered in constructing process of this hybrid methodology can degenerate its performance if the opposite situation occurs. In addition, it cannot be generally guaranteed that the performance of the designed models based on this methodology will not be worse than their nonlinear component. These assumptions are as follows.

- (1) This methodology assumes that the linear and nonlinear patterns of a time series can be separately modeled by different models.
- (2) This methodology assumes that the relationship between the linear and nonlinear components is additive.
- (3) This methodology assumes that the residuals from the linear model will contain only the nonlinear relationship.

In this paper, we propose a new methodology in order to combine the linear and nonlinear models that has no above-mentioned assumptions of traditional hybrid linear and nonlinear models in order to yield the more general and the more accurate forecasting model. Empirical results with Canadian Lynx data set indicate that our proposed methodology can

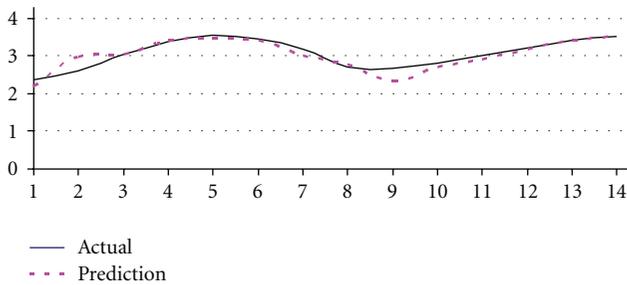


FIGURE 5: ANN model prediction of lynx data (test sample).

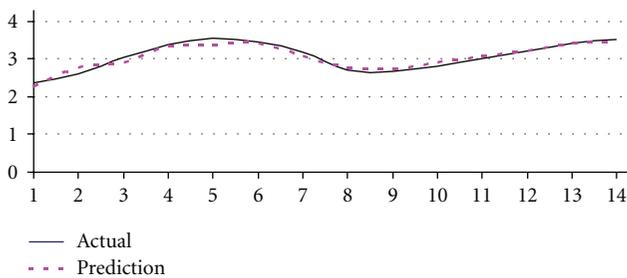


FIGURE 6: Proposed model prediction of lynx data (test sample).

improve the performance of the designed hybrid models by the Zhang's methodology. These results confirm this hypothesis that the aforementioned assumptions considered in constructing process of the traditional hybrid linear and nonlinear methodologies will degenerate their performance if the opposite situation occurs. In addition, in contrast to the traditional hybrid linear and nonlinear methodologies, we can generally guarantee that the performance of the designed models by proposed methodology will not be worse than either of the components used in isolation, so that it can be applied as an appropriate methodology for combination linear and nonlinear models for time series forecasting.

## Acknowledgments

The authors wish to express their gratitude to Seyed Reza Hejazi, Isfahan University of Technology, for their insightful and constructive comments, which helped to improve the paper greatly.

## References

- [1] M. Hibon and T. Evgeniou, "To combine or not to combine: selecting among forecasts and their combinations," *International Journal of Forecasting*, vol. 21, no. 1, pp. 15–24, 2005.
- [2] N. Terui and H. K. van Dijk, "Combined forecasts from linear and nonlinear time series models," *International Journal of Forecasting*, vol. 18, no. 3, pp. 421–438, 2002.
- [3] M. J. Reid, "Combining three estimates of gross domestic product," *Economica*, vol. 35, pp. 431–444, 1968.
- [4] J. M. Bates and C. W. J. Granger, "Combination of forecasts," *Operational Research*, vol. 20, no. 4, pp. 451–468, 1969.
- [5] K. Y. Chen and C. H. Wang, "A hybrid SARIMA and support vector machines in forecasting the production values

- of the machinery industry in Taiwan," *Expert Systems with Applications*, vol. 32, no. 1, pp. 254–264, 2007.
- [6] M. Khashei, M. Bijari, and G. A. Raissi Ardali, "Improvement of Auto-Regressive Integrated Moving Average models using Fuzzy logic and Artificial Neural Networks (ANNs)," *Neurocomputing*, vol. 72, no. 4–6, pp. 956–967, 2009.
- [7] P. F. Pai and C. S. Lin, "A hybrid ARIMA and support vector machines model in stock price forecasting," *Omega*, vol. 33, no. 6, pp. 497–505, 2005.
- [8] L. Yu, S. Wang, and K. K. Lai, "A novel nonlinear ensemble forecasting model incorporating GLAR and ANN for foreign exchange rates," *Computers and Operations Research*, vol. 32, no. 10, pp. 2523–2541, 2005.
- [9] Z. J. Zhou and C. H. Hu, "An effective hybrid approach based on grey and ARMA for forecasting gyro drift," *Chaos, Solitons and Fractals*, vol. 35, no. 3, pp. 525–529, 2008.
- [10] O. Valenzuela, I. Rojas, F. Rojas et al., "Hybridization of intelligent techniques and ARIMA models for time series prediction," *Fuzzy Sets and Systems*, vol. 159, no. 7, pp. 821–845, 2008.
- [11] L. Aburto and R. Weber, "Improved supply chain management based on hybrid demand forecasts," *Applied Soft Computing Journal*, vol. 7, no. 1, pp. 136–144, 2007.
- [12] C. H. Aladag, E. Egrioglu, and C. Kadilar, "Forecasting nonlinear time series with a hybrid methodology," *Applied Mathematics Letters*, vol. 22, no. 9, pp. 1467–1470, 2009.
- [13] P. P. Zhang, "Time series forecasting using a hybrid ARIMA and neural network model," *Neurocomputing*, vol. 50, pp. 159–175, 2003.
- [14] T. Taskaya-Temizel and M. C. Casey, "A comparative study of autoregressive neural network hybrids," *Neural Networks*, vol. 18, no. 5–6, pp. 781–789, 2005.
- [15] M. Khashei and M. Bijari, "A novel hybridization of artificial neural networks and ARIMA models for time series forecasting," *Applied Soft Computing*, vol. 11, pp. 2664–2675, 2011.
- [16] Y. Kajitani, K. W. Hipel, and A. I. Mcleod, "Forecasting nonlinear time series with feed-forward neural networks: a case study of Canadian lynx data," *Journal of Forecasting*, vol. 24, no. 2, pp. 105–117, 2005.
- [17] M. Khashei and M. Bijari, "An artificial neural network (p, d, q) model for timeseries forecasting," *Expert Systems with Applications*, vol. 37, no. 1, pp. 479–489, 2010.
- [18] M. Khashei, S. Reza Hejazi, and M. Bijari, "A new hybrid artificial neural networks and fuzzy regression model for time series forecasting," *Fuzzy Sets and Systems*, vol. 159, no. 7, pp. 769–786, 2008.



**Hindawi**

Submit your manuscripts at  
<http://www.hindawi.com>

