Applied Soft Computing xxx (2014) xxx–xxx

Contents lists available at ScienceDirect

Applied Soft Computing

journal homepage: www.elsevier.com/locate/asoc

A moving-average-filter-based hybrid ARIMA–ANN model for forecasting time series data

C. Narendra Babu*, B. Eswara Reddy

Department of Computer Science & Engineering, JNT University College of Engineering, Anantapuramu, India

ARTICLE INFO

Article history:
Received 26 July 2013
Received in revised form 2 May 2014
Accepted 5 May 2014
Available online xxx

Keywords:
Time series forecasting
ARIMA
ANN
Hybrid model
Box–Jenkins methodology

ABSTRACT

A suitable combination of linear and nonlinear models provides a more accurate prediction model than an individual linear or nonlinear model for forecasting time series data originating from various applications. The linear autoregressive integrated moving average (ARIMA) and nonlinear artificial neural network (ANN) models are explored in this paper to devise a new hybrid ARIMA–ANN model for the prediction of time series data. Many of the hybrid ARIMA–ANN models which exist in the literature apply an ARIMA model to given time series data, consider the error between the original and the ARIMA-predicted data as a nonlinear component, and model it using an ANN in different ways. Though these models give predictions with higher accuracy than the individual models, there is scope for further improvement in the accuracy if the nature of the given time series is taken into account before applying the models. In the work described in this paper, the nature of volatility was explored using a moving-average filter, and then an ARIMA and an ANN model were suitably applied. Using a simulated data set and experimental data sets such as sunspot data, electricity price data, and stock market data, the proposed hybrid ARIMA–ANN model was applied along with individual ARIMA and ANN models and some existing hybrid ARIMA–ANN models. The results obtained from all of these data sets show that for both one-step-ahead and multistep-ahead forecasts, the proposed hybrid model has higher prediction accuracy.

Introduction and related work

Time series forecasting is now a very important research area, owing to the importance of prediction in various applications. Forecasting Internet traffic helps service providers to enhance their services. Forecasting climate change helps the agricultural sector. Forecasting disasters aids in taking necessary precautions and helps mankind to be prepared. Forecasting financial data helps investors to invest safely in the market. However, time series data do not always have the same characteristics. Some time series are seasonal: for example, road traffic is high at some particular times in the day, climate variations repeat according to the seasons, etc. Some other time series are nonseasonal, such as financial and stock market data. Some time series are highly volatile, such as wind speed data, and some are less volatile, such as global temperature and annual rainfall. Some data are almost linear in nature, such as the growth of an animal, plant, or human being, but many other sets of data are nonlinear in nature. Some data are Gaussian in nature, whereas some other data are non-Gaussian. For forecasting such time series data, various prediction techniques have been proposed in the literature, which may use either linear or nonlinear models. A type of linear model, namely the autoregressive integrated moving average (ARIMA), and a type of nonlinear model, the artificial neural network (ANN), were chosen for study in this paper.

ARIMA models assume that the present data are a linear function of past data points and past errors. They also assume that the errors are white in nature, and require that the data be made stationary before fitting a linear equation to the data. In the literature, ARIMA models have been applied to various time series data, such as electricity prices [1,2], sugar prices [3], stock market data [4], and wind speeds [5], for the prediction of future values. ARIMA models can help in understanding the dynamics of the data in a given application. Before forecasting time series data, various preprocessing steps can be applied to the raw data if necessary. In [6], a wavelet transformation was applied before forecasting of global temperature data. In [7], new classification and feature extraction techniques were proposed for electrocardiography data. These preprocessing steps can be applied to the raw data to obtain more accurate predictions. In this paper, the basic ARIMA model was chosen as the linear prediction model.

One efficient nonlinear technique for time series forecasting is ANNs [8]. ANNs are advantageous compared with ARIMA in many

* Corresponding author. Tel.: +91 9035644182.
E-mail address: narendrababu.c@gmail.com (C.N. Babu).

http://dx.doi.org/10.1016/j.asoc.2014.05.028
1568-4946/

applications because ANNs do not assume linearity. They are capable of fitting a nonlinear function to the given data and do not need the data to be made stationary. They are also adaptive in nature. For these reasons, ANN models have become more popular in forecasting. ANNs have been applied to various time series data, such as electricity demand data [9], financial data [10], river flow data [11], and network data [12], for forecasting. In all these cases ANNs were shown to yield good forecasts compared with ARIMA models. In [13], neural networks were used to predict earthquakes in Chile. There is no universal model which can suit all applications. The prediction accuracy can be improved if two different models are applied to the same data rather than a single model. So, many hybrid techniques have been proposed in the literature, which combine the advantages of two or more individual models.

A hybrid ARIMA–ANN model was proposed by Zhang [14], which was shown to give more accurate predictions than the individual models. On Wolf's sunspot data, Canadian lynx data, and exchange rate time series data, this hybrid model was shown to outperform individual ARIMA and ANN models in the case of one-step-ahead prediction. Another hybrid ARIMA–ANN method was proposed by Khashei and Bijari [15], which was shown to give better performance for one-step-ahead forecasting than the method proposed by Zhang [14]. The hybrid method proposed by Zhang was also used for electricity price forecasting in [16] and for water quality time series prediction in [17].

In addition to the ARIMA and ANN models, many other prediction models are available in the literature. Some of them are based on support vector machines [18], and some others on fuzzy models [19]. Modified forms of ARIMA models such as SARIMA [20] and FARIMA [21] are also available. GARCH models have also used for financial forecasting, as in [22]. In [23,24], autoregressive (AR) and GARCH models were used for forecasting financial data. Spectral techniques based on SVD were proposed in [16] and the references therein. Decomposition-based ARIMA, ARIMA–GARCH, and ANN models were also studied in [25–27] and elsewhere. Most of the time it is seen that if hybrid models are used instead of one model the prediction accuracy improves, but if the hybrid uses many decompositions and many models, then the accuracy will degrade after some limit and the model will no longer be successful. So, hybrid models should contain a limited number of individual models to retain the simplicity of the model as well as to retain accuracy.

In this paper, a new hybrid ARIMA–ANN prediction model is proposed which is suitable for both one-step-ahead and multistep-ahead predictions. In this model, the nature of the time series is first explored with the help of a moving-average (MA) filter and then ARIMA and ANN models are suitably applied. Instead of directly applying the ARIMA or ANN model to the given data, the nature of the volatility of the time series data is first understood, and then the ARIMA and ANN models are applied to the data. This proposed method of prediction was applied to a known data set that was generated by adding linear and nonlinear time series data. The performance accuracy of this method was better than the accuracy obtained by the models of Zhang [14] and Khashei and Bijari [15]. On three different time series data, namely sunspot, electricity price, and financial data, for both one-step and multistep predictions, the proposed method gave better accuracy than the other hybrids [14,15].

The rest of the paper is organized as follows. In "ARIMA- and ANN-based modeling techniques" section, the ARIMA, ANN, and some existing hybrid ARIMA–ANN models are described. The proposed method is discussed in detail in "Proposed hybrid ARIMA–ANN model" section. In "Results and discussion" section, the results are discussed in four subsections, along with tables of performance measures and graphs of predicted values. "Conclusion" section ends the paper with a conclusion.

ARIMA- and ANN-based modeling techniques

Detailed descriptions of the ARIMA and ANN modeling techniques are available in the literature. However, to make the discussion complete, a brief description of these modeling methods is presented in this section. Also, some of the popular existing hybrid modeling schemes are outlined.

ARIMA

ARIMA is a linear modeling technique. In this technique, the given time series data are first checked for stationarity. If they are not stationary, a differencing operation is performed. If the data are still nonstationary, differencing is again performed until the data are finally made stationary. If the differencing is performed d times, the integration order of the ARIMA method is said to be d. The resultant data are modeled as an autoregressive moving average (ARMA) time series as follows. The data value at any given time t, say y_t, is considered as a function of the previous p data values, say $y_{t-1}, y_{t-2}, \ldots, y_{t-p}$, and the errors at times $t, t-1, \ldots, t-q$, say n_t, $n_{t-1}, \ldots, n_{t-q}$. The corresponding ARMA equation is shown in (1). In (1), a_1 to a_p are the autoregressive (AR) coefficients and b_1 to b_q are the MA coefficients. Thus the time series model is denoted as ARIMA(p, d, q). The ARMA model assumes that the error sequence n_t is white noise and is Gaussian distributed, so the variance of this error is also a model parameter. The ARIMA modeling procedure has three steps: (a) identifying the model order, i.e., identifying p and q; (b) estimating the model coefficients; and (c) forecasting the data.

Identifying the orders p,q is done using correlation analysis [28], using the nature of the autocorrelation function and partial autocorrelation function. The model coefficients are estimated using the Box–Jenkins method [28]. Of the various estimation approaches other than the nonlinear maximum likelihood approach, which is computationally more complex, Gaussian maximum likelihood estimation (GMLE) approaches [29] are generally used for estimation of the ARIMA model parameters. The model is validated by finding the Akakine information criterion (AIC). The best-suited ARIMA model has the minimum value of the AIC. After all the model coefficients have been estimated, the next values of the time series are predicted using the available past data values and the model coefficients. ARIMA models predict linear time series data with very good accuracy.

$$y_t = a_1 y_{t-1} + a_2 y_{t-2} + \cdots + a_p y_{t-p} + n_t + b_1 n_{t-1} + \cdots + b_q n_{t-q} \quad (1)$$

ANN

This is a nonlinear modeling technique which is suitable for modeling over a very wide range of applications. It is more flexible in terms of architecture. The neural-network architecture bears a high similarity to the neurons in the brain, hence the name "artificial neural network." In this network architecture, there may be two or more layers. For example, a three-layer ANN has three layers, namely an input layer, a hidden layer, and an output layer. The inputs can be of any number. Also, the number of neurons in the hidden layer is flexible. A typical three-layer ANN, with the nomenclature, is shown in Fig. 1. The neurons are processing units which are acyclically linked. Three-layer ANNs are widely used for time series forecasting. To model time series data using such a network, the sequence y_t is considered as a nonlinear function of $y_{t-1}, \ldots, y_{t-N}$. The corresponding equation is shown in (2). In (2), the function g is a nonlinear function, and v_t is a noise or error term. The ANN model output can be represented in terms of input and hidden-layer weight parameters. The transfer function of the hidden layer is generally a sigmoid function, shown in (3), and that of the output

Please cite this article in press as: C.N. Babu, B.E. Reddy, A moving-average-filter-based hybrid ARIMA–ANN model for forecasting time series data, Appl. Soft Comput. J. (2014), http://dx.doi.org/10.1016/j.asoc.2014.05.028

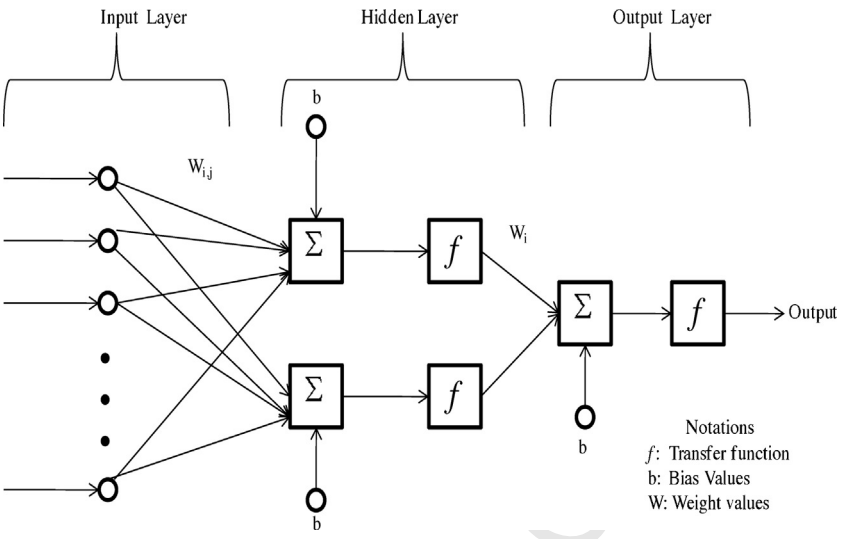

Fig. 1. Three-layer ANN architecture.

layer is a linear function. The model coefficients in an ANN are the weights of each link and the bias values, as shown in Fig. 1.

$$y_t = g(y_{t-1}, y_{t-2}, \ldots, y_{t-N}) + v_t \quad (2)$$

$$\text{Sigmoid}(x) = \frac{1}{1+e^{-x}} \quad (3)$$

To obtain these model coefficients, a known data sequence is first given as input to the ANN, and the ANN is trained with this training sequence. Training a neural network can be considered as minimizing the multivariate global error function formed by the weight values. After training of the ANN, its performance is tested and validated. Many other training algorithms are available in the literature [30]. For example, in [14], a reduced gradient-based algorithm was used for training the ANN. In [17] and [15], a scaled conjugate gradient algorithm and a Levenberg–Marquardt (LM) training algorithm, respectively, were incorporated. In the work presented in this paper, the ANN model was trained using an LM training algorithm.

The performance of ANNs is better than that of the ARIMA method in many applications. This is because ANNs can model nonlinear time series data accurately, as seen from the model equation in (2).

Hybrid models using ARIMA and ANN

Often, the given data may have both linear and nonlinear characteristics. So, hybrid models using both ANN and ARIMA methods are better than individual models for obtaining accurate predictions. Two existing ARIMA–ANN hybrid models considered for discussion in this paper are illustrated as follows.

Zhang's hybrid ARIMA–ANN model

In 2003, Zhang proposed a hybrid ARIMA–ANN model for time series forecasting. According to this model, any time series data are assumed to be the sum of two components, linear and nonlinear. First, an ARIMA model is fit to the given time series data. Then the error sequence is assumed to be the nonlinear component and is modeled using an ANN. The predictions obtained from both the ARIMA model and the ANN model are combined to obtain the final forecast. This is suitable for both one-step-ahead and multistep-ahead predictions. This model was shown to be better than the individual ARIMA or ANN models in many applications in terms of prediction accuracy [16].

Khashei and Bijari's hybrid ARIMA–ANN model

In 2010, Khashei and Bijari proposed another hybrid ARIMA–ANN model for time series forecasting. This model also assumes that any time series data are the sum of two components, linear and nonlinear. In this method, first an ARIMA model is fit to the given time series data and one data value is forecasted. Then the past original data values, the present ARIMA-forecasted data value, and the past error sequence values are all given as inputs to an ANN, and the output of the ANN is the final forecasted value. This method was shown to have better performance than Zhang's method in various applications in terms of prediction accuracy [15].

Proposed hybrid ARIMA–ANN model

In this paper, a new hybrid ARIMA–ANN model is proposed, which is outlined in this section. The technique first characterizes the given data based on the nature of the volatility of the data. Then ARIMA and ANN models are suitably applied. Before describing this technique, we first discuss some interesting facts about ARIMA sequences, which are used in understanding and characterizing the given data.

In the hybrid methods proposed by Zhang [14] and by Khashei and Bijari [15], the data are assumed to be the sum of linear and nonlinear components. But the given data are not decomposed into linear and nonlinear components; instead, a linear ARIMA model is fit directly to the data and the error sequence thus obtained is assumed to be the nonlinear component. Thus both of these hybrid methods explore and use the fact that the ARIMA model is linear. Ideal ARIMA sequences have many interesting properties, two of which are linearity and stationarity. Some other statistical facts about ARIMA and ARMA sequences are the following: first, the error sequence n_t in (1) is Gaussian or normally distributed and white in nature [28], and second, a Gaussian time series represented as a random vector $[\, y_t \quad y_{t-1} \quad y_{t-2} \quad \cdots \,]$ is joint-Gaussian in nature [31].

The second statistical fact can be explored further as follows. A stationary Gaussian time series is always stationary in the strict sense [31]. So, assuming that a given ARMA time series is strictly stationary, one possibility is that this series is a Gaussian time series. Usually, after making the given time series data stationary, estimation of the ARIMA model coefficients is performed using GMLE [29]. In this estimation, the model coefficients are obtained as if the given time series were Gaussian. So, if the given stationary

Please cite this article in press as: C.N. Babu, B.E. Reddy, A moving-average-filter-based hybrid ARIMA–ANN model for forecasting time series data, Appl. Soft Comput. J. (2014), http://dx.doi.org/10.1016/j.asoc.2014.05.028

time series is truly Gaussian, then the estimated ARIMA model is a better fit. So, it can be concluded that if the time series is stationary in the strict sense, an ARIMA model is more suitable for Gaussian time series data. Then the random vector $[\,y_t \quad y_{t-1} \quad y_{t-2} \quad \cdots\,]$ is joint-Gaussian and each random variable y_t is Gaussian distributed.

In general, to diagnose whether a given sequence is normally distributed or not, the Jarque–Bera normality test can be performed. A part of this test checks whether the kurtosis of the sequence, given in (4), is 3 or not:

$$\text{kurtosis} = \frac{E\{(y - E\{y\})^4\}}{(E\{(y - E\{y\})^2\})^2} \tag{4}$$

In (4), y is the random variable for which the kurtosis is being computed, and E stands for the expectation operation. If the kurtosis value is 3, then the sequence is Gaussian; such sequences were considered as low-volatility data in the work described in the present paper. Sequences that did not have a kurtosis value of 3 were considered as highly volatile data. A highly volatile time series is either leptokurtic or platykurtic in nature, which means that the distribution is non-Gaussian. Thus we can conclude that ARIMA models are suitable for any time series data when the data have a kurtosis value of approximately 3. With this understanding, the proposed model is described below.

Mathematically, the proposed model can be described as follows. The time series data y_t are considered as a sum of a low-volatility component l_t and a high-volatility component h_t, as given in (5). After making sure that l_t is stationary, it is modeled as a linear function of past values of the sequence $l_{t-1}, l_{t-2}, \ldots, l_{t-p}$ and the random error sequence $n_t, n_{t-1}, \ldots, n_{t-q}$ using an ARIMA model. This is shown in (6), where f is a linear function. Similarly, h_t is expressed as a nonlinear function of $h_{t-1}, h_{t-2}, \ldots, h_{t-N}$ as shown in (7), and is modeled using an ANN. In (7), g represents the nonlinear function, and ε_t represents the model error. Using the ARIMA-predicted low-volatility component $\hat{l}_t$ and the ANN-predicted high-volatility component $\hat{h}_t$, the predicted time series value $\hat{y}_t$ is obtained as represented in (8).

$$y_t = l_t + h_t \tag{5}$$

$$\hat{l}_t = f(l_{t-1}, l_{t-2}, \ldots, l_{t-p}, n_t, n_{t-1}, \ldots, n_{t-q}) \tag{6}$$

$$\hat{h}_t = g(h_{t-1}, h_{t-2}, \ldots, h_{t-N}) + \varepsilon_t \tag{7}$$

$$\hat{y}_t = \hat{l}_t + \hat{h}_t \tag{8}$$

The steps of the algorithm for the proposed hybrid model are given below and are represented as a flow chart in Fig. 2.

1 Using an MA filter, given in (9), the given time series data are separated or decomposed into two components such that one of the components is less volatile and the other is highly volatile. The length of the MA filter, m, is adjusted so that this decomposition is properly achieved. The first decomposition is y_{tr}, given in (9), which is the smoothed trend component, and has low volatility. The second decomposition obtained from the MA filter is the residual component, given in (10), which has high volatility.
2 The low-volatility component with $k = 3$ is modeled using an ARIMA model and the predictions are obtained as in (6).
3 The high-volatility component with $k \neq 3$ is modeled using an ANN and the predictions are obtained as in (7).
4 The predictions obtained from steps 2 and 3 are added to obtain the final predictions as in (8):

$$y_{tr} = \frac{1}{m} \sum_{i=t-m+1}^{t} y_i \tag{9}$$

$$y_{res} = y_t - y_{tr} \tag{10}$$

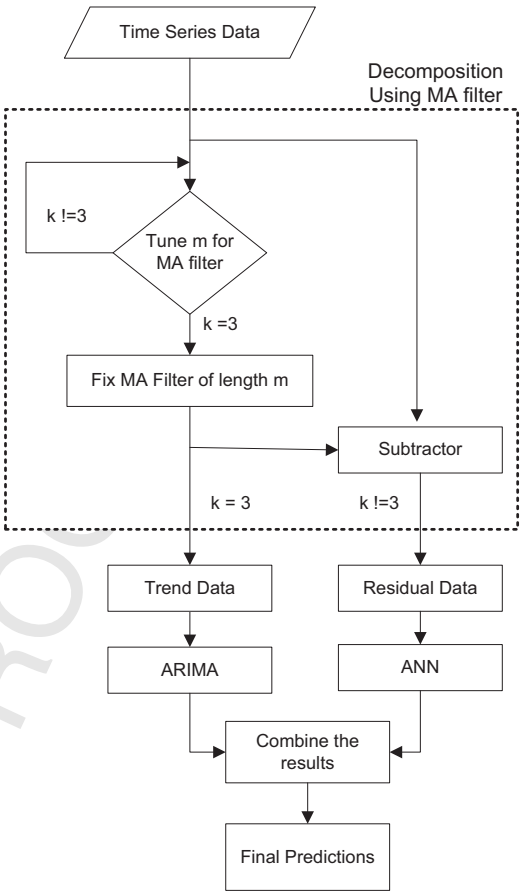

Fig. 2. Proposed method.

The proposed algorithm can give better accuracy compared with the other hybrid models discussed in the previous section, which first directly fit an ARIMA model to the given data. This can be understood from the following reasoning. A linear sequence cannot be accurately modeled by a nonlinear model, and vice versa. A low-volatility series is Gaussian in nature and an ARIMA model suits it better, which implies that it can be modeled accurately using a linear model. Therefore, when the time series is less volatile, it can be considered as a linear sequence. Similarly, if it is highly volatile in nature, it can be considered as a nonlinear sequence. If a linear sequence is modeled by a linear model, the model error will be small. The case for a nonlinear sequence is similar. So, when a given data set is decomposed into low- and high-volatility components, which are almost linear and nonlinear components, respectively, the total model error will be small. On the other hand, if an ARIMA model is directly fit to the data, the separation of linear and nonlinear components is not performed. So, there is a chance that part of the linear component will be modeled by a nonlinear model, resulting in an increase in the model error. Hence, the proposed model can give more accurate results than Zhang's and Khashei and Bijari's models. This fact has been verified using simulated and experimental data sets.

Results and discussion

The proposed method, along with the four modeling techniques discussed in "ARIMA- and ANN-based modeling techniques" section, was applied to simulated data and also to time series data of various kinds. A detailed description of the results is presented in this section. Before any further discussion of the results, however, the performance measures used for comparison of prediction accuracy will be discussed.

Please cite this article in press as: C.N. Babu, B.E. Reddy, A moving-average-filter-based hybrid ARIMA–ANN model for forecasting time series data, Appl. Soft Comput. J. (2014), http://dx.doi.org/10.1016/j.asoc.2014.05.028

Table 1
Performance comparison for simulated data.

	One-step-ahead		Three-step-ahead	
	MAE	MSE	MAE	MSE
ARIMA	0.3925	0.1896	0.4474	0.3070
ANN	0.3582	0.1764	0.3857	0.2372
Zhang model	0.2595	0.0841	0.3487	0.2032
Khashei and Bijari model	0.2365	0.0701	NA	NA
Proposed model	0.1884	0.0507	0.2951	0.1445

Table 2
Performance comparison for sunspot data.

	One-step-ahead		Five-step-ahead	
	MAE	MSE	MAE	MSE ($\times 10^3$)
ARIMA	14.6630	308.1491	22.0840	1.0993
ANN	14.3513	285.1216	21.7042	0.9384
Zhang model	14.2233	298.5670	20.8829	1.0309
Khashei and Bijari model	13.4053	269.5369	NA	NA
Proposed model	9.8718	155.5646	17.7869	0.5769

The two performance measures considered for accuracy comparison in this paper are the mean absolute error (MAE) and mean squared error (MSE), given in (11) and (12), respectively. The MAE signifies the average of the absolute errors over a given prediction horizon, and the MSE signifies the average of the squared errors over the same prediction horizon. The smaller the measure, the better the model. In both (11) and (12), $E\{\cdot\}$ is the expectation operation, ni and nf are the start and end points of the forecast horizon or prediction interval, $y_{j,\text{actual}}$ is the actual value of the time series, and $y_{j,\text{predicted}}$ is the forecasted time series value at the instant j.

To evaluate the performance of a prediction model, the forecast horizon must be sufficiently long. If it is too small, the expectation, i.e., the mean in (11) and (12), will not be accurate enough. If the forecast horizon is very long, the time taken for processing will be high. So, the forecast horizon must be chosen appropriately.

Results for simulated data

A known data set was generated by adding a linear data process AR(2,0,0) to a nonlinear data set simulated by an ANN. The nonlinear model is represented as $N^{x,y,z}$, where x is the number of input nodes, y is the size of the hidden layer, and z is the number of output nodes. In this paper, $z = 1$ was chosen. The simulated ANN data corresponded to $N^{2,2,1}$. The resultant data were modeled using ARIMA and ANN models, Zhang's hybrid model [14], Khashei and Bijari's hybrid model [15], and the proposed hybrid model.

For the ARIMA modeling, a suitable model order was found using the R software package, and then this model was fit to the data using MATLAB. For the ANN and hybrid modeling, MATLAB was used. The total number of data points taken was 100. In the case of one-step prediction, the forecast horizon was 10. The multistep-ahead prediction performed on the data was a three-step-ahead prediction. For this, the forecast horizon considered was 30. The results for the performance measures obtained with this data are shown in Table 1. The actual time series data are shown in Fig. 3, the predictions for one-step-ahead forecasting are shown in Fig. 4, and the predictions for three-step-ahead forecasting are shown in Fig. 5. From Table 1 and the plots in Fig. 4, it can be seen that the proposed method gives better performance than all of the other models used for comparison.

In the case of multistep-ahead prediction, the results from Khashei and Bijari's model are not given, because they were almost same as those from Zhang's model. This is because in Khashei and Bijari's model [15], the ANN inputs should be the ARIMA-predicted present value, the past errors, and the past actual data values, for one-step-ahead prediction. If this model has to be used for multistep-ahead prediction, multiple future values have to be predicted. Consider a data set having 100 points. For a five-step-ahead prediction, the 101st to 105th values have to be predicted based on only the first 100 data points. In this case, if Khashei and Bijari's model is used, the 101st point can be predicted. To predict the 102nd point, the ARIMA-predicted 102nd point and the past errors are available, but among the past actual values needed by the model, the 101st actual value will not be known, so the model is not suitable for multistep-ahead prediction. One way to overcome this problem would be to use the 101st model prediction instead of the 101st actual value, but this reduced the model accuracy and it was observed that the model accuracy was no better than that of Zhang's model. So, for multistep-ahead prediction, the results from this model have not been included, as they do not provide an apt comparison.

From the results in Table 1, it can be verified that the proposed model gives better performance than the other models for the simulated data, i.e., a known time series data set. With this success in mind, the discussion now progresses towards applying the model to real time series data sets obtained from various applications. The time series data considered in this paper are sunspot data, electricity price data from the Australian National Electricity Market, and the close prices of stock from the New York Stock Exchange. Each one of these is discussed in the following sections.

Results for sunspot data

Sunspot time series data from 1700 to 1987 were considered for this study; these data were a set of 288 points. For one-step-ahead prediction, the forecast horizon was chosen as 25 data points. The multistep prediction performed in this case was five-step-ahead prediction. The forecast horizon considered was 50. The prediction performance results for all the models are tabulated in Table 2 for

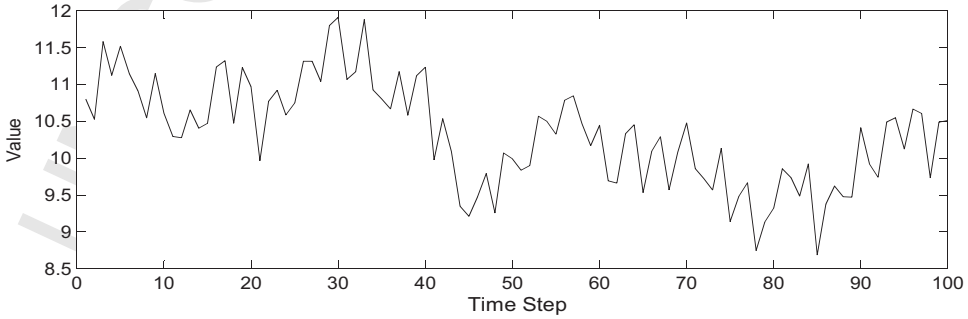

Fig. 3. Simulated time series data.

Please cite this article in press as: C.N. Babu, B.E. Reddy, A moving-average-filter-based hybrid ARIMA–ANN model for forecasting time series data, Appl. Soft Comput. J. (2014), http://dx.doi.org/10.1016/j.asoc.2014.05.028

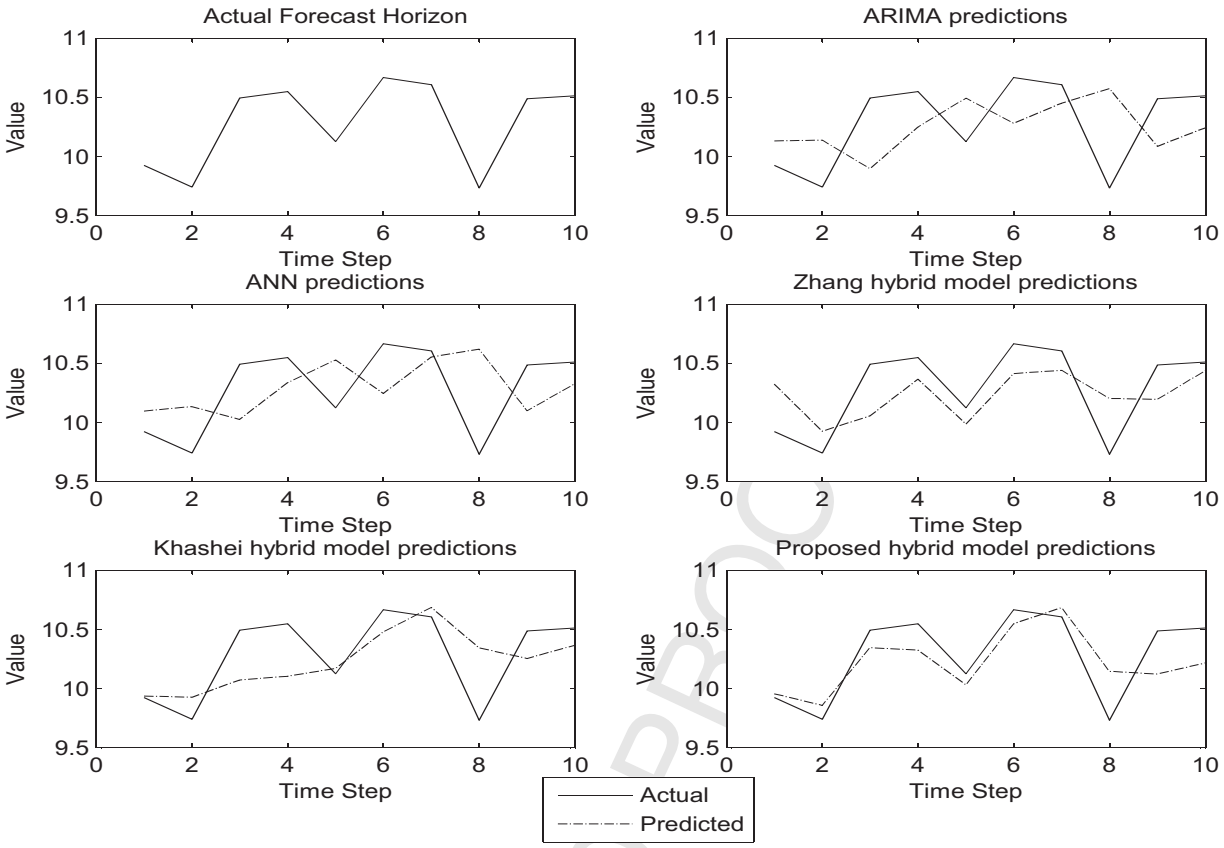

Fig. 4. Results for simulated time series data.

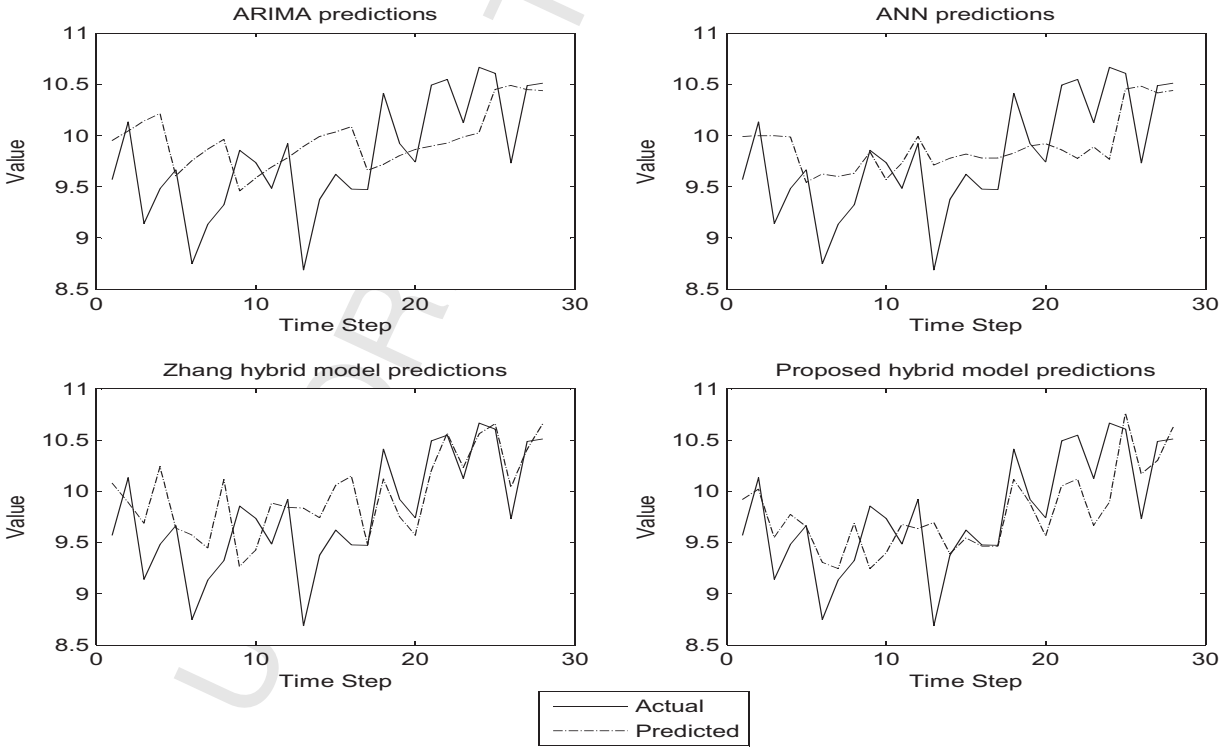

Fig. 5. Three-step-ahead predictions for simulated data using various models.

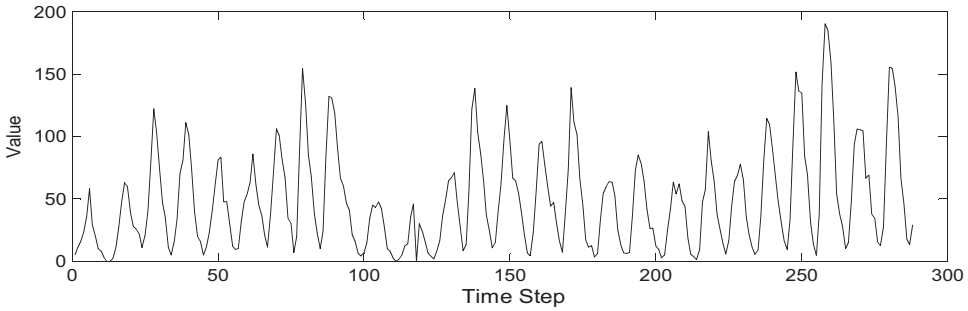

Fig. 6. Sunspot time series data.

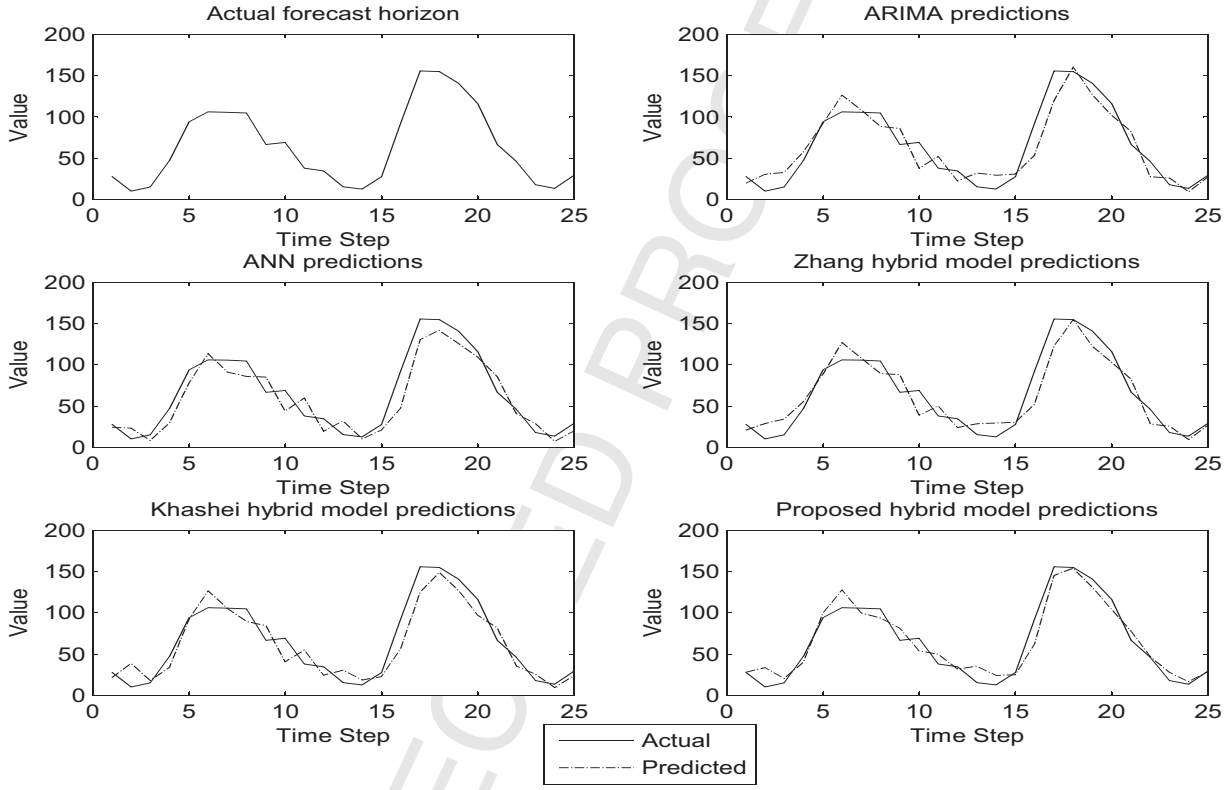

Fig. 7. One-step-ahead predictions for sunspot data using various models.

both of these cases. The original time series data are shown in Fig. 6. The predictions for the one-step-ahead forecast are shown in Fig. 7, and those for the five-step-ahead forecast are shown in Fig. 8. From the table and the figures shown, it can be seen that the proposed method outperforms all of the other models used for comparison in terms of MSE and MAE.

$$MAE = E\{|y_{\text{actual}} - y_{\text{predicted}}|\} = \frac{1}{nf - ni + 1}\left(\sum_{j=ni}^{nf} |y_{j,\,\text{actual}} - y_{j,\,\text{predicted}}|\right) \quad (11)$$

$$MSE = E\{|y_{\text{actual}} - y_{\text{predicted}}|^2\} = \frac{1}{nf - ni + 1}\left(\sum_{j=ni}^{nf} |y_{j,\,\text{actual}} - y_{j,\,\text{predicted}}|^2\right) \quad (12)$$

In the proposed method, when the MA filter was used, the length was fixed at 37. The given time series data had a kurtosis value of 3.6, indicating that it was highly volatile. After using the filter, the smoothed component, which we call the trend component, had a kurtosis of 3, which indicated that it had low volatility and the ARIMA method was suitable for modeling. The residual component obtained from the filter had a kurtosis of 3.2, indicating that it was a highly volatile component and the ANN method was suitable. Thus the proposed model was applied as per the discussion in "Proposed hybrid ARIMA–ANN model" section. Multistep forecasting generally has less accuracy than one-step-ahead forecasting. As seen from the results tabulated in Table 2, this can be observed to be true in this case.

Results for electricity price data

The electricity price data studied here were data for New South Wales from the Australian National Electricity Market [32] for the month of May in 2013. The data were available for every half-an-hour. This was converted first to one-hour data, so that there were

Please cite this article in press as: C.N. Babu, B.E. Reddy, A moving-average-filter-based hybrid ARIMA–ANN model for forecasting time series data, Appl. Soft Comput. J. (2014), http://dx.doi.org/10.1016/j.asoc.2014.05.028

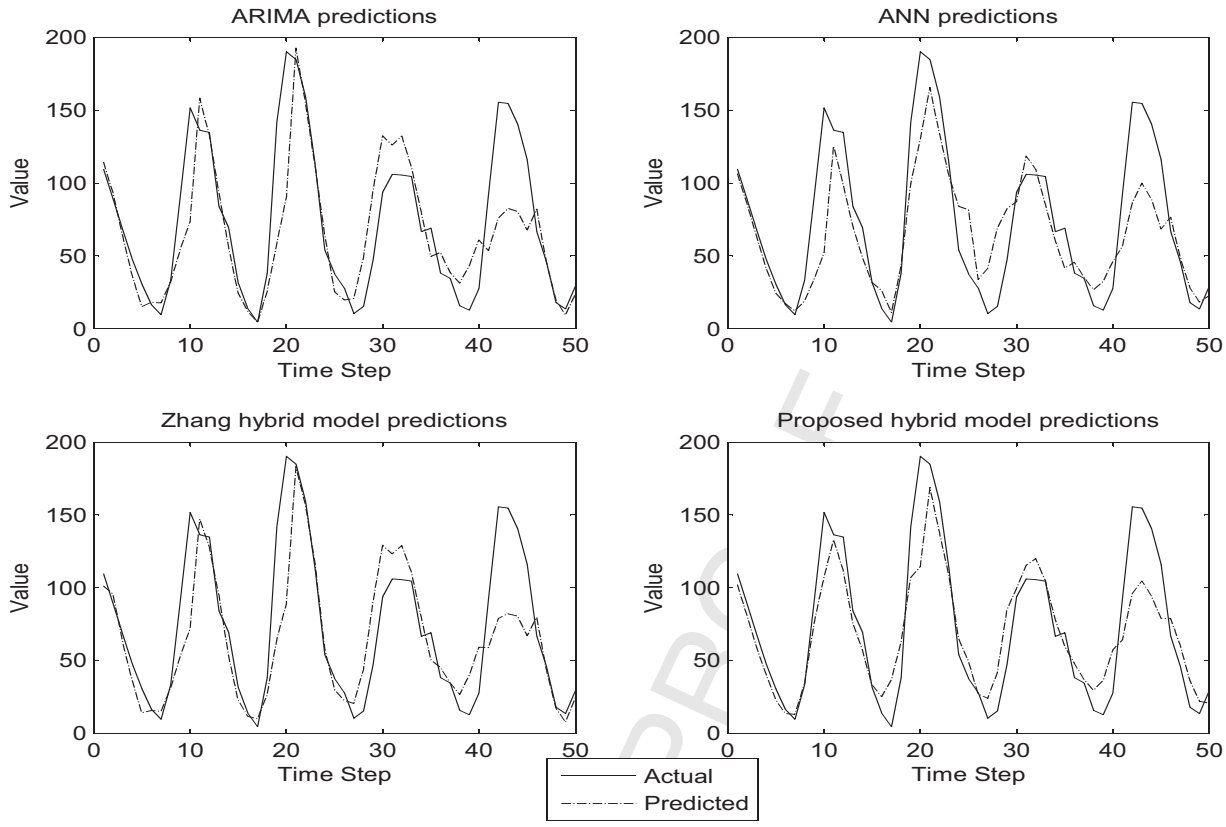

Fig. 8. Five-step-ahead predictions for sunspot data using various models.

Table 3
Performance comparison for electricity price data.

	One-step-ahead		24-step-ahead (1 day ahead)	
	MAE	MSE	MAE	MSE
ARIMA	4.8233	36.4544	8.6682	121.5435
ANN	3.7374	22.4304	8.4138	109.8774
Zhang model	3.9204	27.0377	7.9437	107.3933
Khashei and Bijari model	3.8346	26.1396	NA	NA
Proposed model	3.2342	18.2793	5.3219	53.0071

24 data points for 1 day. So, for 1 month, 744 data points representing hourly electricity price data were taken as the given time series data set for forecasting. In one-step-ahead forecasting, the forecast horizon considered was 24 data points. Also, 24-step-ahead (1 day ahead) forecasting was performed, where the forecast horizon was taken as 7 days, which means 168 data points. The prediction performance results for all of the models for both one-step-ahead and one-day-ahead forecasts are shown in Table 3. The original data set is shown in Fig. 9. The one-step-ahead predictions are shown in Fig. 10, and the 1-day-ahead predictions in Fig. 11. From the table and the figures, it can be seen that the proposed method outperforms the others for both one-step-ahead and multistep-ahead forecasting.

When the proposed prediction model was used on the data, the data had a kurtosis of 28.4, indicating that the data were very highly volatile in nature. When these data were passed through an MA

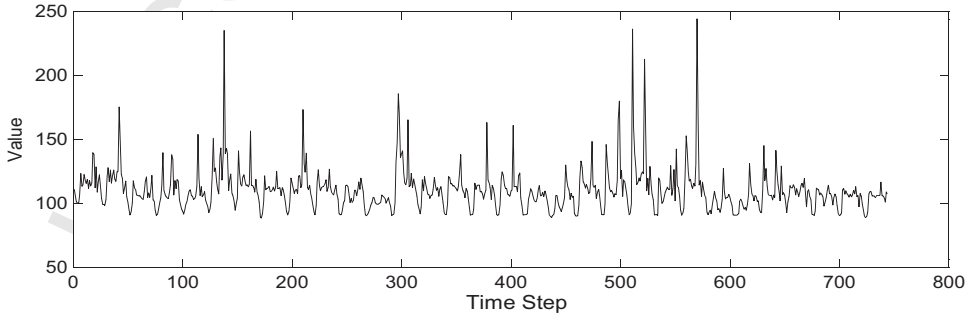

Fig. 9. Electricity price time series data.

Please cite this article in press as: C.N. Babu, B.E. Reddy, A moving-average-filter-based hybrid ARIMA–ANN model for forecasting time series data, Appl. Soft Comput. J. (2014), http://dx.doi.org/10.1016/j.asoc.2014.05.028

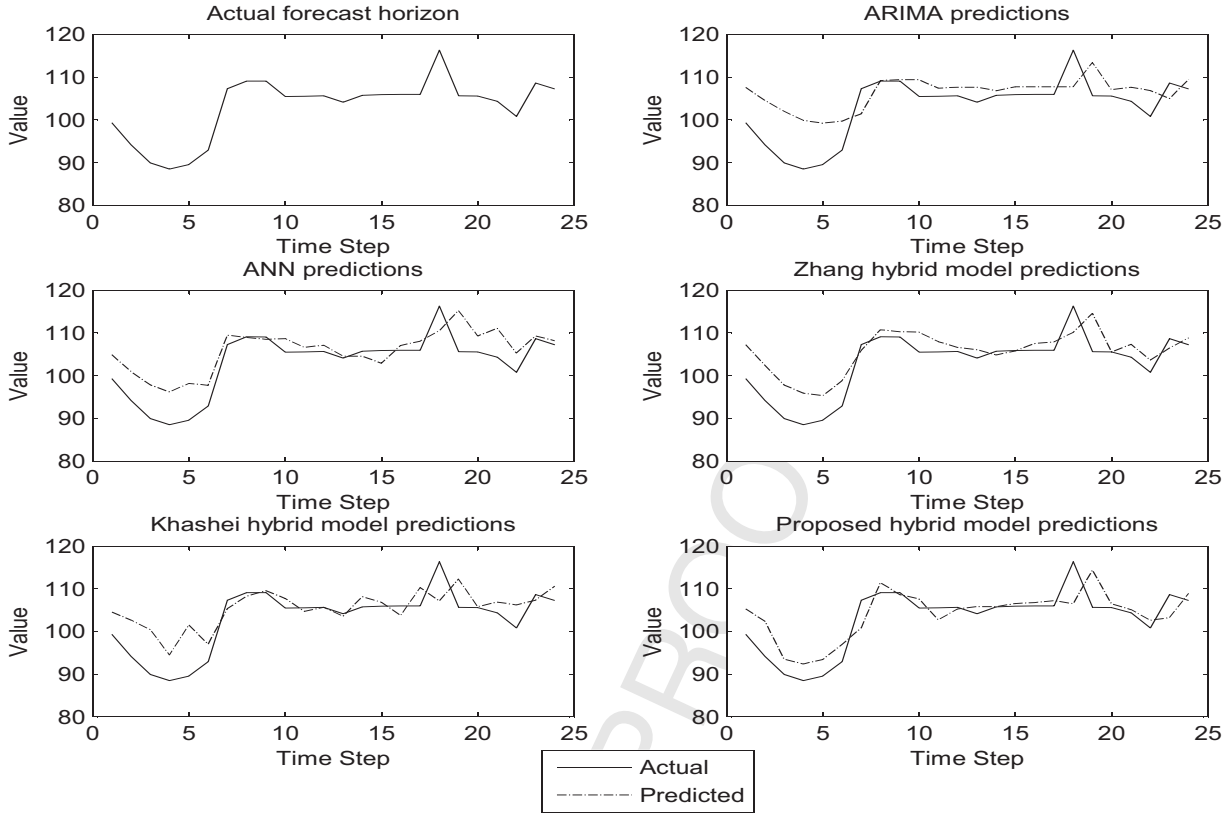

Fig. 10. One-step-ahead predictions for electricity price data using various models.

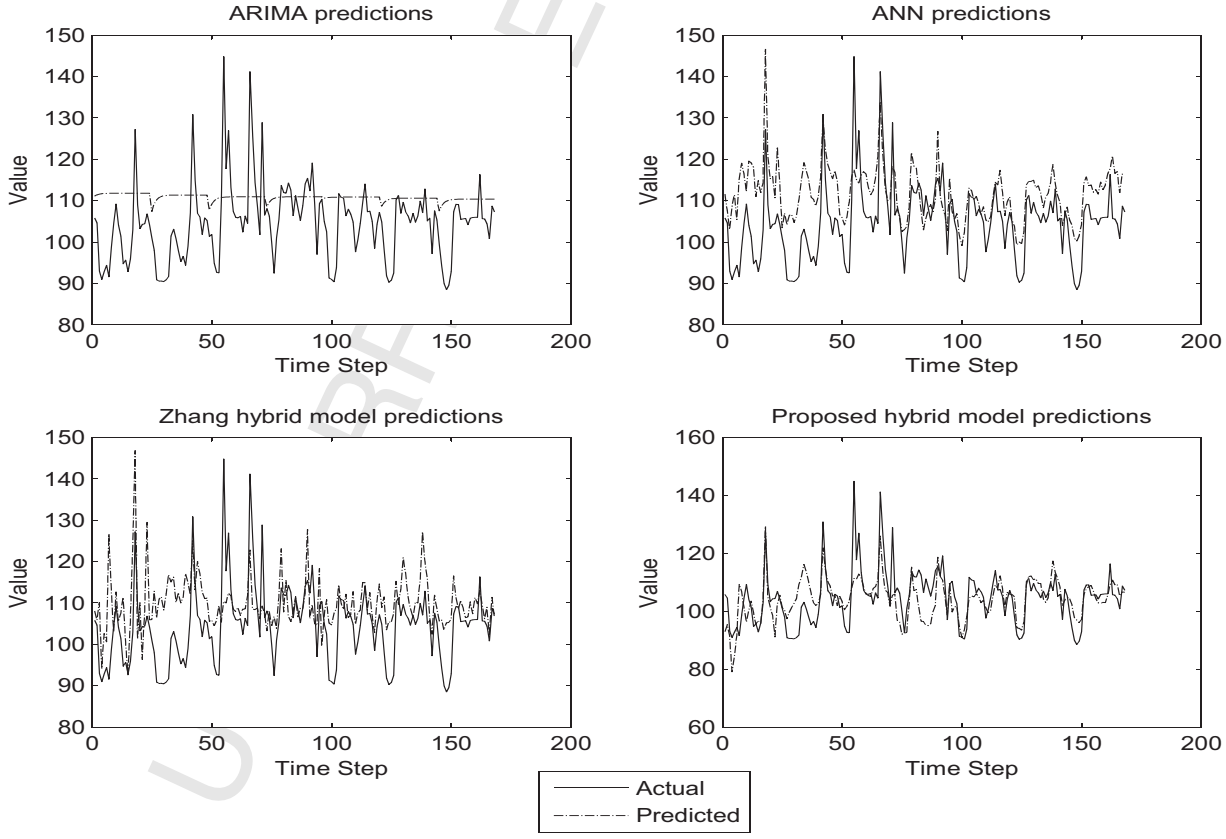

Fig. 11. Five-step-ahead predictions for electricity price data using various models.

Please cite this article in press as: C.N. Babu, B.E. Reddy, A moving-average-filter-based hybrid ARIMA–ANN model for forecasting time series data, Appl. Soft Comput. J. (2014), http://dx.doi.org/10.1016/j.asoc.2014.05.028

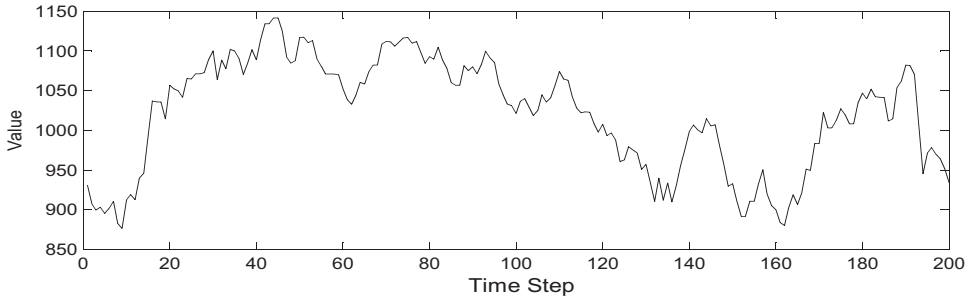

Fig. 12. L&T stock market time series data.

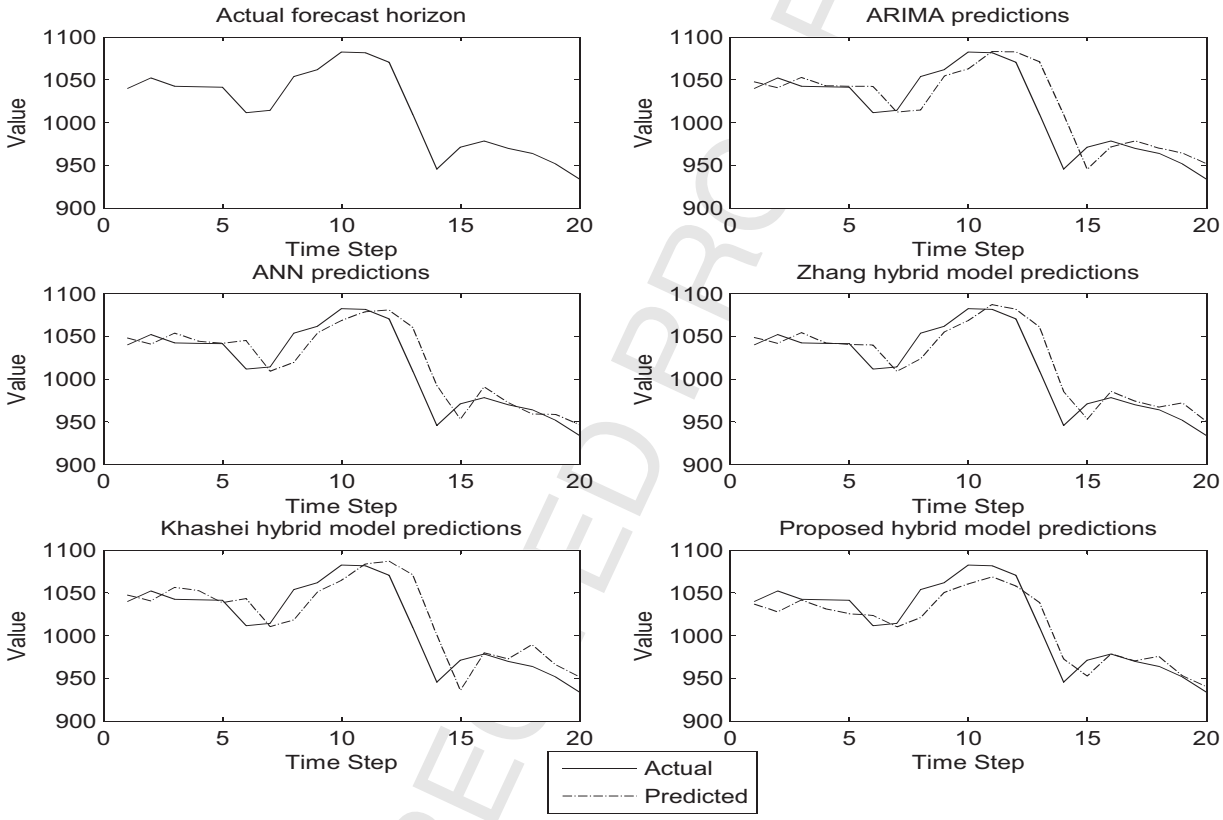

Fig. 13. One-step-ahead predictions for financial data using various models.

filter of length 25, the trend component had a kurtosis of 3 and the residual component had a kurtosis of 18.2. An ARIMA model was fit to the trend component and an ANN was fit to the residual component. Note that when an ARIMA model is fit according to either Zhang's or Khashei and Bijari's model, the order of the model is same. But in the proposed method, the order of the ARIMA model is different. For example, in this case, for Zhang's and Khashei and Bijari's models, the ARIMA model used was ARIMA(1,0,1), but for the proposed method the ARIMA model was ARIMA(1,1,1). This was because when the trend component was separated, the ARIMA model fit to the data was entirely different from the ARIMA model fit directly to the data. From the results, it can be concluded that the proposed method outperforms the other models discussed in this paper.

Results for stock market data

The close prices of the Larsen and Turbo (L&T) company stock for 200 trading days before May 31, 2013 were chosen as the time series data for study. The data set was taken from [33]. For one-step-ahead prediction, the forecast horizon was chosen as 20 data points. Three-step-ahead forecasting was also performed, for which the forecast horizon was taken as 21 data points. The five models discussed in this paper were applied to these time series data, and the prediction performance results are tabulated in Table 4. The original data are shown in Fig. 12. The one-step-ahead predictions are shown in Fig. 13, and the three-step-ahead predictions

Table 4
Performance comparison for L&T stock time series data.

	One-step-ahead		Three-step-ahead	
	MAE	MSE	MAE	MSE
ARIMA	17.4141	629.7310	26.8604	1.5685
ANN	14.8967	428.7039	22.8979	1.0514
Zhang model	14.7062	389.8686	18.9346	0.9732
Khashei and Bijari model	14.8407	401.2017	NA	NA
Proposed model	12.8226	261.6390	15.9454	0.7422

Please cite this article in press as: C.N. Babu, B.E. Reddy, A moving-average-filter-based hybrid ARIMA–ANN model for forecasting time series data, Appl. Soft Comput. J. (2014), http://dx.doi.org/10.1016/j.asoc.2014.05.028

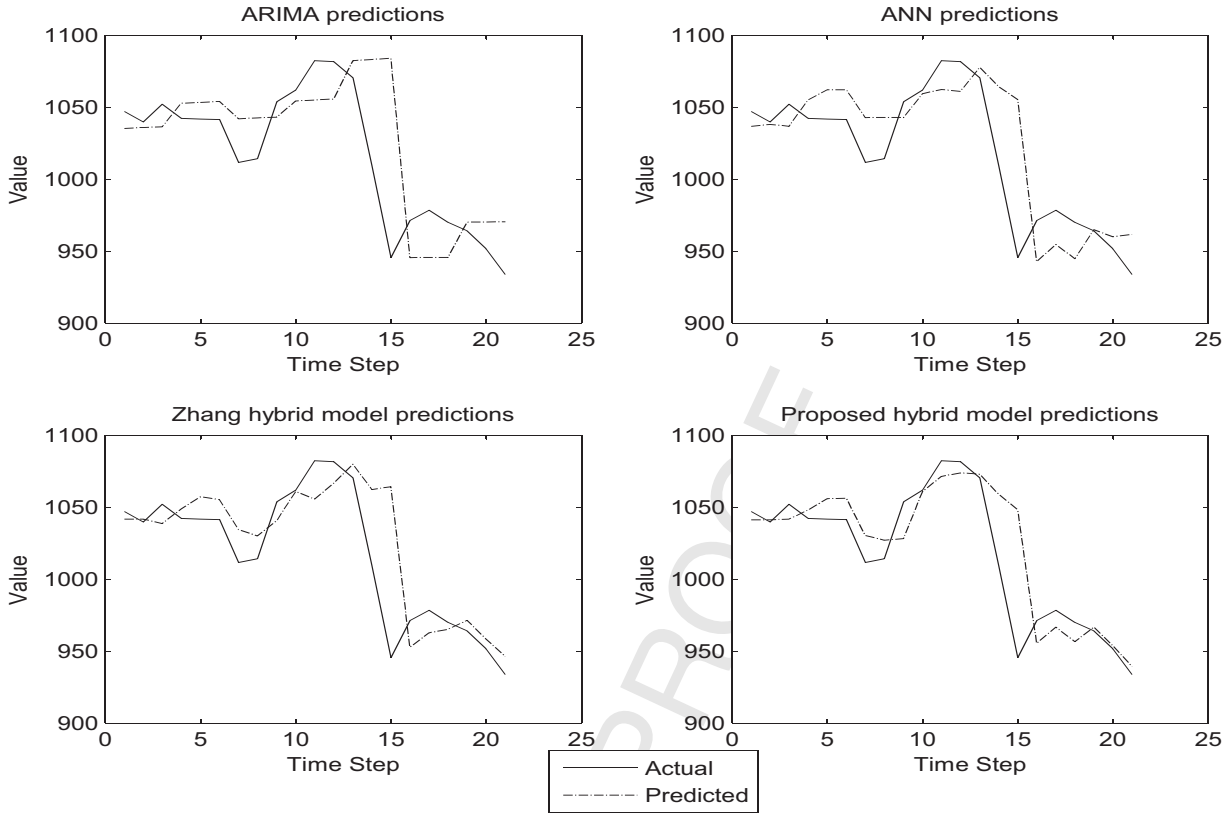

Fig. 14. Five-step-ahead predictions for financial data using various models.

in Fig. 14. From the table and the results, it can be seen that the proposed method outperforms the other models. In the case of one-step-ahead prediction, the performance of Zhang's and Khashei and Bijari's models showed a very small improvement, whereas the proposed model showed significant improvement compared with Zhang's model. Also, in this case, the accuracy of Zhang's model was better than that of Khashei and Bijari's model for one-step-ahead prediction. But the accuracy of the proposed method was much better than that of the others in terms of both MSE and MAE, as can be seen from the table.

The MA filter length in the proposed method was 90. The given time series data had a kurtosis of 2.04. The trend component of the MA filter had a kurtosis of 3, and the kurtosis of the residual component was 2. It can be seen that the original data had a kurtosis less than 3, so it was considered as highly volatile. Whenever the kurtosis is not 3, the data are highly volatile. If the kurtosis is greater than 3, the data are both highly outlier-prone and highly volatile. If the kurtosis is less than 3, the data are still highly volatile but less outlier-prone [34]. The data set considered in this section was less outlier-prone, whereas the sunspot and electricity price data were highly outlier-prone. Irrespective of whether or not the data were highly outlier-prone, the proposed model outperformed the other models, as seen from all of the results.

Conclusion

Time series data originating from various applications, in general, comprise both linear and nonlinear variations. Linear ARIMA models and nonlinear ANN models cannot individually model such data accurately. Hybrid models which combine the strengths of ARIMA and ANN models are better than the individual types of models, as they are capable of exploiting the advantages of both types of models simultaneously. In the present study, a new hybrid ARIMA–ANN-based prediction model has been developed based on the statistical properties of ARIMA sequences. The model uses an MA filter to decompose the given time series into two data sets. Then ARIMA and ANN models are applied suitably to these decompositions. The forecasts from the hybrid model are obtained by adding the forecasts from the two individual models. This hybrid model is capable of both one-step-ahead and multistep-ahead prediction. The model was applied to simulated time series data and to three available data sets of different kinds, namely sunspot data, electricity price data, and financial data. For both one-step-ahead and multistep-ahead prediction, the proposed hybrid model has higher prediction accuracy in terms of MAE and MSE than several other models, such as ARIMA and ANN models and some existing hybrid ARIMA–ANN models. Thus the hybrid model proposed in the present paper becomes a simple and accurate prediction model in many applications.

References

[1] E. Gonzalez-Romera, M. Jaramillo-Moran, D. Carmona-Fernandez, Monthly electric energy demand forecasting based on trend extraction, IEEE Trans. Power Syst. 21 (4) (2006) 1946–1953, http://dx.doi.org/10.1109/TPWRS.2006.883666.
[2] J. Contreras, R. Espinola, F. Nogales, A. Conejo, ARIMA models to predict next-day electricity prices, IEEE Trans. Power Syst. 18 (3) (2003) 1014–1020, http://dx.doi.org/10.1109/TPWRS.2002.804943.
[3] K. Suresh, S. Krishna Priya, Forecasting sugarcane yield of Tamilnadu using ARIMA models, Sugar Tech. 13 (1) (2011) 23–26, http://dx.doi.org/10.1007/s12355-011-0071-7.
[4] J.-J. Wang, J.-Z. Wang, Z.-G. Zhang, S.-P. Guo, Stock index forecasting based on a hybrid model, Omega 40 (6) (2012) 758–766, http://dx.doi.org/10.1016/j.omega.2011.07.008.
[5] E. Cadenas, W. Rivera, Wind speed forecasting in three different regions of Mexico, using a hybrid ARIMA–ANN model, Renew. Energy 35 (12) (2010) 2732–2738, http://dx.doi.org/10.1016/j.renene.2010.04.022.
[6] C. Babu, B. Reddy, Predictive data mining on average global temperature using variants of ARIMA models, in: 2012 International Conference on Advances

Please cite this article in press as: C.N. Babu, B.E. Reddy, A moving-average-filter-based hybrid ARIMA–ANN model for forecasting time series data, Appl. Soft Comput. J. (2014), http://dx.doi.org/10.1016/j.asoc.2014.05.028

in Engineering, Science and Management (ICAESM), 2012, pp. 256–260 http://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=6215607
[7] U. Orhan, Real-time CHF detection from ECG signals using a novel discretization method, Comput. Biol. Med. 43 (10) (2013) 1556–1562, http://dx.doi.org/10.1016/j.compbiomed.2013.07.015 http://www.sciencedirect.com/science/article/pii/S0010482513001893
[8] R.F.A. Lapedes, Nonlinear signal processing using neural networks: Prediction and system Modeling, Tech. Rep. Technical Report LAUR-87-2662, Los Alamos National Laboratory, Los Alamos, NM, 1987.
[9] D. Singhal, K. Swarup, Electricity price forecasting using artificial neural networks, Int. J. Elec. Power Energy Syst. 33 (3) (2011) 550–555, http://dx.doi.org/10.1016/j.ijepes.2010.12.009.
[10] W.-S. Chen, Y.-K. Du, Using neural networks and data mining techniques for the financial distress prediction model, Expert Syst. Appl. 36 (2 Pt 2) (2009) 4075–4086, http://dx.doi.org/10.1016/j.eswa.2008.03.020.
[11] C.H.F. Toro, S.G. Meire, J.F. Glvez, F. Fdez-Riverola, A hybrid artificial intelligence model for river flow forecasting, Appl. Soft Comput. 13 (8) (2013) 3449–3458, http://dx.doi.org/10.1016/j.asoc.2013.04.014.
[12] B.R. Chang, H.F. Tsai, Novel hybrid approach to data-packet-flow prediction for improving network traffic analysis, Appl. Soft Comput. 9 (3) (2009) 1177–1183, http://dx.doi.org/10.1016/j.asoc.2009.03.003.
[13] J. Reyes, A. Morales-Esteban, F. Martnez-lvarez, Neural networks to predict earthquakes in Chile, Appl. Soft Comput. 13 (2) (2013) 1314–1328, http://dx.doi.org/10.1016/j.asoc.2012.10.014.
[14] G. Zhang, Time series forecasting using a hybrid ARIMA and neural network model, Neurocomputing 50 (0) (2003) 159–175, http://dx.doi.org/10.1016/S0925-2312(01)00702-0.
[15] M. Khashei, M. Bijari, A novel hybridization of artificial neural networks and ARIMA models for time series forecasting, Appl. Soft Comput. 11 (2) (2011) 2664–2675, http://dx.doi.org/10.1016/j.asoc.2010.10.015.
[16] H.H. Arash Miranian, Majid Abdollahzade, Day-ahead electricity price analysis and forecasting by singular spectrum analysis, IET Gener. Transm. Dis. 7 (4) (2013) 337 L 346, http://dx.doi.org/10.1049/iet-gtd.2012.0263.
[17] D. mer Faruk, A hybrid neural network and ARIMA model for water quality time series prediction, Eng. Appl. Artif. Intell. 23 (4) (2010) 586–594, http://dx.doi.org/10.1016/j.engappai.2009.09.015 http://www.sciencedirect.com/science/article/pii/S0952197609001390
[18] Y. Sai, Z. Yuan, K. Gao, Mining stock market tendency by RS-based support vector machines, in: Proceedings of the 2007 IEEE International Conference on Granular Computing, GRC '07, IEEE Computer Society, Washington, DC, USA, 2007, p. 659, http://dx.doi.org/10.1109/GRC.2007.99.
[19] M.R. Hassan, A combination of hidden Markov Model and fuzzy model for stock market forecasting, Neurocomputing 72 (16-18) (2009) 3439–3446, http://dx.doi.org/10.1016/j.neucom.2008.09.029.
[20] M. Olsson, L. Soder, Modeling real-time balancing power market prices using combined SARIMA and Markov processes, IEEE Trans. Power Syst. 23 (2) (2008) 443–450, http://dx.doi.org/10.1109/TPWRS.2008.920046.
[21] J. Ilow, Forecasting network traffic using FARIMA models with heavy tailed innovations, in: 2000 IEEE International Conference on Acoustics, Speech, and Signal Processing, ICASSP '00, vol. 6, 2000, pp. 3814–3817, http://dx.doi.org/10.1109/ICASSP.2000.860234.
[22] W. Liu, B. Morley, Volatility forecasting in the Hang Seng Index using the GARCH approach, Asia-Pacific Financial Markets 16 (1) (2009) 51–63, http://dx.doi.org/10.1007/s10690-009-9086-4.
[23] S.Y. Sohn, M. Lim, Hierarchical forecasting based on AR-GARCH model in a coherent structure, Eur. J. Operat. Res. 176 (2) (2007) 1033–1040.
[24] G.M. Ferenstein, Elzbieta, Modelling stock returns with AR-GARCH processes, SORT 28 (1) (2004) 55–68.
[25] A. Conejo, M. Plazas, R. Espinola, A. Molina, Day-ahead electricity price forecasting using the wavelet transform and ARIMA models, IEEE Trans. Power Syst. 20 (2) (2005) 1035–1042, doi:10.1109/TPWRS.2005.846054.
[26] C. Chen, J. Hu, Q. Meng, Y. Zhang, Short-time traffic flow prediction with ARIMA-GARCH model, in: 2011 IEEE Proceedings on Intelligence Vehicles Symposium IV, 2011, pp. 607–612, http://dx.doi.org/10.1109/IVS.2011.5940418.
[27] A. Jain, A.M. Kumar, Hybrid neural network models for hydrologic time series forecasting, Appl. Soft Comput. 7 (2) (2007) 585–592.
[28] G.E.P. Box, G. Jenkins, Time Series Analysis, Forecasting and Control, Holden-Day Incorporated, 1990.
[29] Gaussian maximum likelihood estimation for ARMA models. I. Time series, J. Time Series Anal. 27(6) (2006) 857–875.
[30] S.O. Haykin, Neural Networks and Learning Machines, 3rd edition, Prentice Hall, 2008.
[31] M.H. Hayes, Statistical Digital Signal Processing and Modeling, John Wiley & Sons, 1996.
[32] Australian Energy Market Operator (AEMO): Price and Demand Data Sets. Available at: http://www.aemo.com.au/data/price_demand.html
[33] The NSE India Data. Available at: http://www.nseindia.com/products/content/equities/equities/eq_security.htm
[34] L.T. DeCarlo, On the meaning and use of kurtosis, Psychol. Methods 2 (3) (1997) 292–307.

Please cite this article in press as: C.N. Babu, B.E. Reddy, A moving-average-filter-based hybrid ARIMA–ANN model for forecasting time series data, Appl. Soft Comput. J. (2014), http://dx.doi.org/10.1016/j.asoc.2014.05.028